Medical Oncology (2026) 43:22
https://doi.org/10.1007/s12032-025-03127-z

REVIEW ARTICLE

Deep learning in bone marrow cytomorphology: advances in segmentation, classification, and clinical translation

Shahid Mehmood[1,2] · Muhammad Zubair[2] · Farman Matloob Khan[3,4] · Asghar Ali Shah[5] · Sagheer Abbas[6] · Khan Muhammad Adnan[7]

Received: 22 June 2025 / Accepted: 10 November 2025

Abstract

Bone marrow cytomorphology analysis has long been a prerequisite for diagnosing hematologic disorders. It is often a tedious and subjective activity through conventional manual methods. The recent advancement in deep learning (DL) has a bright future with potential automation in cell classification, segmentation, and diagnosis workflows. This review outlines the current state-of-the-art application of DL for bone marrow analysis, focusing mainly on challenges such as data scarcity, class imbalance, and inter-center variability. It includes an evaluation of publicly available datasets together with strategies for overcoming limitations, including synthetic data generation and federated learning. This review analyzes segmentation techniques, ranging from the classical watershed algorithm to novel U-Net (a specialized neural network for image segmentation) and Vision Transformer hybrids, for their efficacy in isolating cells, tissue subsystems, and subcellular structures. DL classification models, including convolutional neural networks (CNNs), attention-based architectures, and ensembles, demonstrate expert-level accuracy in detecting malignancies like acute myeloid leukemia (AML) and myelodysplastic syndromes (MDS). Clinical applications involve AI-driven platforms that integrate into digital pathology workflows, with early reports suggesting reductions in diagnostic turnaround time and inter-observer variability. For crowded images of the bone marrow, context-aware analysis is enhanced by hybrid models combining CNN and Transformer architectures. However, there are challenges with generalizability and interpretability, as well as difficulties in integrating multimodal data. Therefore, future directions emphasize validation, explainable AI, and integrating cytomorphology with genetic and flow cytometry data. This interdisciplinary work aims to bridge the fields of AI and hematopathology towards standardization and precise diagnostics, which would improve clinical decision-making and patient outcomes.

Keywords Bone marrow · Artificial intelligence · Deep neural network · Aute lymphoblastic leukemia · Convolutional neural network

Abbreviations

AI	Artificial intelligence
AML	Acute myeloid leukemia
ALL	Acute lymphoblastic leukemia
BMA	Bone marrow aspirate
BM	Bone marrow
BMMC	Bone marrow mononuclear cells
CBC	Complete blood count

✉ Asghar Ali Shah
asghar.ali.shah@kateb.edu.af

✉ Khan Muhammad Adnan
adnan@gachon.ac.kr

[1] Department of Computer Science, Bahria University, Lahore 54000, Pakistan

[2] Department of Computer Science, Riphah International University, Islamabad 45000, Pakistan

[3] College of Pharmacy, University of Sharjah, Sharjah, UAE

[4] College of Pharmacy, Universiti Teknologi MARA (UiTM), 42300 Puncak Alam, Selangor, Malaysia

[5] Department of Computer Science, Kateb University, Kabul, Afghanistan

[6] Department of Computer Science, Prince Mohammad Bin Fahd University, Alkhobar, Saudi Arabia

[7] Department of Software, Faculty of Artificial Intelligence and Software, Gachon University, Seongnam-si 13557, Republic of Korea

Published online: 24 November 2025

CNN	Convolutional neural network
DL	Deep learning
DNN	Deep neural network
DICOM	Digital imaging and communications in medicine
FISH	Fluorescence in situ hybridization
GPU	Graphics processing unit
H&E	Hematoxylin and eosin
HSV	Hue saturation value
HL7	Health level seven (data standard)
LIS	Laboratory information system
M:E	Myeloid to erythroid (ratio)
MDS	Myelodysplastic syndrome
ML	Machine learning
PACS	Picture archiving and communication system
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SVM	Support vector machine
TAT	Turnaround time
UMAP	Uniform manifold approximation and projection
WBC	White blood cells
WSI	Whole slide imaging

Introduction

Bone marrow cytomorphological examination remains the gold standard for diagnosing hematologic diseases, such as leukemias and myelodysplastic syndromes [1]. However, the traditional manual review of BM aspirate smears is labor-intensive and prone to significant inter-observer variability [2]. Hematopathologists must identify and count dozens of cell types in crowded smears, where subtle continuous maturation changes can blur distinctions between cell classes [3, 4]. These challenges, compounded by the need for expert training and the subjective nature of morphological assessments, motivate a pressing need for automated, reliable analysis methods. An efficient computational system could help overcome issues of time consumption and inconsistency in bone marrow evaluations, ultimately improving diagnostic speed and accuracy.

Artificial intelligence (AI), and DL in particular, has rapidly emerged as a powerful tool for image-based analysis in medicine [5]. In fields such as radiology and digital pathology, DL algorithms already excel at recognizing complex patterns in images, often rivaling human expertise [6, 7]. Similarly, for hematology, modern DL, especially CNNs, offers the ability to automatically learn discriminative morphological features from BM cell images, obviating the need for manual feature engineering [8]. This data-driven feature learning has led to remarkable improvements in performance: DL models have achieved high accuracy in detecting and classifying BM cells, in many cases surpassing the accuracy of traditional manual methods [9]. For example, recent studies report that CNN-based systems can reliably distinguish normal and malignant cells in BM smears, showing diagnostic performance that approaches that of expert hematologists [10]. Such promising results underscore the potential of AI to augment BM morphology analysis and reduce the diagnostic burden on clinicians.

Importantly, these advances are beginning to translate into clinical practice. Notably, a fully automated digital microscopy platform for BM smears (incorporating AI for cell analysis) was recently granted FDA approval, the first for BM morphology, highlighting a significant milestone in the clinical adoption of DL models [3]. Nevertheless, significant challenges remain before AI-based BM analysis can be widely and safely implemented. One key hurdle is data availability and variability: achieving robust, generalizable performance across different laboratories and patient populations requires large and heterogeneous training datasets to capture the diversity of BM presentations [3]. In practice, many studies still rely on small or single-center image datasets, and the class imbalance inherent in BM data (with some rare cell types) can lead to biased models that underperform on uncommon but critical cell classes [11]. Another technical challenge lies in cell segmentation. BM aspirate slides typically contain densely packed, overlapping cells, making it difficult for algorithms to accurately isolate individual cells for analysis [12]. Errors at the segmentation stage can propagate and affect downstream cell classification and diagnosis, so robust segmentation methods are crucial [13]. Furthermore, considerations such as staining variability, image focus/quality, and the need for explainability and clinician trust are all areas of concern as these AI systems transition from research to clinical use [14]. Addressing these issues will be essential to realize the potential benefits of DL in BM morphology.

Box 1: Key Technical Terms for Clinical Readers

To facilitate understanding for readers from clinical backgrounds, we provide brief definitions of essential deep learning and computer vision terms used throughout this review:

Convolutional Neural Network (CNN): A type of artificial intelligence model specifically designed for image analysis. CNNs automatically learn to recognize patterns in images (e.g., cell shapes, textures) through training on labeled examples, without requiring manual feature specification.

U-Net: A neural network architecture with an encoder-decoder structure resembling the letter 'U'. Widely used for medical image segmentation (identifying boundaries of cells or tissues), U-Net excels at precise localization while maintaining context.

Dice Coefficient (Dice Score): A metric quantifying overlap between two sets, commonly used to evaluate segmentation accuracy. Values range from 0 (no overlap) to 1 (perfect overlap); scores greater than 0.7 generally indicate good performance.

Intersection over Union (IoU): Another overlap metric for segmentation evaluation, calculated as the intersection divided by the union of predicted and true regions. Closely related to the Dice score but more sensitive to size discrepancies.

Transfer Learning: A strategy where a model pre-trained on large general image datasets (e.g., ImageNet) is adapted to specialized medical imaging tasks. This approach requires fewer labeled medical images and often improves performance.

Precision and Recall: Classification performance metrics. Precision measures the proportion of correct positive predictions (e.g., of cells labeled as blasts, how many truly are blasts?). Recall measures the proportion of actual positives correctly identified (of all true blasts, how many were detected?).

AUC (Area Under the Receiver Operating Characteristic Curve): A threshold-independent metric for classification performance, ranging from 0.5 (random guessing) to 1.0 (perfect classification). AUC $\geq$0.9 is generally considered excellent.

F1-Score: The harmonic mean of precision and recall, providing a balanced single metric for classification performance.

These terms will be briefly contextualized when first introduced in each section.

Given this field's growing interest and rapid developments, we present a comprehensive review of DL techniques for BM morphology analysis. We survey the current state-of-the-art and ongoing challenges across several key aspects: dataset challenges (data scarcity, class imbalance, and annotation issues), image segmentation techniques for detecting and isolating BM cells, and classification methods for identifying cell types and detecting malignancies. We also examine how these advances are being applied in clinical diagnostics, highlighting AI-driven tools for assisting in leukemia diagnosis and other hematological applications. Finally, we compare architectural approaches, ranging from single CNN-based models to hybrid CNN and Transformer networks, as well as ensemble strategies. We also discuss their relative merits in tackling BM analysis tasks. By bringing together perspectives from computer vision, machine learning, and hematopathology, this review aims to inform both technical and clinical audiences of the latest achievements in DL for BM morphology. This interdisciplinary effort at the intersection of AI and hematology holds great promise for more standardized, precise, and rapid BM diagnostic workflows in the near future.

Data landscape in bone marrow analysis

Available datasets

Despite these obstacles, recent efforts have produced several noteworthy BM morphology analysis datasets. Table

Table 1 Comparison of publicly available and proprietary datasets used for DL-based BM morphology analysis

Dataset name	Type	Size/Samples	Cell types/Labels	Stain/Modality	Public access
MLL Helmholtz Fraunhofer	Single-cell images (aspirate)	171,375 cells / 945 pts	21 classes	Pappenheim (MGG)	Yes (TCIA)
Target Recognition (Cheng et al.)	Object detection (aspirate)	13,059 cells / 1,204 fields	15 classes	Wright-Giemsa	Yes (Figshare)
BMCD-FGCD	Fine-grained single-cell images (aspirate)	92,335 cells	~40 classes	Giemsa	Yes (Google Drive)
BoMBR	Biopsy—Reticulin segmentation	201 images	Reticulin fibers + tissue	Reticulin stain	Yes (pre-print + repo)
BaMBo	Biopsy—Cellularity estimation	185 images	4 compartments (bone, fat, hematopoiesis, background)	H&E	Soon (bioRxiv preprint)
High-Res WBC (Bodzas et al.)	Single-cell WBC images (blood + marrow)	16,027 cells / 78 pts	9 WBC types, incl. blasts	MGG	Yes (Springer Nature)
Morphogo System	Large-scale aspirate smear dataset (proprietary)	> 2.8 million cells / 508 cases	25 + marrow cell classes	MGG (clinical routine)	No (Proprietary)
FlowCyt	Flow cytometry (non-image)	~30 samples; millions of cell events	5 cell types (T, B, Mono, Mast, HSPC)	Flow cytometry (12 markers)	Yes (GitHub)

The table summarizes dataset types (aspirate or biopsy), image modality and staining techniques, number of samples or cells, number of annotated cell types or compartments, and accessibility status. These datasets support various tasks such as cell classification, object detection, segmentation, and multimodal integration and collectively serve as critical benchmarks for training and evaluating AI models in hematopathology

1 summarizes key public (or soon-to-be public) datasets, highlighting their size, scope, and characteristics:

Bone marrow cytomorphology MLL Helmholtz Fraunhofer dataset

This is the largest publicly available dataset of BM cell images to date. It contains 171,375 single-cell images extracted from BM aspirate smears of 945 patients collected at the Munich Leukemia Laboratory (MLL) [15]. Samples were stained with May–Grünwald–Giemsa (Pappenheim) and imaged under a brightfield microscope at 40 × oil immersion. The dataset encompasses a broad spectrum of hematologic diagnoses and 21 distinct BM cell categories (covering all major lineages and maturation stages). Hematopathologists expertly labeled each cell, providing an exceptionally curated reference set. All data were acquired and annotated through a partnership of MLL (sample processing), Fraunhofer IIS (slide scanning technology), and Helmholtz Munich (post-processing). This collection is accessible via The Cancer Imaging Archive (TCIA) under a CC BY-4.0 license (https://doi.org/10.7937/TCIA.AXH3-T579). It was introduced by Matek et al. [16], who demonstrated that deep CNNs trained on this dataset achieved high accuracy in classifying BM cell morphologies, underscoring the value of large-scale data for automated hematopathology.

Dataset for target recognition of bone marrow aspirate cells

Released in 2023 by Cheng et al. [17] (associated with an improved YOLOv7 model study), this dataset focuses on object detection in BM smears. It comprises 1,204 cropped field images (each ~ 600 × 600 pixels) of Wright–Giemsa-stained BM aspirate smears. Within these field images, 13,059 individual cells are annotated with bounding boxes and class labels spanning 15 BM cell types. The categories include a mix of myeloid lineage cells (e.g., neutrophils, eosinophils at various maturation stages) and erythroid precursors (e.g., polychromatic and orthochromatic erythroblasts), among others. Macrophages and plasma cells were excluded due to the limited number of examples. The images were obtained by scanning whole slides at ultra-high resolution and then programmatically extracting cell-containing regions. Annotations were performed with expert guidance using the LabelMe tool [18]. This dataset is available on Figshare (https://doi.org/10.6084/m9.figshare.23805324.v1) and represents one of the first publicly available BM aspirate detection datasets. It has enabled the training of object-detection networks (e.g., YOLOv7-CTA), which have achieved ~ 88.6% mAP in detecting 15 types of BM cells [17], a significant step toward automated differential counts.

BMCD-FGCD dataset (bone marrow cell dataset-fine-grained cell dataset)

The BMCD-FGCD is a recent large-scale dataset developed in collaboration with Zhejiang Provincial Hospital in China [19]. It contains 92,335 BM cell images, each labeled into one of nearly 40 distinct cell categories. This fine-grained taxonomy encompasses all major white blood cell lineages and their corresponding maturation stages, as well as nucleated red cell precursors and less common cell types, offering

a detailed classification scheme. Images were acquired from BM aspirate smears prepared and stained in the Department of Hematology at Zhejiang Hospital using Olympus microscopes under high magnification. The dataset underwent rigorous quality control, where low-quality images were filtered or standardized, and three senior hematologists (with more than 20 years of experience) manually verified and annotated each cell's class [19]. The developers split the data into a training set (73,877 images) and a test set (18,458 images) for benchmarking algorithms. The dataset is currently publicly available to researchers via https://drive.google.com/file/d/1NrbK-OZCgTiFhWqeboaq8OVFXPNONegv/view?usp=sharing. It was used to evaluate a new "Kansformer" model for fine-grained classification, demonstrating that modern networks can handle high intra-class similarity and inter-class imbalance when provided with sufficient data [19]. BMCD-FGCD has the potential to serve as an invaluable benchmark for fine-grained BM cell classification.

BoMBR dataset (bone marrow biopsy reticulin)

BoMBR is an annotated dataset of BM *biopsy* images aimed at automated reticulin fiber quantification (a key feature for grading myelofibrosis). Raina et al. [20] presented that BoMBR includes 201 high-resolution images of BM trephine biopsy sections specially stained for reticulin fibers. Each image has been meticulously labeled for semantic segmentation, with a focus on delineating reticulin fiber networks. In addition to reticulin, the annotations mark other tissue components, notably bony trabeculae, adipocytes (fat spaces), and hematopoietic cellular areas, so that the entire image is segmented into meaningful classes. Two experienced hematopathologists performed the segmentation, starting from preliminary machine-generated masks that they refined to ensure accurate fiber borders. BoMBR is the first publicly available dataset for reticulin fiber segmentation in marrow biopsies. Using this dataset, the authors trained a multi-class U-Net model that achieved a mean Dice score of ~ 0.92 in identifying reticulin, and they could automatically predict marrow fibrosis grades with promising accuracy. The study provides a preprocessing code repository, filling an important gap for computational pathology in myelofibrosis. It enables researchers to develop and compare histological fiber detection and quantification algorithms, a task previously limited by a lack of data.

BaMBo dataset (bone marrow biopsy cellularity)

The BaMBo dataset, introduced by Singh et al. [21], is a curated collection of 185 BM core biopsy images with annotations targeting cellularity assessment. Each image is a high-resolution whole slide or large field from an H&E-stained marrow biopsy, and each has been expertly annotated with semantic segmentations for four tissue components: (1) bony trabeculae, (2) adipose (fat) marrow, (3) hematopoietic cellular regions, and (4) background. These labels enable precise computation of the BM cellularity (the ratio of hematopoietic cells to fat), an important diagnostic parameter. Two experienced hematopathologists created the annotations, leveraging DL assistance to pre-segment images and correcting them for accuracy. The BaMBo dataset provides ground truth for marrow cellularity in a far more objective and reproducible way than manual visual estimates. A U-Net model trained on BaMBo achieved an average Dice score of 0.83 across classes and predicted cellularity with an error of 6% compared to experts. This dataset addresses the lack of standardization in biopsy readings by offering a reference for algorithm development. While currently in preprint, the authors have made their code publicly available and intend to make the dataset available (https://doi.org/10.1101/2024.10.02.616393). BaMBo represents a step toward computationally derived, quantitative BM biopsy metrics.

High-resolution large-scale white blood cell dataset

Bodzas et al. [22] published a comprehensive dataset of normal and pathological white blood cells, which partly overlaps with BM cytology (particularly for blasts). This dataset comprises 16,027 annotated white blood cell images covering nine cell classes. These classes encompass all major leukocyte types, including neutrophils (segmented and band forms), eosinophils, basophils, lymphocytes, monocytes, as well as nucleated red blood cells (normoblasts) and pathological blasts from both the myeloid and lymphoid lineages. The images were obtained from 78 patient samples, including 33 patients with acute myeloid or lymphoblastic leukemia and 45 without malignant findings. All smears were stained with May–Grünwald–Giemsa and imaged under a 100 × oil objective using a high-end microscope camera, yielding an excellent resolution of ~ 42 pixels/µm. Blast cells in this dataset were cross-confirmed by flow cytometry immunophenotyping to ensure accurate labeling of AML vs. ALL blasts. The dataset is open access (https://doi.org/10.1038/s41597-023-02378-7) and hosted via a Springer Nature repository [23]. It provides one of the highest-quality collections of single-cell images for blood/BM cytology and has been used to train and test algorithms for leukocyte recognition. By including clinically significant rare cells (e.g., blasts) in a large, labeled set, this resource helps address both high-resolution image analysis and class imbalance issues in hematology.

Morphogo system dataset

Morphogo is a commercial AI-driven BM analysis platform, rather than a freely downloadable dataset; however, its development is backed by an enormous dataset worth noting. As described by Lv et al. [24], Morphogo's CNN models were trained on over 2.8 million BM nucleated cell images on an unprecedented scale. These images were derived from digitized BM aspirate slides processed by an extensive diagnostics network (KingMed Diagnostics in China), capturing routine clinical diversity. The system recognizes 25 + distinct cell categories spanning all major lineages and maturation stages, including myeloblasts, promyelocytes through neutrophils (segmented), eosinophil and basophil series, monocytes and their precursors, lymphocytes, plasma cells (at different maturation stages), and even smudge cells. All cells were annotated under expert supervision, following the WHO morphology criteria. The slides were scanned in a two-step process: low-power (40 ×) whole-slide mapping and high-power (100 × oil) targeted imaging of cells. Although this dataset of millions of labeled cells is not publicly accessible, the authors report that Morphogo achieved an overall accuracy of ~ 99% in classifying cells into 25 types with high specificity. The system's performance was validated on 508 independent BM cases (385,207 cells) across various disease categories. Morphogo's success underscores what is possible with extensive data, highlighting the gap between proprietary data volumes and public datasets. The existence of this 2.8 million-image dataset (annotated with expert-level precision) provides a benchmark for the community, even if researchers outside the company cannot directly use it. It also demonstrates the practicality of AI-assisted workflow: Morphogo integrates automated scanning, classification, and expert review in a clinical setting.

FlowCyt dataset (flow cytometry benchmark)

The FlowCyt dataset Bini et al. [25] is a multimodal benchmark that, while not image-based, is relevant for DL on BM cells. It consists of flow cytometry data from 30 BM aspirate samples, with each cell described by 12 colorimetric marker values. Every cell in the dataset has a ground-truth label identifying it as one of five key hematopoietic cell types: T-lymphocytes, B-lymphocytes, monocytes, mast cells, or hematopoietic stem/progenitor cells (HSPCs). The dataset encompasses millions of single-cell events (up to 1 million cells per patient) spanning healthy and abnormal BM immunophenotypes. The labels were obtained by expert gating and immunophenotypic definitions, effectively providing a reference standard for each cell. FlowCyt is notable as the first public flow cytometry classification benchmark in hematology, allowing the direct application of DL to cytometry profiles. The dataset is freely available via the authors' repository. While it does not include cell images, it addresses the same problem of multi-class discrimination in BM populations and can complement morphology datasets. For example, one can train models that integrate cytological images with corresponding flow cytometric data for more robust classification. FlowCyt has been used to evaluate various machine learning approaches (from XGBoost to Graph Neural Networks) on single-cell data, and it supports the exploration of cell population dynamics in a way pure image datasets do not. As such, it broadens the horizon of "dataset" in BM analysis beyond microscopy, emphasizing the potential of multimodal learning.

Dataset challenges

Data scarcity and limited public datasets

DL models require large, diverse, and well-annotated datasets, but BM cytomorphology data are notoriously scarce. Until recently, only a few public BM morphology datasets existed, each far smaller than needed for robust training. Strict privacy regulations and the specialized nature of BM exams hinder data sharing across centers [19]. Unlike radiology, which has established archives, hematopathology lacks extensive digitized image repositories [26]. This dearth of big data has meant that BM cell classification is still often done "manually… thousands of times every day" in clinical practice [16]. In short, obtaining high-quality, labeled BM images for AI is a fundamental bottleneck.

Class imbalance in rare cell types

BM smear differentials are inherently imbalanced. Common mature cells (e.g., neutrophils or normoblasts) vastly outnumber rare but clinically critical cells, such as blasts or plasma cells. In a typical dataset, some cell categories may have orders of magnitude more examples than others. For instance, one study's database reveals "large differences in the number of different types of cell targets," with specific progenitors and reactive cells being significantly underrepresented [17]. This class imbalance complicates model training—networks can become biased toward abundant cell types and struggle to recognize rare cells that might indicate disease. Ensuring sufficient samples of all categories (e.g., leukemic blasts, mast cells) or applying rebalancing techniques is essential to avoid overlooking diagnostically important minority classes.

Inter-center variability

Significant variability arises from differences in sample preparation and imaging across institutions. BM aspirate smears can vary in staining method (Wright-Giemsa vs. May–Grünwald–Giemsa), reagent quality, smear thickness, and artifact presence. Likewise, imaging hardware and settings differ; a slide scanned on one system may have a distinct color tone or focus compared to another center's microscope. These inconsistencies lead to domain shifts in the image data [27]. A model trained on one lab's slides may falter on another due to "image discrepancies due to diverse collection and staining processes" [28].

Furthermore, institutional protocols in counting or class definitions might not align. Such inter-center variability means that models need careful normalization, calibration, or training on multi-center data to generalize well. Addressing this challenge is difficult because sharing patient data between centers is restricted; however, emerging solutions like federated learning (see conclusion) aim to mitigate this limitation.

Annotation challenges and expertise requirements

Creating ground truth labels for BM cells requires expert hematopathologists, as subtle distinctions (e.g., a myeloblast vs. a promyelocyte) necessitate extensive training. This dependency on specialized experts makes annotation slow and costly. Preparing a single large dataset can entail tens of thousands of manual annotations, a labor-intensive process prone to fatigue [2, 29]. Additionally, human annotations are subject to inherent variability, as different pathologists may disagree on the classification of borderline cells or dysplastic changes. Studies note that manual microscopy reviews are subjective and can yield "inconsistencies and discrepancies among hematologists" [27, 28]. Inter-observer variability means that even "ground truth" labels in morphology may have noise or require consensus from multiple experts. Finally, because annotation is time-consuming and expensive, public datasets often have limited label granularity (e.g., only cell class labels without detailed subtypes or segmentation). They may lag behind the evolving standards of classification. All these factors, scarcity of experts, time/resource constraints, and variability, make compiling large, high-quality annotated BM datasets a formidable challenge.

Limitations of This Review: Our review focuses on English-language publications and publicly documented datasets. Proprietary datasets (e.g., Morphogo's 2.8 million images [24]) are described based on published reports but could not be independently evaluated. Additionally, rapidly evolving commercial platforms may possess capabilities that are not yet reflected in the peer-reviewed literature.

Emerging strategies to overcome dataset challenges

In response to the challenges outlined, researchers are pursuing several strategies to improve data availability and quality for DL in BM morphology:

- **Data Augmentation and Synthetic Data Generation:** To bolster limited datasets, conventional augmentations (rotations, color jitter) are routinely applied. More recently, generative adversarial networks (GANs) [30] have been utilized to generate realistic synthetic BM cell images that can augment or even replace real samples. For example, a StyleGAN2 model was able to generate convincing artificial blast cell images, thereby expanding training sets for leukemia detection [26]. Such synthetic data helps address class imbalance and scarcity, provided it is nearly indistinguishable from real data (validated via expert "Turing tests") [26].
- **Federated Learning for Multi-Center Data:** Federated learning (FL) allows training a shared model on data from multiple hospitals without centralizing the data [31]. In an FL setup, each center's data remains on-site; only model updates are aggregated. This privacy-preserving paradigm directly tackles the legal and ethical barriers to pooling patient data. By leveraging FL, one can obtain a model that is effectively trained on a large, multi-institutional dataset, thereby improving generalizability to diverse staining and imaging styles [32]. Early studies in digital pathology have shown that FL can achieve performance close to traditional pooled training, and ongoing efforts are extending this to hematopathology. Federated approaches will enable the inclusion of rare disease cases scattered across institutions, helping to mitigate batch effects while complying with data privacy regulations.
- **Collaborative Annotation Platforms:** To alleviate the annotation burden, collaborative web-based platforms and annotation drives are emerging. These enable multiple experts (from different centers) to contribute labels in a shared online environment, often within the context of grand challenge competitions. Such platforms can incorporate consensus-building tools (to handle inter-observer variability) and possibly active learning, where an algorithm suggests the most informative fields for experts to label. By distributing the work across multiple hematopathologists and automating certain parts of the process, the community can assemble high-quality, labeled datasets more efficiently. For example, several recent datasets (like the MLL and BMCD-FGCD sets) were the product of considerable team efforts rather than a single-lab endeavor [19]. Continued development of user-friendly annotation software, standardized labeling

protocols, and incentive structures (e.g., co-authorship or data-sharing agreements) will encourage multi-institution participation in dataset construction.

- **Standardized Benchmarks and Data Sharing Initiatives:** There is a growing call for standardization in hematology image datasets [21]. Just as the ImageNet and COCO datasets drove progress in general computer vision, the hematopathology field would benefit from agreed-upon benchmark tasks and reference datasets. Recognizing this, recent works have released their data and defined evaluation protocols (train/test splits, metrics) to facilitate fair comparisons of algorithms [19]. Efforts by organizations (e.g., the International Council for Standardization in Hematology) to digitize and share reference slides are also underway. The inclusion of datasets in public archives (such as TCIA or Dryad) with DOIs and documentation is increasing [15]. Such standardized, high-quality datasets accompanied by consensus annotations and perhaps periodic challenge evaluations are crucial for objectively benchmarking new DL models in BM morphology. Over time, these community datasets will help drive algorithm performance to clinical-grade levels, ensuring that researchers tackle clinically relevant problems with comparable assumptions and confidence.

While data-related challenges in BM DL are significant, combining new datasets and innovative strategies steadily closes the gap. With continued collaboration between clinicians and AI researchers, the field is moving toward robust, generalizable models that can assist in hematopathology diagnostics built on a foundation of ever-improving datasets [33, 34].

Segmentation techniques

Importance of segmentation in marrow image analysis

Accurate image segmentation is a foundational step in BM morphology analysis for smears (aspirate slides) and biopsies. By delineating cell and tissue structures from the background, segmentation enables the reliable extraction of features for downstream tasks such as cell classification and quantification [12, 35]. In dense BM samples, precise segmentation is critical, and errors propagate to cell identification and disease diagnosis at this stage [13]. For example, an automated U-Net segmentation of nucleated cells in whole-slide marrow smears achieved a 71% Dice coefficient (58% IoU (Intersection over Union)) on the validation set [36], allowing over 500,000 single-cell images to be cropped for lineage classification in an MDS study. This demonstrates how good segmentation directly facilitates large-scale single-cell analysis. In biopsy histology, the segmentation of tissue compartments (e.g., marrow cells, adipose tissue, and bone) provides objective measures, such as cellularity, which is clinically crucial for diagnosing conditions like aplastic anemia or myelofibrosis. Automated tools now utilize segmentation to estimate marrow cellularity with accuracy within a few percentage points of expert assessments [37]. In short, segmentation "significantly influences subsequent processes" in marrow image analytics [35], underpinning the accuracy of classification, counting, and diagnosis.

Classical image processing segmentation methods

Traditional computer vision techniques have been applied to BM smears and histology for decades, with mixed success. Thresholding methods (e.g., Otsu's method) were employed to separate cells or nuclei based on their intensity [38]. Su et al. [12] combined adaptive thresholding and region growing to segment BM nuclei and cytoplasm, utilizing color deconvolution and K-means clustering to refine the cytoplasm masks. Such approaches can segment many cells in clean areas but struggle when stain intensity varies or cells overlap. Marker-controlled watershed algorithms have been widely employed to split dense cell clusters in aspirate smears [39]. The watershed transform treats the image as a topographic surface and can delineate touching cells by finding "catchment basins" for each cell. Osowski et al. [40] pioneered a watershed-based pipeline for segmenting myeloblasts in marrow smears, achieving the separation of tightly clustered blast cells (appearing as distinct dark regions) with only minor boundary errors at the cytoplasmic level. Mathematical morphology (e.g., iterative erosion/dilation) has been used to enhance the separation of adhered cells [35, 41]. Arslan et al. [42] developed a color- and shape-based algorithm that combines morphological operations in the HSV color space to segment white blood cells in marrow and blood images. Their method achieved high accuracy in identifying isolated leukocytes and merged cell clusters by integrating color and geometric cues. Similarly, Theera-Umpon [43] applied fuzzy clustering and edge detection to segment whole cells and their nuclei in marrow smears, an early attempt that yielded "commendable" accuracy for its time. Active contour models (also known as snakes) have been adapted for use in marrow histology images. For instance, Song et al. [44] proposed a dual-channel active contour to delineate megakaryocyte nuclei and cytoplasm in BM trephine biopsy slides. By leveraging separate stain color channels (e.g., H&E channels) and an initial machine-learned nuclei detection, their method accurately traced cell boundaries even in complex tissue

backgrounds. Overall, classical techniques can be effective in controlled scenarios. Watershed and morphological filters excel at separating touching cells, and active contours can precisely refine boundaries. However, they require careful parameter tuning and often falter with the challenges of real-world marrow images (variable staining, dense overlaps, artifacts).

Deep learning–based segmentation approaches

Modern DL has dramatically improved BM image segmentation in recent years. CNN models (specialized algorithms that automatically learn image features through training) learn rich features that overcome the limitations of manual thresholds or simple filters. The dominant paradigm for biomedical segmentation is the encoder-decoder CNN [45]. U-Net, a Fully Convolutional Network (FCN) architecture, has been widely adopted for marrow cell and tissue segmentation [46, 47]. Even with modest training datasets, the U-Net can learn to "delineate the boundaries of nucleated cells and segment the cell area of interest from the background" in marrow smears. In one study, a U-Net detected and segmented nucleated cells in BM WSI with a 71% Dice (58% IoU), as noted, which was sufficient to drive an automated dysplasia classification pipeline [36]. U-Net variants have been tailored to marrow data, for example. Gated U-Net (GCT-Unet) introduces channel-attention mechanisms to better handle "dense numbers and high adhesion" of marrow cells.

Qin et al. [48] demonstrated that incorporating gated channel transformers enhanced the U-Net's performance on tightly clustered BM cells, particularly when training data are limited. Deep models can also segment multiple classes simultaneously; for example, a custom U-Net trained on the BaMBo biopsy dataset (185 H&E images) can segment four tissue classes (trabecular bone, fat, hematopoietic cells, and background) with an average Dice score of 0.83 [21]. This multi-class segmentation enables the automated computation of marrow cellularity with an error rate of only 5.9% compared to pathologists. Beyond U-Net, researchers have explored other CNN architectures, including SegNet (another FCN) and GAN-based models. Zhang et al. [49] introduced an adversarial model (AMLcGAN) for myeloblast segmentation in AML smears, which achieved a mean Dice score of 82.5%, outperforming the standard U-Net (77.4%) and SegNet (78.2%) on the same task. Notably, combining GAN's feature learning with segmentation improved the precision and recall of identifying blast cell regions. For instance, segmentation (differentiating each cell) and object-detection-based models, such as Mask R-CNN, have been applied in analogous cytology domains. They can be used for BM smears: Mask R-CNN can localize individual cells and output a binary mask for each cell. While specific Mask R-CNN results in the marrow are scarce in the literature, detection-driven approaches are emerging.

One study utilized the YOLOv3 detector to rapidly identify nucleated cells in BM smears, highlighting its real-time speed but noting slightly lower precision compared to segmentation models [50]. Researchers sometimes combine deep networks with classical post-processing, for example, using CNN probability maps to get the best of both worlds

Table 2 Summary of key studies utilizing DL architectures for BM image analysis segmentation tasks

Study/Model	Data type	Architecture	Task	Dataset/Size	Performance (Dice/IoU/Accuracy)	Notable strengths
Qin et al. (2023) [48]	Smear cells (aspirate)	Gated U-Net (GCT-Unet)	Single-cell segmentation	Private BM aspirate images (~N/A)	Dice ~ 0.83 (dense clusters)	Channel-attention modules for adhesion handling
Zhang et al. (2023) [49]	Smear cells (AML)	AMLcGAN (Adversarial GAN)	Myeloblast segmentation	Private dataset (N/A)	Dice 82.5% (vs U-Net 77.4%)	Improved fine boundary delineation
Lee et al. (2022) [36]	WSI Marrow Aspirates	U-Net	Nucleated cell segmentation	Private WSI (~500,000 cells)	Dice 71%, IoU 58%	Enabled large-scale dysplasia classification
Singh et al. (2024) [21]	Marrow Biopsy (H&E)	U-Net (multi-class)	Tissue compartment segmentation	BaMBo dataset (185 images)	Dice ~ 0.83 overall	Accurate cellularity measurement (<6% error)
Raina et al. (2024) [20]	Reticulin Biopsy Images	Multi-class U-Net	Reticulin fiber segmentation	BoMBR dataset (201 images)	Dice ~ 0.92 (fibers)	First automated fibrosis grading benchmark
Fu et al. (2024) [53]	Biopsy (CD138 stain)	Segmentation CNN	Plasma cell detection and counting	Private biopsy slides	ICC > 0.97 with pathologists	Reliable plasma cell quantification

This table outlines the data type, network architecture, specific segmentation task, dataset size, reported performance metrics (e.g., Dice coefficient, IoU, accuracy), and the distinctive strengths of each model. The studies span applications from single-cell segmentation in aspirate smears to tissue and fiber segmentation in biopsies, highlighting the adaptability of convolutional and adversarial networks in various hematopathological contexts

as markers for watershed segmentation. The latest trend is the incorporation of Vision Transformers into segmentation networks. Transformer-based models leverage global self-attention, which can be advantageous for complex histology images where long-range context (e.g., spatial arrangement of cells and stromal elements) matters. Early examples in pathology include transformer-enhanced U-Nets, which improved the segmentation of tissues such as kidney glomeruli [51]. In BM histology, we are beginning to see analogous approaches: for instance, a Swin Transformer U-Net backbone was recently used to quantify plasma cells in CD138-stained marrow biopsies [52]. While pure transformer architectures for marrow segmentation are not yet common, the expectation is that they will help capture subtle contextual cues in crowded marrow images that CNNs might miss. DL approaches have demonstrated improved robustness to staining variation and cell heterogeneity compared to classical methods [42], though challenges remain with extreme variations. They learn features that make segmentation "unaffected by variations in background, staining techniques, and morphological disparities" to a far greater degree than manual methods [42]. As a result, deep models routinely achieve higher overlap scores (Dice, IoU) on benchmark datasets than classical algorithms, and they generalize better to new cases when trained on extensive data. Table 2 summarizes key studies utilizing DL architectures for marrow segmentation.

Challenges in segmenting bone marrow images

Despite advances, BM image segmentation faces unique challenges due to the complex nature of marrow preparations:

- **Dense Cell Clusters & Overlapping Cells:** Marrow aspirate smears exhibit a significantly higher cell density than peripheral blood, characterized by frequent cell overlap and clustering [13]. This makes separating individual cell boundaries difficult. Classical algorithms often over-segment (splitting one cell into pieces) or under-segment (merging adjacent cells). Watershed segmentation helps split touching cells [39] but can produce spurious boundaries if not carefully initialized. Deep networks can implicitly learn to separate crowded cells, but even they may confuse tightly adherent cells as one object. Some works address this issue by incorporating shape priors or multi-step processing, for example, by identifying nuclei first as seed points and then expanding to whole-cell regions [54]. Instance segmentation models (Mask R-CNN, Cellpose) explicitly aim to solve overlaps by predicting distinct masks, but they require extensive annotated examples of overlapping configurations. Handling dense clusters remains a key benchmark for any segmentation technique's robustness.
- **Smear Artifacts and Debris:** BM aspirates often contain staining artifacts, broken cells (smudge cells), platelet clumps, and other debris. These can confound segmentation algorithms. A stray ink or stain blob might be detected as a "false cell." Traditional filters can remove some noise (e.g., small objects or out-of-range colors), but artifacts are diverse. DL can be more resilient if trained on enough augmented data—the model can learn to classify debris as background. For instance, a recent robust segmentation method explicitly modeled red blood cell regions and background noise, then excluded them before segmenting the white cells [55]. This two-stage approach (artifact removal followed by cell segmentation) improved accuracy in mixed blood/marrow smear images. Still, artifact-heavy fields sometimes need manual cleanup or specialized preprocessing (like thresholding out slide pen marks).
- **Background Heterogeneity in Aspirate Smears:** Unlike cultured cell images, marrow smears have a complex background matrix. Erythrocytes often form a red "haze" across the slide; cytoplasm from burst cells can create a pale, granular background, and variations in slide staining or scanning can alter the color balance across an image. This heterogeneity can break simple thresholding; what works in one region may fail in another. Color-space-based segmentation (e.g., using HSI or LAB color space) has been employed to more effectively distinguish nuclei/cells from the background by hue [56]. Arslan's color-shape method explicitly processed color information to remain robust against background variation [42]. Deep CNNs inherently handle some color variation via learned invariances, but they can still be fooled by unusual background appearances if these are not represented in the training data. Techniques such as color normalization (standardizing the appearance of stains) or training with diverse staining examples are employed to mitigate this issue.
- **Histology Complexity and Tissue Texture:** In core biopsies or trephine sections, the **marrow tissue** is an intricate mesh of cells, fat spaces, blood vessels, and bony trabeculae. Segmenting this at a tissue level involves differentiating regions that may have similar colors or textures. For example, adipocyte (fat cell) spaces appear as white voids in H&E. They are easily identified by their intensity, whereas distinguishing hematopoietic cell regions from connective tissue or artifacts can be more challenging. Background in biopsies may include slide artifacts (folds, tearing) that mimic tissue. Moreover, reticulin fiber networks (seen in reticulin-stained slides) are delicate, thread-like structures that

weave through the marrow, segmenting these thin fibers and pushing the resolution limits of both manual and automated methods. DL has shown promise: the new BoMBR dataset for reticulin fiber quantification enables the training of a multi-class segmentation model that distinguishes between reticulin fibers, cells, and bone. They report a high mean Dice (~0.92) for fiber segmentation [20], indicating that deep models can indeed capture these filamentous structures that classical edge detectors might miss. Nonetheless, achieving such performance required meticulous annotations and probably specialized network layers or loss functions to handle extremely thin objects. In summary, BM segmentation must cope with challenges such as cell crowding, overlap, artifacts, and variable histology—necessitating sophisticated, often hybrid, solutions.

Segmentation at different scales: single cells, tissue compartments, and subcellular structures

Single-cell segmentation (aspirate smears)

A significant focus in marrow cytology is segmenting individual nucleated cells in smear images (semantic segmentation). The goal is often to count and classify every myeloid, erythroid, or blast cell in the field. Classical pipelines typically segment nuclei (as they have high contrast) and then include surrounding cytoplasm to define each cell [12]. This two-step approach works because most diagnostic features (e.g., blast identification) rely heavily on nuclear morphology. However, overlapping cytoplasm or smeared cell boundaries make complete cell segmentation tricky. Modern CNN-based approaches can perform end-to-end semantic segmentation of all cells vs background, producing a mask of all cell regions [36]. Instance segmentation is more challenging; some studies utilize detection models to localize cells and then crop or mask them (as done in the MDS study, where U-Net masks located cells that were then cropped for classification [36]). To improve single-cell segmentation, researchers have leveraged large datasets. The Morphogo system, for example, was trained on over 2.8 million marrow cell images and can identify more than 25 cell types with 99% accuracy. Morphogo's pipeline involves detecting and isolating single cells from smear images (though details are proprietary)—highlighting that robust segmentation/generalization comes from enormous training data [24]. Another line of work utilizes multi-scale or texture-aware models, as marrow smears exhibit fine details (nuclear chromatin) and coarse context (cell size, neighborhood). Networks incorporating multi-scale feature fusion can therefore better separate touching cells of different types. For instance, a Texture-U-Net was proposed to segment BM WSI regions by combining global texture analysis with local segmentation [57]. Single-cell segmentation performance is evaluated by how well individual cell regions are delineated (Dice/IoU per cell) and its downstream impacts, such as cell counting accuracy. Deep models now achieve detection rates of over 90% for nucleated cells in many smear datasets. However, distinguishing truly overlapping cells (e.g., a cluster of erythroblasts) still sometimes requires manual correction or sophisticated instance strategies (such as iterative segmentation or contour refinement).

Tissue-level segmentation (biopsies)

In BM biopsy histology, the aim is to segment broader regions, including marrow, bone, and fat, as well as specific stromal components. This compartment segmentation is directly tied to clinical metrics, such as marrow cellularity (the percentage of the area occupied by hematopoietic cells). Traditional histomorphometry methods used point counting or manual region drawing to estimate these. Now, automated segmentation offers an objective alternative. The BaMBo dataset (BM Biopsy dataset) establishes a benchmark with expert-annotated regions for cellular marrow, adipose tissue, and bone [21]. A deep U-Net model trained on BaMBo could predict marrow cellularity with < 6% error, closely matching pathologist estimates. Similarly, MarrowQuant, a digital pathology tool, segments H&E-stained biopsy images into bone, fat (represented by adipocyte "ghosts"), hematopoietic marrow, and vessels [58]. Its algorithm (which combines color thresholding and region growing in QuPath scripts) achieved around 86% average mask accuracy, including ~92% for hematopoietic areas [59]. These high-level segmentations enable the automated computation of the fat-to-cell ratio and the detection of fibrosis or granulomas. DL further refines this: one study utilized a CNN to segment and count plasma cells in biopsy sections stained for CD138, thereby enabling precise estimates of plasma cell percentage for myeloma diagnostics [52]. Another effort by van Eekelen et al. [60] applied DL to segment and quantify cellularity and differentiate lineage compartments (e.g., erythroid vs myeloid regions) in trephine biopsies. Tissue segmentation faces challenges, such as variable histology preparation and bone spicule fragmentation (bone can scatter in the section). However, networks can be trained to recognize even partially preserved trabeculae versus background. The output of tissue-level segmentation is often evaluated by overlap with pathologist-drawn regions or by correlation with clinical measures. For instance, automated cellular area fraction correlating strongly with manual cellularity counts validates the approach [21]. As public resources grow (BaMBo provides open data for this task),

we can expect rapid improvement and standardization in tissue-level marrow segmentation.

Subcellular structure segmentation

Segmenting components within cells or the microenvironment is important in some applications. One example is the segmentation of the nucleus versus the cytoplasm in marrow cells. Certain dysplastic features (like cytoplasmic granule changes or nuclear-cytoplasmic ratio) require identifying these sub-regions. Song et al.'s active contour method explicitly produced separate masks for megakaryocyte nuclei and their cytoplasm [44]. They first used a supervised classifier to detect the nucleus, then grew an active contour from it in a different stain channel to capture the cell body. Compared to one-step methods, this two-step segmentation yielded notably better boundary accuracy for both the nucleus and cytoplasm. In digital pathology workflows, one might use a nuclear segmentation model (common in histopathology) on marrow biopsy slides to count all nuclei, then classify those nuclei by cell type; however, distinguishing overlapping nuclei in marrow (especially in cluster-forming cells like megakaryocytes or lymphoid aggregates) is non-trivial. Another subcellular target is the segmentation of reticulin fibers. BM fibrosis grading relies on assessing reticulin fibers (stained black by a reticulin silver stain). Traditionally, fibrosis is graded by visual estimation of the density of fibers. Automated segmentation can make this quantitative. The BoMBR dataset (201 reticulin-stained biopsy images) was created to advance this task. A DL model trained on BoMBR performs multi-class segmentation, identifying reticulin fibers separate from cells and other structures, achieving a high Dice score of ~ 0.92 for the fiber class. This allows the computation of fiber area percentage or length, which can be mapped to the Myelofibrosis (MF) grading scale. Indeed, the BoMBR team demonstrated that using the segmentation output to predict the MF grade yielded a weighted F1-score of ~ 0.656, showing moderate agreement with pathologists and indicating room for further improvement [20]. Other sub-components that might be segmented include iron deposits in Prussian blue stains and osteoblasts or osteoclasts lining bone trabeculae in biopsies. In summary, subcellular segmentation extends the analysis to finer granularity by separating nuclei, cytoplasm, or stromal fibers, allowing the extraction of morphological metrics (nuclear size, N: C ratio, fiber length) that inform more refined diagnostic or prognostic models.

Evaluation metrics and performance outcomes

Studies commonly report the Dice similarity coefficient (also known as the Dice score, a metric that measures how well predicted cell boundaries match expert annotations) and the Intersection over Union (IoU) for segmented regions to quantify segmentation accuracy. These overlap metrics compare the model's predicted mask with ground-truth annotations (using the Dice score, which is defined as $2 \cdot TP/(2 \cdot TP + FP + FN)$). Dice scores in the range of 0.75–0.9 are now typical for DL models, as demonstrated in recent studies on BM segmentation [20, 21, 36, 48, 49]. In contrast, classical methods often achieve lower overlap or require case-by-case tuning. For example, the U-Net model for cell detection in MDS smears achieved Dice ≈0.71 and IoU ≈0.58 on validation—decent given the difficulty (segmenting varied nucleated cells on a noisy background) [36]. On the biopsy side, the BaMBo U-Net achieved a Dice score of ~ 0.83 across four classes [21], correctly labeling most pixels in each compartment. An adversarial CNN for blast segmentation (AMLcGAN) achieved a Dice score of approximately 0.825, outperforming conventional networks by around 5 percentage points [49]. Such improvements in overlap translate to more accurate counts and measurements. For example, a Dice gain from 0.77 to 0.82 for blast cell masks led to better precision in automatic blast counting, which is crucial for AML diagnostics. Beyond overlap metrics, studies also utilize Precision/Recall (especially in scenarios such as segmentation or object detection, where ensuring that most cells are found with few false alarms is crucial), and Pixel Accuracy (the overall fraction of correctly classified pixels).

Correlation measures are used for segmentation that feeds into clinical metrics: the automated cellularity from BaMBo's model had an intraclass correlation (ICC) of 0.78 with pathologist estimates [60], indicating good agreement. Likewise, in reticulin segmentation, the automated fiber quantification can be validated by how well it reproduces the fibrosis grade that an expert would assign. High segmentation performance also yields tangible gains in classification and diagnosis pipelines [61]. By isolating individual cells, one can apply specialized classifiers (such as CNNs per cell type) with higher accuracy than if the cells were unsegmented in a cluttered image [36]. Eckardt et al. [9] demonstrated that an automated segmentation and classification system can flag dysplastic cells in marrow smears, aiding in MDS diagnosis and reducing review time. In cell counting tasks, good segmentation ensures that automated differential counts (the proportions of neutrophils, erythroid precursors, blasts) closely match manual counts, which is essential for reliable diagnoses. Ultimately, the practical impact of improved segmentation is seen in more standardized and reproducible morphologic assessments. Systems like Morphogo, which rely on CNN-based segmentation and identification, have demonstrated the ability to match expert-level identification across dozens of cell classes [24].

This consistency at scale (Morphogo examined thousands of cells per slide with 99% accuracy) can shorten diagnostic turnaround and help eliminate observer variability [24, 62]. Similarly, automated cellularity measurements provide quantitative tracking of marrow recovery or aplasia, which can guide treatment decisions. In fibrosis assessment, image segmentation enables quantitative grading (e.g., fiber area percentage), which can be more sensitive to change than coarse manual grades [20]. As these techniques mature, we expect DL segmentation to become an integral part of BM pathology workflows, powering everything from computer-aided differential counts to objective tissue assessments and ultimately improving the precision of hematologic diagnoses.

Classification techniques

Traditional vs. deep learning approaches

Early BM cell classification efforts relied on classical machine learning algorithms (e.g., k-NN, SVM, decision trees) applied to handcrafted features extracted from microscopy images. These traditional pipelines required explicit preprocessing, cell segmentation, and feature engineering (such as nuclear shape descriptors or color histograms) before classification [63, 64]. Such methods were time-consuming, operator-dependent, and often struggled with inter-observer variability and inconsistent feature definitions. In contrast, modern DL, particularly CNNs, learns discriminative features directly from raw images, eliminating manual feature extraction [8]. CNN-based models have demonstrated substantially higher accuracy and robustness than traditional image analysis techniques for BM cytomorphology [65]. These advances, coupled with the ability of deep models to generalize across variations in staining and imaging, have established DL as the preferred approach for BM cell classification tasks.

CNN-based classification of bone marrow cells

Single-cell image classification

The most common application of DL in marrow morphology is classifying individual nucleated cells into their respective lineages or malignancy categories. A microscope slide is digitized, and individual cell images (segmented cell crops or patches) are fed to a CNN classifier. Standard architectures, such as VGG, ResNet, DenseNet, and EfficientNet (often pre-trained on ImageNet), have been fine-tuned for this task [66]. For example, Matek et al. achieved human-level performance in recognizing blast cells of AML using a CNN, matching the accuracy of expert hematologists in identifying these malignant, immature cells [67]. In a follow-up study with a large dataset of BM cells spanning 21 morphological classes, their ResNeXt-50 CNN model attained high overall accuracy in differentiating cell types; however, it was noted that performance on rare cell classes like myeloblasts was lower (e.g., ~75% precision and 65% recall for blast identification) [16]. This highlights that even with DL, morphologically similar cells can be confused, underscoring the need for ample training data and specialized models for challenging classes. Recent studies leveraging large-cell image datasets (tens of thousands to millions of single-cell images) have improved single-cell classification performance. For instance, a deep CNN trained on 41,595 BM cell images from healthy individuals (covering various normal leukocyte classes) achieved a mean precision and recall of about 0.89 for each cell type, with an area under the ROC curve (AUC) of 0.99 [68]. Such high accuracy approaches pathologist-level differential counts. DL has also been applied to detect malignant cells in marrow or blood smears, for example, distinguishing ALL (acute lymphoblastic leukemia) blast cells from normal lymphocytes. Shafique and Tehsin showed that a pre-trained CNN could accurately detect ALL and even classify its subtypes, significantly outperforming earlier rule-based methods [69]. Likewise, Wu et al. developed a BMSNet model that achieved hematologist-level accuracy in classifying single-nucleated BM cells as normal or leukemic, demonstrating the feasibility of automated expert-like cell recognition [13].

Some researchers have explored hierarchical classification reflecting the hematopoietic lineage tree, in which a first CNN separates broad categories (e.g., myeloid vs. lymphoid), and subsequent models classify finer subtypes. While intuitive, such multi-step classifiers have not shown decisive gains over one-step models. Modern CNNs can often learn the hierarchy implicitly, given enough data [70]. Nonetheless, hierarchical or cascaded models can be helpful when certain classes are very similar; a coarse-to-fine approach can reduce misclassification by focusing the model on smaller decision spaces at each step. Ensemble strategies have also been employed at the single-cell level to enhance robustness, for example, by combining the predictions of multiple CNNs trained with different architectures or input preprocessing methods. By averaging or voting on predictions, ensembles reduce variance and improve reliability by a few percentage points, which is valuable for critical distinctions, such as blast detection [71]. Matek et al. [16] effectively employed a dual-CNN approach, where one network focused on cell morphology classification and another on genetic feature prediction, to improve overall system performance. This illustrates how multiple models can complement each other in classification tasks.

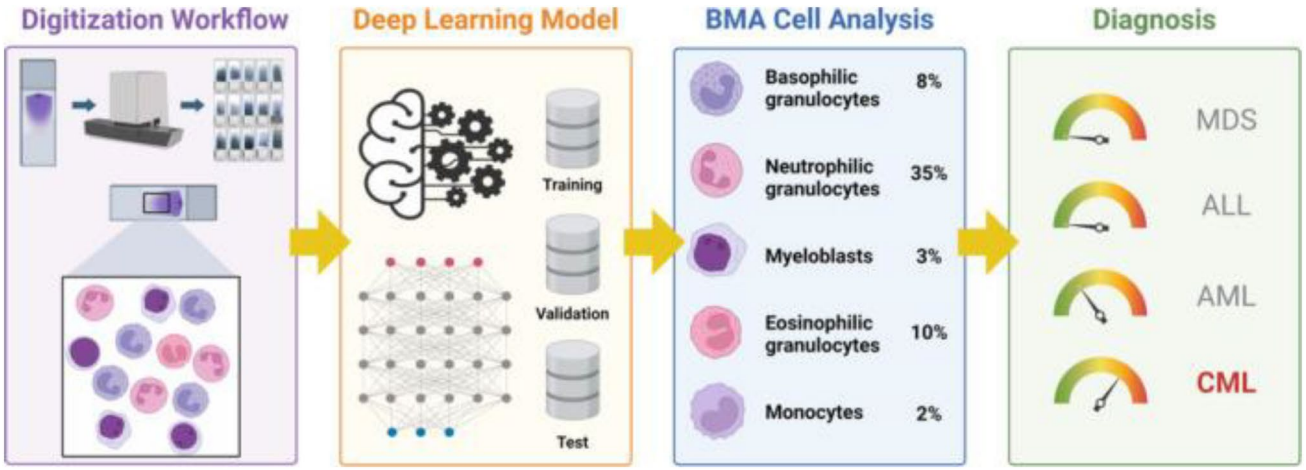

Fig. 1 End-to-end DL workflow for automated BM aspirate (BMA) analysis. The digitization workflow captures high-resolution images of stained BM smears, which are then processed by a DL model trained on annotated cell images. The model performs cell type classification (e.g., basophilic granulocytes, myeloblasts) and quantifies their proportions. These quantitative outputs contribute to diagnostic predictions for hematologic diseases such as MDS, ALL, AML, and CML, with the system flagging the most likely diagnosis based on cellular composition

Slide-level classification and diagnosis

Beyond single-cell analysis, DL is leveraged to classify whole BM aspirate smears or biopsy slides to assist in diagnoses (e.g., identifying leukemia or myelodysplastic syndrome from an entire slide). Whole-slide image (WSI) analysis is challenging due to the complex mixture of numerous cells, varied regions (e.g., fatty vs. cellular areas), and artifacts. A common approach is an end-to-end pipeline in which the WSI is first processed to detect or crop individual cells or regions of interest. Then, those are classified and aggregated to produce a slide-level diagnosis [44, 72]. Figure 1 illustrates an example of such a pipeline, where a CNN analyzes a digitized marrow smear to yield a differential count of cell types, informing the final diagnosis.

Early DL systems for slide-level analysis employed a sequential two-step approach: one CNN (or a classical algorithm) to detect/segment cells, and another CNN to classify each cell. Chandradevan et al. [73] implemented such a system for automated differential counts on BM aspirates, focusing on non-neoplastic cells, and reported high accuracy in classifying normal cell types when evaluated on initial datasets. However, running two separate networks can be computationally intensive and may propagate errors (e.g., a missed cell cannot be classified). Researchers have designed integrated models that perform both detection and classification within a single network to address this issue. Song et al. [44] proposed a unified DL model that simultaneously localizes and classifies cells in BM histology images, using an autoencoder-based architecture with a novel spatial probability prior to better handling overlapping or irregularly shaped cells. This end-to-end model proved more efficient and accurate than the traditional cascade of two CNNs, demonstrating the advantage of joint learning. For instance, it improved overall classification accuracy and speed by avoiding duplicate computations for feature extraction. Another noteworthy approach is the use of multiple instance learning (MIL) and attention mechanisms for slide-level diagnosis. Rather than classifying every cell, MIL-based models treat the slide as a bag of many cell instances and learn to predict the patient-level label (e.g., “AML” vs. “healthy”) directly from the bag, often highlighting which instances (cells) are most indicative of the outcome.

An example is the MIL framework MILLIE (Multiple Instance Learning for Leukocyte Identification), which was able to distinguish APL (acute promyelocytic leukemia, a subtype of AML) from other AML subtypes with an AUC of 0.99 by automatically attending to promyelocytes with Auer rods—a key diagnostic feature [74]. Overall, slide-level CNN models have shown impressive performance. Wang et al. [72] developed a hierarchical DL system that can process an entire BM aspirate WSI in seconds, detecting all nucleated cells and classifying them to compute the marrow differential counts. Their framework achieved over 90% accuracy in classifying cell lineages and very high recall, outperforming earlier methods that analyzed smaller fields of view. In terms of clinical diagnoses, DL models can classify marrow slides into disease categories such as AML, ALL, MDS, or aplastic anemia. Zhou et al. [10] presented a system for multi-disease classification that, in a real-world clinical test, correctly identified cases of MDS, ALL, AML, and chronic myeloid leukemia (CML) with an overall accuracy of around 94%. Similarly, Wang et al. [75] demonstrated accurate differentiation of aplastic anemia vs.

MDS vs. AML using DL on marrow smears. A specialized model by Eckardt et al. [9] focused on detecting APL blasts in the BM and achieved near-perfect specificity, offering a rapid tool to flag this medical emergency. These examples underscore that DL classifiers can operate at the whole-slide level by analyzing cells individually and aggregating results or directly learning slide-level patterns. In practice, many systems combine approaches, such as using a CNN detection model (like YOLO or Faster R-CNN) to identify cell locations and a classification model (ResNet/DenseNet) to label cell types [76, 77], followed by the computation of summary statistics or diagnoses from the per-cell predictions. Some commercial platforms (e.g., Morphogo and others) have implemented such end-to-end solutions. One report documented an automated marrow smear analyzer that achieved greater than 95% accuracy for every major cell category and 99% specificity, matching expert manual differential counts [62, 78]. The continued development of slide-level AI aims to assist pathologists by handling tedious tasks (like counting hundreds of cells) and providing decision support for diagnoses like leukemia or MDS.

Training strategies and transfer learning

A critical challenge in DL for BM cytomorphology is the limited size of labeled datasets since annotating thousands of cells by expert morphologists is labor-intensive. Transfer learning has, therefore, become a cornerstone of training strategies. In most studies, CNN models are pre-trained on large general image datasets (such as ImageNet) and then fine-tuned on domain-specific BM images. This approach leverages learned low-level features (edges, textures, shapes) from natural images, which surprisingly transfer well to hematology imagery. Huang et al. [8] demonstrated that using ImageNet-pretrained CNNs significantly improved leukemia cell classification accuracy on a small BM dataset, making it feasible to perform well with limited samples. The fine-tuning process involves replacing and retraining only the top layers of the network (or using a reduced learning rate for the pre-trained layers) on the BM data, which speeds up convergence and often boosts metrics by several points compared to training from scratch. Recent evaluations have systematically compared training from scratch vs. different transfer learning schemes for BM cell classification; they generally conclude that initializing with pre-trained weights yields superior accuracy and faster training, especially for models like ResNet or DenseNet on datasets of only a few thousand images (typical in this field) [79, 80]. Another approach to cope with limited labels is semi-supervised learning and active learning. Nakamura et al. [81] combined self-training with active learning to enable the model to progressively label new, unlabeled cell images and selectively add them to the training set. In their workflow, CNN was first trained on a small, labeled set and then used to predict labels on a larger pool of unlabeled images. High-confidence predictions were accepted as pseudo-labels (self-training), while uncertain cases were flagged for human expert labeling (active learning), thereby efficiently expanding the training data. This strategy effectively increased the training set size without requiring an entirely manual annotation of all images, improving classification performance. Class imbalance in BM data (e.g., few blast cells but many neutrophils) is another training concern; researchers address this issue with techniques such as data augmentation, class-weighted loss functions, or oversampling of rare cell images to ensure the CNN learns minor classes adequately. In hierarchical models, staged training can be employed: for example, first train a model to separate broad classes, then train specialized models for sub-classes. The weights of each stage can be initialized from those of the previous stage to retain shared feature representations. However, one must avoid error propagation across stages and validate the end-to-end performance on the final targets.

Ensemble training is also explored to maximize accuracy: multiple CNN models are combined with different random initializations or architectures. For example, one might train three variant classifiers, ResNet50, DenseNet121, and EfficientNet, on the same data and then average their outputs or use a voting scheme. Such ensembles yield a more robust final classifier as uncorrelated errors are smoothed out. Zhou et al. [10] developed an ensemble model that utilized three integrated ResNet CNNs to classify 20 diverse types of BM cells. The model achieved an overall accuracy of approximately 82.9%, with an AUC of ~ 0.99 and an average F1-score of 0.829. While ensembles can improve metrics (often by a few percent in F1-score or AUC), they increase computation and complexity; therefore, a balance is struck based on the requirements. Overall, careful training protocols leveraging transfer learning, semi-supervised data augmentation, and sometimes multi-model ensembles are key to building high-performing classifiers in this domain.

Performance of recent deep learning models

DL classifiers for BM cytomorphology have reported strong performance in recent studies (2020–2025), though results vary with task complexity and dataset size. Single-cell classification models typically achieve high accuracy and precision. For instance, Alsharani et al. [82] achieved over 97% accuracy in classifying five types of marrow cells using a fine-tuned DenseNet121 with attention mechanisms. Likewise, multiple groups report per-class accuracies in the 90–99% range for normal leukocyte differentiation tasks [83]. In a multi-class problem with 10 + cell categories, one

large-scale study achieved an average accuracy of ~ 95% across all classes [84], a level comparable to that of experienced human microscopists. Slide-level diagnostic models also demonstrate excellent efficacy: one system evaluating entire BM aspirate smears achieved an overall classification accuracy of 94.4% in distinguishing among five diagnostic categories (AML, ALL, chronic myeloid leukemia, essential thrombocythemia, and normal/reactive marrow) [62]. Many deep models report precision and recall values in the 0.8–0.95 range. For example, a slide-level CNN for leukemia subtyping had a precision of ~ 0.83 and a recall of ~ 0.82 across classes [62]. At the same time, a cell-level classifier often shows even higher precision/recall (above 0.90) for the dominant classes [82, 83]. Reported F1-scores (the harmonic mean of precision and recall) are correspondingly high, frequently 0.85–0.95 for well-defined tasks. For critical binary decisions like "malignant vs benign," the models achieve very high discriminative ability, with AUC values of 0.95–0.99 noted. In one study, an attention-based MIL model distinguishing APL from other leukemias achieved an AUC of 0.99, effectively making zero false-negative APL cases in their test set [74]. It's important to note that performance can vary significantly by cell type or subclass. Deep models excel on abundant, distinctive cell types (e.g., segmented neutrophils or mature lymphocytes), but rare or visually confusable cells remain challenging. As mentioned, even a leading CNN struggled with reliably separating myeloblasts from other progenitors, yielding a recall of

Table 3 Overview of DL models applied to BM cell classification tasks

Study/Model	Data type	Architecture/Method	Task	Dataset/Size	Performance	Notable strengths
Wu et al. (2020) [13]	Single-cell (aspirate)	BMSNet (CNN) with YOLO	Normal vs. Leukemic cell classification	Private dataset	AUC 0.948	Clinician-level leukemic cell recognition
Matek et al. (2021) [16]	Single-cell (aspirate)	ResNeXt-50 CNN	Multi-class cell classification (21 types)	MLL dataset (171,375 cells)	Mean precision 0.51 Mean recall 0.689	Large public benchmark; human-level AML blast detection
Shafique & Tehsin (2021) [69]	Single-cell (smear)	Pretrained CNN (ResNet/DenseNet)	ALL blast classification	Public WBC dataset	Accuracy ~ 98%	Early success with transfer learning
Ananthakrishn et al. (2022) [85]	Single-cell (aspirate)	Probabilistic Siamese network with triplet loss function	Multi-class cell classification (20 types)	MLL dataset (171,375 cells)	Accuracy 0.84 F1 score 0.91	Siamese neural network for BM classification
Wang et al. (2022) [72]	Whole-slide Aspirates	Hierarchical CNN Pipeline	Cell detection, classification, slide diagnosis	Private dataset	> 90% classification accuracy	End-to-end full WSI processing
Zhou et al. (2022) [10]	Whole-slide Aspirates	3 × ResNet Ensemble	Multi-disease slide diagnosis (AML, MDS, ALL)	Private dataset	94% diagnostic accuracy	Ensemble improved consistency and rare class detection
Alsharani et al. (2023) [82]	Single-cell (aspirate)	DenseNet-121 + Attention	5-type marrow cell classification	Private dataset	> 97% accuracy	Attention improved interpretability
Goldgof et al. (2023) [68]	Single-cell (aspirate)	DeepHeme, ResNeXt-50	cell classification of 23 distinct types	41,595 single-cell images from BMA smears of 50 patients	Mean precision 0.89 Mean recall 0.89 Mean AUC 0.99	expert-level accuracy, exceptional speed, and robust generalizability
Manescu et al. (2023) [74]	Slide (no cell labels)	Attention-based MIL Framework	Leukemia subtyping and APL detection	Private dataset	AUC 0.99 for APL	No single-cell labels are needed, excellent for rare subtypes
Glüge et al. (2024) [70]	Single-cell (aspirate)	Regnet_y_32gf	BM cell classification of 21 distinct types	MLL dataset (171,375 cells)	Mean precision 0.787 ± 0.060 Mean recall: 0.755 ± 0.061 Mean F1 score: 0.762 ± 0.050	Optimizing DL training, enhancing model interpretability
Su et al. (2024) [86]	Whole-slide Aspirates	Cell Detection and Confirmation Network (CDC-NET)	Diagnosis of AML	270 BMA smear images	Precision 0.917 ± 0.031 Recall 0.785 ± 0.042 F1 score 0.846 ± 0.038	Ensemble DL framework that emulates expert consensus

Models span single-cell and whole-slide classification settings, with architectures ranging from standard CNNs to attention-based frameworks and ensembles. Notable advancements include using public benchmarks (e.g., MLL), incorporating attention mechanisms for interpretability, and label-efficient methods such as multiple instance learning (MILLIE)

around 65% for blasts in a mixed population [16]. Similarly, differentiating subtypes of dysplastic cells in MDS, or reactive vs leukemic lymphocytes, can reduce the model's precision. To gauge real-world utility, authors now often compare AI performance to the inter-expert agreement, finding that AI accuracy is approaching the variability range of human pathologists for many tasks. However, direct comparison of metrics across publications should be done cautiously. Different works use datasets (often private), class definitions, and evaluation protocols. Few studies perform external validation on independent cohorts, so a method with 95% accuracy on one dataset might not generalize to another without further tuning [3]. Efforts are underway to create public benchmark datasets and competitions to enable head-to-head comparisons of algorithms.

In summary, recent DL models frequently achieve accuracies of 90–95% on benchmark datasets [16, 68, 82], though performance varies by cell type complexity and dataset characteristics. Continued data diversity and model design improvements are expected to narrow the gap in the remaining complex cases, bringing these techniques closer to routine clinical deployment. Table 3 summarizes key studies based on their data type, network architecture, classification task, dataset size, performance metrics (accuracy, AUC, F1-score), and unique contributions.

Limitation

Direct comparison of reported performance metrics across studies should be interpreted cautiously. Differences in dataset composition, training and testing split methodologies, cell type definitions, and evaluation protocols limit comparability. The absence of standardized benchmark tasks analogous to ImageNet in general computer vision makes cross-study comparisons particularly challenging.

Clinical integration, workflow, and interpretability

The translation of deep learning tools from research into routine hematopathology practice requires careful consideration of how these systems fit into the clinical workflow. In bone marrow cytomorphology, AI-powered solutions must integrate with existing laboratory infrastructure, deliver results promptly, support effective human–AI collaboration, and comply with regulatory requirements. This section discusses key aspects of clinical integration and interpretability for bone marrow AI tools, including system compatibility, real-time deployment, human–AI interaction, and regulatory considerations.

System compatibility (DICOM, LIS, PACS)

Integrating AI-based bone marrow analysis into the clinical environment mandates compatibility with standard medical informatics systems. Digital Imaging and Communications in Medicine (DICOM) has emerged as an important format for digital pathology images, similar to its established role in radiology. Recent extensions to the DICOM standard support whole-slide imaging, enabling large pathology images (such as bone marrow aspirate slides) to be stored and managed in Picture Archiving and Communication Systems (PACS) just like radiology scans [87, 88]. Embracing DICOM for hematopathology slides means that digitized BM smears can be archived, retrieved, and viewed through enterprise PACS, facilitating multi-modality case review and long-term storage alongside other patient imaging.

Equally critical is integration with the Laboratory Information System (LIS) or specialized Anatomic Pathology LIS (AP-LIS). AI tools should ideally communicate results (e.g., differential counts, cell classifications) directly to the LIS, so that pathologists can review AI findings within the existing case management workflow. In practice, however, achieving seamless LIS integration has proven to be a challenging task. Many commercial AI solutions for digital morphology rely on proprietary viewers or even cloud-based platforms, making it non-trivial to interface them with the hospital LIS and workflow [89]. Establishing a protocol to integrate computational pathology tools into the LIS is crucial. However, many AI systems running on cloud servers with closed viewers cannot easily plug into existing lab information systems [90, 91]. This lack of interoperability can silo the AI analysis outside the routine workflow, limiting its clinical utility.

To address these compatibility issues, standards and frameworks are being developed to enhance interoperability. One approach is using Health Level 7 (HL7) messaging (or newer FHIR standards) as a bridge between the LIS and AI modules. For example, an open-source integration framework in 2025 demonstrated how the AP-LIS can send an HL7 message to an AI decision support system (AI-DSS) when a slide is ready for analysis; the AI-DSS runs the deep learning model and returns results to the LIS, where the pathologist can visualize the output in their slide viewer or LIS interface [92]. Such implementations demonstrate that, with the right middleware, AI results (such as cell classifications or heatmaps) can be seamlessly integrated into the standard reporting workflow. In addition, some commercial platforms are beginning to offer LIS connectivity, for instance, generating standardized digital reports that can be attached to the case record. Scopio's full-field morphology system, for example, produces a digital report of every assessment (with annotations and flagged abnormalities)

that is easily shareable across the care team [93, 94], which suggests it could be integrated or appended in the LIS or patient record.

Overall, system compatibility means that AI tools must integrate seamlessly with both the existing imaging and data systems. Using DICOM for image format and PACS for storage allows bone marrow slides to be handled with enterprise imaging tools. Likewise, integrating with LIS (via HL7/FHIR or vendor-specific APIs) ensures that AI-generated differentials or alerts appear in the context of the patient's case, rather than on a separate platform.

Real-time and edge inference requirements

For AI-assisted bone marrow analysis to be practical in busy clinical labs, the system must meet real-time or near-real-time performance requirements. Turnaround time (TAT) is crucial in hematopathology workflows, particularly for urgent cases, such as acute leukemias. An AI system that significantly slows down slide review would hinder rather than help. Fortunately, modern digital morphology platforms are designed to rapidly scan slides and run AI inference. For example, the Scopio X100 full-field scanner can digitize and pre-classify a peripheral blood smear in around 4–7 min per slide, which is comparable to manual review speeds [93]. The recently FDA-cleared Scopio bone marrow application utilizes AI to rapidly analyze thousands of cells in each bone marrow sample, providing clinicians with decision support without long delays [94]. In fact, their peripheral blood smear application has been reported to reduce review turnaround time by 60% compared to the traditional manual method [94], demonstrating how AI can expedite morphology analysis once slides are digitized.

Achieving this kind of performance often requires edge computing, i.e., performing the AI inference on local hardware (within the lab or on the device) rather than sending images to a remote server. Bone marrow aspirate slides are extremely large, high-resolution images (often scanned at 100 × oil immersion), so transmitting these multi-gigabyte files to an off-site cloud for analysis can be impractical and slow. Moreover, privacy regulations and hospital IT policies may limit the use of cloud services. Therefore, many hematology AI platforms process data on-site: either directly on the scanning instrument or on a local server equipped with GPUs. For instance, commercial digital morphology analyzers, such as Scopio's devices, come with built-in AI software that immediately processes the slide as it is scanned, outputting results securely to the lab network [95].

Another aspect of real-time performance is the ability to handle high throughput. Large hematology labs might process dozens of BM smears a day, alongside hundreds of blood smears. AI systems must therefore be scalable, for example, high-throughput slide scanners or slide loaders that can batch-process multiple slides. The Scopio X100HT (high-throughput model) can scan up to 30 slides in one run and analyze up to 40 samples per hour [93, 96], which accommodates the workload of larger laboratories. Ensuring that AI inference keeps pace with scanning (or even becomes the rate-limiting step rather than human review) is key to real-time workflow integration.

Finally, real-time use also implies enabling timely remote collaboration and consultation. A fully digital workflow enables multiple specialists to examine the same slide simultaneously. Some systems capitalize on this by providing real-time remote viewing over the hospital network [96]. Edge inference and efficient networking together ensure that AI results for bone marrow cytology are available when and where they are needed, without interrupting the clinical workflow.

Human–AI interaction and decision support

Introducing AI into bone marrow morphology should augment, not replace, the human expert. Human–AI interaction design is therefore crucial to ensure that pathologists and laboratory professionals can intuitively work with the AI system. A central concept is that these tools function as decision support systems (DSS), which pre-analyze slides and provide suggestions or preliminary results that the human can then verify, refine, or override. The interface between human and AI must be ergonomic, transparent, and conducive to trust.

A well-designed AI-assisted workflow typically presents the AI findings in a clear, organized manner. For bone marrow aspirates, this might mean the system highlights all the nucleated cells and groups them by lineage. From there, the human expert can interact with the results. Modern platforms allow pathologists or technologists to correct the AI's classification for any given cell easily. Studies have found that such human-in-the-loop approaches can maintain or improve accuracy while greatly increasing efficiency [97].

Transparency and interpretability of AI suggestions are vital for user acceptance. One concern with "black-box" AI models is that they might offer a prediction without any rationale, leaving the pathologist uncertain whether to trust the result. To mitigate this, developers are incorporating explainability features into hematology AI tools. For example, the Scopio bone marrow application emphasizes that it provides an in-depth explanation for every suggestion offered, aiming to engender user trust through transparency [98]. Early efforts in explainable AI (XAI) for pathology have demonstrated that showing users why the algorithm made a call can bridge the trust gap and facilitate adoption [36, 99, 100].

Another facet of human–AI interaction is the integration of AI outputs into workflows for decision support. Rather than having AI operate in isolation, the outputs are most useful when they directly inform clinical decisions. For instance, if the AI detects a high proportion of blasts in a BM smear, it could prompt additional diagnostic steps (immunophenotyping, molecular tests) faster. Ultimately, the decision support role of AI in bone marrow cytomorphology is about enhancing the human's capability by standardizing the screening process, reducing oversight of rare events, and providing quantitative data, all while keeping the human expert in control of the final interpretation.

In summary, effective human–AI interaction in hematopathology tools means intuitive user interfaces, the ability for experts to correct and guide the AI, and making the AI's "thinking" as interpretable as possible. Early clinical use of digital morphology analyzers has demonstrated that when AI's findings are presented with clarity and the option for human adjustment, collaboration can significantly enhance workflow efficiency and diagnostic confidence [24].

Regulatory and implementation considerations

The deployment of AI tools in bone marrow diagnostics must navigate regulatory approval processes and practical implementation challenges. On the regulatory front, these AI-driven systems are typically classified as medical devices or in vitro diagnostic tools, which require rigorous validation of safety and effectiveness. Notably, 2024 marked a milestone with the FDA's De Novo clearance of the first AI-powered bone marrow aspirate analysis software (Scopio Labs' Full-Field BMA application) [94]. Similarly, in other regions, regulatory bodies have begun approving these technologies. For example, the Morphogo system (developed in China) obtained a CE mark in Europe in 2020 and was authorized by China's NMPA as a medical device in 2021 [101]. These approvals signal growing trust in AI for hematopathology. However, they also impose ongoing responsibilities: manufacturers must ensure quality control, post-market surveillance, and (for learning algorithms) manage updates in line with regulatory guidance.

Implementation considerations in clinical labs are multifaceted. First, adopting an AI-driven workflow necessitates a fully digital environment, something not yet universal in pathology. Glass slides must be scanned at high quality, and many pathology departments are still transitioning to digital systems. In fact, only a minority of anatomic pathology labs are fully digitized, due to cultural resistance and technical hurdles. This means that before AI can even be introduced, hospitals may need to invest in slide scanners, servers, and network upgrades, as well as training staff in digital pathology. Such investment is substantial, and currently, there is often no direct reimbursement for digital pathology or AI analysis in many healthcare systems. The lack of reimbursement mechanisms (e.g., insurance billing codes for AI-assisted analysis) makes it more challenging to justify the cost, so institutions must consider the value proposition in terms of efficiency gains, faster turnaround times, or improved diagnostic quality.

Another key implementation step is the clinical validation and integration of workflow. Even after regulatory approval, laboratories typically perform their own validation studies to ensure the AI's performance on their patient population. Guidelines for evaluating digital cell morphology systems (e.g., CLSI H20-A2 for differential counts) recommend testing a broad range of sample types, including normal and various pathology cases, to ensure the system is robust [102]. Many early studies have indeed been pilot evaluations, often revealing that while AI differentials correlate well with manual counts for common cells, rare or dysplastic cells can be missed or misclassified [103–105]. Thus, labs must be aware of the failure modes and implement procedures for pathologist review of flagged or uncertain cases.

Regulatory compliance also means maintaining proper documentation and quality management. AI systems should be integrated into the lab's quality assurance program. Any changes to the AI algorithm may require re-validation and possibly notification to regulators, depending on the device's regulatory classification. The FDA and other agencies are actively developing frameworks for adaptive AI algorithms; however, most approved systems are currently locked models that do not change without approval. Laboratories implementing AI must stay informed about these regulatory expectations to remain compliant.

Finally, user training and change management are nontechnical but crucial considerations. Hematopathologists and laboratory technologists need to be trained not just in operating the new software, but in understanding its output and limitations. Early adopters have noted that having champion users and continual feedback loops with the vendor can smooth the implementation [92]. Moreover, clear protocols should define how discrepancies between AI and human assessments are resolved and how final decisions are made.

In summary, turning AI-powered bone marrow cytomorphology from an exciting concept into daily practice involves crossing the chasm of regulatory clearance, financial justification, and workflow adaptation. The recent FDA and CE approvals are encouraging signs of regulators' openness to these tools. To implement them successfully, pathology departments must invest in digital infrastructure, validate the tools in their own environment, and train their staff to work in a human–AI partnered workflow. When

these conditions are met, AI has the potential to significantly streamline bone marrow examinations, offering faster and more standardized results while ensuring that pathologists remain at the helm, utilizing these advanced tools to enhance clinical decision-making rather than replace it.

Diagnostic applications

DL is emerging as a powerful tool for diagnosing hematologic diseases from BM cytomorphology. AI models can analyze digitized BM aspirate smears to detect malignant cells, classify different cell lineages, and even predict genetic subtypes from morphology. These applications span a range of diagnoses from acute leukemias and BM failure syndromes to plasma cell neoplasms and promise to augment the diagnostic workflow by improving throughput and consistency in clinical hematology labs.

Acute leukemias (AML and ALL)

One of the foremost applications of AI in BM analysis is the detection and classification of acute leukemias. DL models have shown high accuracy in identifying AML blasts on BM smears, often rivaling expert performance [24]. For instance, Eckardt et al. [9] developed a convolutional neural network that detects AML with high confidence and predicts the presence of an NPM1 gene mutation directly from blast cell morphology. Another approach, Multiple Instance Learning for Leukocyte Identification (MILLIE), differentiates ALL and AML in peripheral blood and marrow smears without needing single-cell annotations. Notably, MILLIE can flag promyelocytes to recognize acute promyelocytic leukemia (APL), achieving outstanding accuracy (area under the ROC curve, AUC ≈0.94 in blood and 0.99 in marrow) [74]. These AI systems excel at recognizing leukemic blasts and abnormal cells that define acute leukemia, supporting early diagnosis and risk stratification. By rapidly scanning entire slides, AI can act as a pre-screening tool, alerting pathologists to blast-rich samples for priority review and potentially expediting critical diagnoses, such as APL, that require urgent treatment.

Myelodysplastic syndromes and aplastic anemia

Beyond acute leukemia, AI-driven image analysis has been applied to BM failure and MDS, where subtle dysplastic changes are diagnostically important. Recent studies demonstrate that DL can identify dysplastic cells across erythroid, granulocytic, and megakaryocytic lineages. Lee et al. [36] developed a model to distinguish normal versus dysplastic cells in MDS patient smears, achieving AUC values of 0.945–0.996 with an overall accuracy of ~ 91–99% in classifying dysplasia. Similarly, Wang et al. [75] introduced a CNN that automatically differentiates aplastic anemia (AA), MDS, and AML cases based on marrow smears. Their system attained excellent performance, with an AUC of 0.985 for detecting MDS (binary classification) and 0.968 for multi-class discrimination among AA/MDS/AML, corresponding to over 90% accuracy. These models address the subjectivity and labor intensity of identifying dysplastic features, such as hypo-segmented neutrophils or abnormal megakaryocytes, which are often subtle and difficult to detect. By standardizing the detection of dysplasia, AI can assist hematopathologists in diagnosing MDS more consistently and earlier. AI-based assessment of neutrophil morphology has been shown to accurately flag MDS-related changes (e.g., pseudo–Pelger–Huët anomalies), with one study reporting a sensitivity of over 95% for detecting MDS via dysplastic neutrophil features [106]. Although current algorithms for MDS/AA do not yet fully match the expertise of expert hematologists in all aspects (e.g., some have lower recall for specific dysplastic cell subtypes), they are proving valuable as an auxiliary diagnostic aid. Explainability techniques, such as heatmaps (Grad-CAM), are also being explored to highlight cell regions that influenced the AI's classification (e.g., nuclei vs. cytoplasm), which helps build clinician trust in these AI decisions.

Plasma cell neoplasms (multiple myeloma)

In plasma cell neoplasms such as multiple myeloma, the percentage of plasma cells in BM is a critical diagnostic and prognostic indicator. DL has been applied to improve the precision of plasma cell quantification on BM slides. For example, Fu et al. [53] trained a segmentation CNN to identify CD138-positive plasma cells in BM biopsy images. The AI's plasma cell counts showed excellent agreement with those of pathologists, with an intraclass correlation coefficient (ICC) greater than 0.97 between the model and expert annotations. The overall accuracy of the CNN in labeling individual cells was comparable to that of hematopathologist labels. Once validated, the model was deployed on whole slide images to output plasma cell percentages in a workflow-friendly manner. Such AI tools can reduce variability in plasma cell counts and aid in the objective diagnosis of conditions like myeloma or plasmacytosis. In addition, AI-based systems have been explored to detect abnormal plasma cells in peripheral blood or BM aspirates. For instance, an AI system was used to recognize circulating plasma cells in blood, aiding in the detection of plasma cell leukemia or high tumor burden myeloma [107]. By automating plasma cell identification, DL supports more standardized diagnostics for plasma cell dyscrasias and can facilitate

disease monitoring (e.g., assessing treatment response by measuring residual plasma cells).

Integration into digital pathology workflows

A key advantage of AI in BM diagnostics is its integration into digital pathology workflows. Whole-slide imaging (WSI) of BM aspirate smears allows thousands of cells to be captured in high resolution, which AI algorithms can then analyze in bulk. Systems like Morphogo and Scopio Labs combine specialized hardware (automated slide scanners) with AI software to enable end-to-end analysis of BM smears. In practice, a glass slide is scanned into a virtual slide, and the AI model locates and classifies each nucleated cell on the smear. This yields an automated differential count of the marrow, enumerating blasts, promyelocytes, myelocytes, erythroid precursors, and plasma cells, much faster than a manual count under the microscope. For example, the Morphogo system can run continuously and scan up to 27 slides 24/7, producing digital images for AI analysis, which significantly improves laboratory throughput [24]. AI-based pre-screening is also used to triage slides: the software can flag slides with probable abnormalities (e.g., excess blasts or dysplastic cells) for immediate human review. This integration ensures that critical cases are prioritized, addressing scenarios such as acute leukemia, where rapid diagnosis significantly impacts patient management. Moreover, digital platforms allow remote consultation; specialists can review the AI-annotated marrow slides from anywhere. Scopio's system, for instance, enables 100 × oil-equivalent digital viewing, allowing hematopathologists to examine cell details online if needed, thereby facilitating telepathology and second opinions without requiring physical slide shipment. By embedding AI into the routine workflow, labs can streamline the diagnostic process: Scopio reports that its AI-augmented peripheral blood review cut turnaround time by 60% in practice, and similar efficiency gains are anticipated for BM review [94]. AI integration into digital pathology transforms BM examination from a manual, subjective art into a faster, standardized, and telemedicine-enabled process.

AI platforms and regulatory milestones

Several AI-driven platforms for BM morphology have now moved from research into real-world clinical deployment. Notably, in 2024, the Scopio Labs Full-Field Bone Marrow Aspirate application became the first FDA-cleared AI software for BM analysis. This De Novo FDA clearance was a significant regulatory milestone, indicating the tool's safety and effectiveness for clinical use. The Scopio system combines a high-throughput slide scanner with AI algorithms to identify and quantify cells in BM aspirates. It provides decision support by highlighting diagnostically relevant cells (such as blasts, lymphoma cells, and plasma cells) and generating a comprehensive digital report for the case. The FDA's approval underscores that the platform met rigorous validation standards, giving confidence in its diagnostic accuracy. In parallel, other platforms have gained traction globally. The Morphogo system (China) is another AI-powered BM analyzer with impressive performance in validation studies, accurately classifying over 25 types of marrow cells with ~ 99% overall accuracy [24]. Morphogo's AI can automatically locate and label each cell in a whole-slide smear image, effectively performing a complete differential count and generating preliminary diagnostic opinions for review. Such systems are being adopted in hospitals to reduce manual workload and inter-observer variability in morphology assessments. Aside from commercial products, cutting-edge research prototypes, such as MILLIE [74], illustrate the potential of AI in diagnosis. While not a commercial platform, MILLIE's success in identifying APL and distinguishing AML from ALL in an annotation-free manner highlights how AI can be trained on patient-level labels to perform slide-level diagnosis. We are also seeing the integration of these AI tools with laboratory information systems and partnerships with major diagnostics companies. For example, Scopio's technology is being integrated with hematology analyzers for seamless digital review of flagged samples. With several platforms now CE-marked in Europe and FDA-cleared in the US, AI in BM diagnostics is transitioning from experimental to mainstream. These regulatory clearances and commercial deployments mark a new era where AI becomes a co-pilot to the hematologist, providing rapid preliminary evaluations that human experts can confirm and act upon.

Performance compared to human experts

A crucial aspect of deploying AI for diagnosis is ensuring performance comparable to that of experienced morphologists. DL models for BM analysis have been reported to have high sensitivity and specificity in detecting disease-specific findings. In the detection of acute leukemia, AI algorithms have achieved sensitivity and specificity in the 90–99% range, approaching the accuracy of skilled hematopathologists. For instance, a DL model for classifying BM cells showed 80.95% sensitivity and 99.48% specificity across a broad set of cell types. It rarely misses blasts or dysplastic cells and has an extremely low false-positive rate [24]. Another study's CNN could identify MDS cases with 99.2% sensitivity; virtually all true MDS cases were detected while maintaining over 97% specificity [36]. Regarding overall diagnostic accuracy, many AI models report values above

90%, and AUC values (a threshold-independent performance measure) often exceed 0.95 for key classifications [3]. These numbers indicate expert-level discrimination of cell abnormalities, comparable to that achieved by human review. In practice, side-by-side evaluations have shown encouraging concordance. Morphogo's automated counts, for example, correlated strongly with manual differential counts by pathologists (intra-class correlation > 0.9 for major lineages) [24]. Likewise, in plasma cell detection, the AI's cell-by-cell labeling was nearly as accurate as that of a pathologist, with an agreement coefficient of 0.975 between the model and an expert, compared to 0.994 between two independent experts [53]. Such reader studies suggest AI can reproduce what a trained hematologist sees under the microscope. Nevertheless, it is notable that AI performance can vary by cell type and disease: for rare cell classes or subtle dysplastic changes, models may show slightly lower F1 scores or recall. Continuous evaluation and external validation are, therefore, critical. Formal comparisons with human experts in diagnostic trials (measuring reading time, error rates, inter-observer agreement) are underway or anticipated as these tools enter clinical use. Well-trained AI systems have demonstrated performance comparable to that of human experts on specific, well-defined tasks [24, 68], although comprehensive clinical validation across diverse settings remains ongoing.

Current limitations and challenges

Inadequate external validation

The majority of studies were conducted using single-institution data, and often on internal validation sets derived from the same population used for training. Very few models have been tested on an actual external cohort of other hospitals, geographical areas, or other scanning equipment. For example, the MLL dataset [16] is one of the largest publicly available datasets; however, models trained on these data may not generalize to other institutions that use different staining solutions or microscopy systems [108]. This lack of external validation is a significant gap between the stated performance metrics and clinical utility in the real world.

Size and composition of datasets

Large datasets, despite developments, generally, most public BM datasets are modest relative to the general computer vision benchmarks. The largest publicly available single-cell dataset comprises around 171,000 cells [16], whereas commercial systems, such as Morphogo, are trained on more than 2.8 million images [24]. This discrepancy suggests that the publicly available resources may not be sufficient to train generalizable models. Moreover, the composition of datasets typically reflects the prevalence of the disease and the demographics of the contributing institutions' populations, which may limit their applicability to underrepresented populations or rare disease subtypes.

Class imbalance persistence

Although methods, such as oversampling and weighted loss functions, can decrease class imbalance during training [11], the actual performance achieved on rare cell types (e.g., blasts in non-leukemic samples, mast cells, megakaryoblasts) remains suboptimal. In the case of Matek et al. [16], there was only a 65% recall of myeloblast identification, despite high overall performance, indicating that some rare yet clinically important cell types remain difficult to identify.

Computational resource requirements

The clinical deployment of DL models requires significant computational resources. A single BM aspirate can generate gigabytes of data in whole-slide image analysis, which requires acceleration on the GPU to process promptly. Some healthcare facilities, especially those with limited resources, lack the necessary hardware (high-end GPUs), storage facilities (for archiving WSI), or the technical expertise to implement and maintain such systems. The times of inference to classify single cells are low (milliseconds), but individual slides of 50,000 cells to 100,000 cells can be processed in minutes to hours, depending on the model structure and hardware used [24, 72].

Image quality dependencies

The existing AI systems are sensitive to changes in image quality, such as focus cases [14], staining, and scanning artifacts. In contrast to human pathologists, who can compensate for poor preparation, DL models can generate inaccurate results if the input images deviate from the characteristics of the training data. The automated quality control systems that identify unsuitable images prior to analysis are not yet standardized; thus, there is a risk of false classification if poor-quality images are processed.

Black box nature and interpretability challenges

Despite the developments in explainable AI, DL models remain, to some extent, opaque. Although visualization methods can be used to indicate areas of interest, it does not necessarily identify the decision logic, especially in complex hierarchical or ensemble models. Pathologists may

struggle in understanding the reason why a model classified a borderline cell in a specific manner, making it difficult to trust and accept clinically. Furthermore, the interpretability tools themselves need to be validated, as there is limited evidence that regions of emphasis are consistently related to the same morphological features upon which human experts base their diagnoses.

Workflow integration problems

Incorporating AI into the current laboratory workflow necessitates a significant organizational transformation. The existing Laboratory Information Systems (LIS) and digital pathology systems may lack an API or data standard to facilitate seamless integration with AI. Whole-slide imaging DICOM-WSI standards are relatively new and are not universally implemented, thus posing an interoperability challenge. Moreover, the workflow of pathology should be reengineered to incorporate AI pre-screening, quality control checkpoints, and AI-pathologist conflict management simultaneously, thereby preventing regulatory violations and maintaining diagnostic precision.

Regulatory and reimbursement uncertainty

Although the Scopio system has received FDA clearance [3], AI regulatory in hematopathology is still under development. There are still liability concerns when AI is involved in diagnosis, including the need for continuous monitoring of the model and retraining it, as well as how to handle algorithm updates. Moreover, the reimbursement of AI-assisted diagnostics remains unclear in most healthcare systems, which presents a financial obstacle to its adoption.

Limited disease coverage

Current AI technologies are mainly centered on typical applications (AML/ALL detection, normal cell classification). There are many hematologic disorders, such as chronic lymphoblastic proliferative disorders, rare storage diseases, and parasitic infections in the marrow, for which there is a lack of data in AI development. AI methods for making unusual diagnoses may need years of further data gathering and verification.

User acceptance and training

Pathologists and lab technicians require training to utilize AI tools, interpret confidence scores, and recognize the limitations of the system. The experience of early adopters suggests that there is a learning curve for determining when to trust AI recommendations and when it is best to review them manually. Acceptance rates may be influenced by the presence of generational gaps in technology adoption, concerns about deskilling, and the apprehension of job loss. To establish trust, it is necessary to be transparent in model training, to monitor performance continuously, and to communicate the intended use cases clearly.

Over-reliance risk

On the other hand, too much faith in AI systems might lead to accepting incorrect results, especially when AI assigns a high confidence score to incorrect classifications. Setting the right levels of skepticism and ensuring pathologist monitoring is very important but challenging to standardize across institutions.

Architectural approaches

BM cytomorphology analysis has leveraged a variety of DL architectures, from classic convolutional neural networks (CNNs) to modern hybrid models and ensembles. This section reviews the major architectural approaches and their performance in cell classification, segmentation, and disease prediction tasks.

Single CNN models

Early and foundational BM image analysis work adopted single CNN architectures as end-to-end classifiers. Models such as VGG, ResNet, DenseNet, Inception, and EfficientNet have all been explored for classifying BM cells in microscope images. For example, Matek et al. [16] introduced a ResNeXt-50 CNN to classify 21 types of BM cells, achieving a significant advance in automated differential counts (identifying blast cells with precision ~ 0.75 and recall ~ 0.65). Subsequent studies built on this, often using ResNet family networks (e.g., ResNet-50 or variants like ResNeXt) as backbones for multi-class cell recognition. Notably, a comparative study by Glüge et al. [70] tested VGG, ResNet, RegNet, and a Transformer-based model on the same dataset and found they performed as well as or better than Matek's original ResNeXt-50 model. Interestingly, the simplest ResNet (with the fewest parameters) achieved the second-highest accuracy in that comparison, indicating that increasingly complex CNNs (with more layers/parameters) did not necessarily yield superior accuracy on that task. This suggests that well-designed, moderate-sized CNNs can suffice for BM cell classification, especially when sufficient training data and augmentation are available.

Several CNN architectures have attained expert-level performance in cell classification. DenseNet-121, for

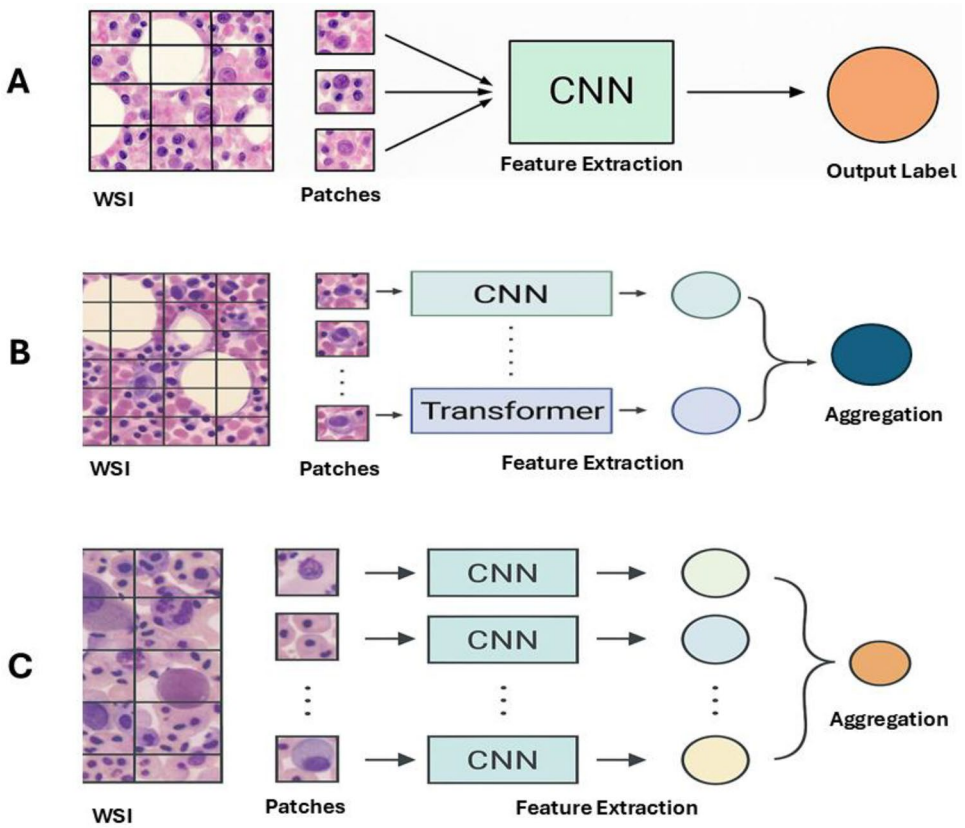

Fig. 2 Comparison of DL model architectures for whole-slide BM analysis. (**A**) A single CNN model processes WSI patches for direct classification. (**B**) A hybrid model combines CNN and Transformer-derived features, which are aggregated before classification. (**C**) An ensemble model independently processes patches through multiple CNNs and aggregates their outputs for robust prediction

instance, has been used for cell-level and region-of-interest classification. Peng et al. [109] proposed an attention-gated DenseNet (DAGDNet) to suppress background noise in single-cell images, achieving a mean precision of ~ 0.88 on the Matek et al. [15] dataset. In another study, an InceptionV3 CNN was applied to detect dysplasia in BM smears from MDS patients, achieving outstanding AUCs between 0.945 and 0.996 for distinguishing between dysplastic and normal cells [36]. Similarly, Wu et al. [13] developed a one-stage CNN-based detector/classifier (BMSNet, utilizing a YOLOv3 architecture with a SE-ResNeXt-50 backbone) that can localize and identify nucleated BM cells in a single model. Such models approached hematologist-level accuracy in identifying cell types and could process images rapidly, highlighting the feasibility of real-time analysis in clinical settings. Overall, single CNN models (whether plain or with attention mechanisms) form the backbone of many BM AI systems, excelling in tasks ranging from classifying individual cell images to screening whole-slide fields for cells of diagnostic relevance.

Hybrid CNN + transformer models

Given the success of CNNs, emerging approaches have started integrating self-attention mechanisms and transformer architectures with CNN backbones. The motivation is to combine the strengths of CNNs in local feature extraction with those of Transformers in modeling long-range dependencies and global context. Vision Transformers (ViT) and hybrid models have shown promise in hematology image analysis. For example, Tripathi et al. [110] introduced HematoNet, a hybrid model using a CoAtNet architecture (which merges convolution and attention layers). This CNN-Transformer hybrid outperformed pure CNN baselines (EfficientNetV2 and ResNeXt-50) in BM cell classification, achieving an accuracy exceeding 95% across morphological classes. In another report, a combination of CNN and ViT was used as a unified backbone, yielding higher prediction accuracy for classifying BM smear cells than traditional CNNs alone [111]. These results suggest that incorporating self-attention helps the model capture subtle morphological nuances and the cell-to-cell context in crowded marrow images.

Transformers have also been applied in segmentation and tissue-level analysis relevant to BM. Chen et al.'s [112] TransUNet architecture, which cascades a CNN with a Transformer, demonstrated state-of-the-art performance in medical image segmentation. In the BM domain, Qin et al. [113] proposed MAD-Net, a dense multi-attention network, noting that Transformer-based models, such as TransUNet and Swin-Unet, have outstanding segmentation accuracy. These architectures could facilitate tasks such as segmenting

BM biopsy histology images (e.g., identifying megakaryocytes or tumor infiltrates). Indeed, hierarchical vision Transformers, such as Swin Transformer, have been used to extract patch features in WSIs of BM, enabling attention-based pooling of information across an entire slide [114]. The combination of convolution and self-attention thus appears beneficial for cell-level and slide-level analyses, especially as datasets grow larger in 2020–2025. Figure 2 illustrates three representative architectural strategies for whole-slide image (WSI) analysis in BM cytomorphology. Panel A depicts a single CNN model that extracts features from image patches and outputs classification labels directly. Panel B demonstrates a hybrid approach, where both CNN and Transformer architectures extract aggregated features to improve context-aware classification. Panel C shows an ensemble model where multiple CNNs independently process patches, and their outputs are aggregated to enhance robustness and reduce model variance. These designs reflect evolving efforts to optimize performance, interpretability, and generalization in digital hematopathology.

Ensemble models

Researchers have explored ensemble modeling in BM cytomorphology analysis to enhance robustness and accuracy further. Ensemble approaches combine multiple models or classifiers in parallel or stages to leverage their complementary strengths. In some cases, ensembles involve training multiple CNNs and averaging their predictions (or using a majority vote), which can reduce variance and improve generalization. For instance, Zhou et al. combined three ResNet-based classifiers into an ensemble to distinguish 20 different BM cell types, achieving an accuracy of 0.829, AUC of 0.987, and F1-score of 0.829 on a large test set. This represented an improvement over any single model, especially in correctly recognizing rare cell classes (e.g., megakaryoblasts and proerythroblasts), which are critical for diagnosing acute leukemias. In another approach, ensemble logic was applied in a two-step system: one model first identified and excluded improperly prepared ("crushed") cells, and a second model then classified the remaining cells, effectively creating an ensemble pipeline that improved the final diagnosis of acute lymphoblastic leukemia with sensitivity 0.86 and specificity 0.95 [10].

Ensembles can also mix different architecture types. For example, one might ensemble a DenseNet, a ResNet, and a ViT-based model, thereby combining diverse "perspectives" on the data. In practice, many high-performing systems in digital pathology use ensembles, for instance, by training multiple CNNs on different data augmentations or splits and fusing their outputs to smooth out idiosyncrasies. The downside of ensembles is increased computational cost; however, since classifying a single cell image is relatively fast (often in the order of milliseconds on a GPU), even an ensemble of several networks can process an entire BM smear within a reasonable time.

Architectural limitations

While hybrid CNN-Transformer models show promise [92, 93], they increase computational costs and model complexity, which may limit their clinical deployment. The benefits of architectural sophistication must be balanced against practical constraints of inference time, memory requirements, and interpretability. Most deployed systems [24, 78] still rely on established CNN architectures rather than cutting-edge transformer models, suggesting a gap between research innovations and clinical implementation.

Conclusion

DL has demonstrably advanced BM cytomorphology analysis, achieving high accuracy in cell identification and subtype classification. The architectures reviewed contribute unique strengths, from standalone CNNs (ResNet, DenseNet) to transformer-infused models and ensembles. CNN-based systems offer strong performance in isolating and recognizing cell features, capturing pathologists' knowledge in a data-driven manner. Newer hybrid models further enhance this by integrating global attention and ensembling strategies to push performance toward (and in some cases beyond) the human expert level. These successes highlight the interdisciplinary synergy between AI and hematology, where computer scientists contribute state-of-the-art algorithms. At the same time, clinicians and pathologists bring domain expertise and curate high-quality data.

Despite this progress, important challenges and future directions remain before such models see routine clinical adoption:

- **Generalization Across Centers:** Models often exhibit a decline in accuracy when applied to external data from different hospitals or scanners. Variations in slide preparation, staining, and imaging can shift the data distribution. Robust training on heterogeneous, multi-center datasets is needed to ensure generalizable performance [3, 108]. Future research should focus on domain adaptation and augmentation techniques to ensure that an algorithm trained on one laboratory's slides performs reliably on another. Rigorous external validations (including prospective trials) are necessary before deployment.
- **Interpretable and Trustworthy AI:** Pathologists must understand and trust AI decisions in a clinical setting.

Thus, interpretability is key. Methods such as saliency maps and Grad-CAM highlight which cell regions influence a model's prediction, explaining why the AI classified a cell as either a blast or dysplastic [36]. Such visualization tools can uncover failure modes (e.g., a network focusing on an artifact instead of the nucleus [36, 115]. Ongoing work in explainable AI and uncertainty estimation will be vital to gaining clinician confidence and meeting regulatory transparency requirements.

- **Multi-Modal Data Integration:** Cytomorphology is only one piece of the diagnostic puzzle. In hematology, cell morphology is often combined with immunophenotyping (using flow cytometry), molecular genetics, and clinical data to facilitate an accurate diagnosis. A promising future direction is multi-modal learning that fuses image-based features with other data (e.g., combining a marrow smear image with the patient's cytogenetics or blood counts). Early studies indicate that morphological networks can even predict genetic mutations, such as NPM1, directly from cell images [116], suggesting the rich information content in cell morphology. By integrating modalities, AI models could, for example, refine a leukemia diagnosis by considering both the cell image and corresponding flow cytometry results, improving accuracy and providing insights into genotype–phenotype correlations.
- **Clinical Integration and Regulatory Approval:** Finally, translating these AI tools into routine practice requires meeting safety, efficacy, and regulatory standards. AI-powered analysis systems must be rigorously clinically validated and cleared by regulatory bodies. Encouragingly, we saw the first instances of this when the FDA recently cleared an AI-assisted BM smear analysis platform [3, 94], marking an important milestone. To build on this, researchers and clinicians should collaborate on prospective studies to demonstrate improved efficiency and diagnostic outcomes with the aid of AI. Equally important is workflow integration: the goal is for AI to augment hematologists' work, for example, by pre-screening slides and flagging abnormal cells for review, thus saving time [24]. Ensuring user-friendly interfaces and seamless incorporation into laboratory information systems will facilitate adoption.

In summary, DL-based architectural innovations propel BM cytomorphology analysis into a new era of precision. CNNs and transformers can assist in rapidly identifying cells and patterns in ways that complement expert morphological assessment, while ensemble and multi-modal approaches promise robust and comprehensive diagnostic support. By continuing to address the remaining challenges and improving generalization, interpretability, and clinical validation, the community can fully realize the potential of AI in hematology. The convergence of interdisciplinary expertise in machine learning and clinical pathology will drive further breakthroughs, ultimately leading to AI-augmented diagnostic workflows that improve patient care in hematology.

Author Contributions Shahid Mehmood, Muhammad Zubair, and Farman Matloob Khan have collected data from various resources and contributed to the preparation of the original draft. Shahid Mehmood, Asghar Ali Shah, Sagheer Abbas and Khan Muhammad Adnan performed formal analysis; writing—review and editing, Asghar Ali Shah, and Khan Muhammad Adnan; performed supervision, Muhammad Zubair and Farman Matloob Khan.; drafted pictures and tables, Asghar Ali Shah, Sagheer Abbas and Khan Muhammad Adnan,; performed revisions and improve the quality of the draft. All authors have read and agreed to the published version of the manuscript.

Funding No funding is involved in this study.

Data Availability The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding authors.

Declarations

Conflicts of Interest The authors declare no competing interests.

References

1. Lin Y, Chen Q, Chen T. Recent advancements in machine learning for bone marrow cell morphology analysis. Front Med (Lausanne). 2024;11:1402768. https://doi.org/10.3389/fmed.2024.1402768.
2. Fuentes-Arderiu X, Dot-Bach D. Measurement uncertainty in manual differential leukocyte counting. Clin Chem Lab Med. 2009;47(1):112–5. https://doi.org/10.1515/CCLM.2009.014.
3. Ghete T, Kock F, Pontones M, Pfrang D, Westphal M, Höfener H, et al. Models for the marrow: a comprehensive review of AI-based cell classification methods and malignancy detection in bone marrow aspirate smears. Hemasphere. 2024;8(12):e70048. https://doi.org/10.1002/hem3.70048.
4. 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):349–64.
5. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering. 2023;10(12):1435.
6. Hijazi A, Bifulco C, Baldin P, Galon J. Digital pathology for better clinical practice. Cancers. 2024;16(9):1686.
7. Shamir SB, Sasson AL, Margolies LR, Mendelson DS. New frontiers in breast cancer imaging: the rise of AI. Bioengineering. 2024;11(5):451.
8. 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 (Baltimore). 2020;99(45):e23154. https://doi.org/10.1097/MD.0000000000023154.
9. Eckardt JN, Schmittmann T, Riechert S, et al. Deep learning identifies acute promyelocytic leukemia in bone marrow smears. BMC Cancer. 2022;22(1):201.

10. Zhou M, Wu K, Yu L, Xu M, Yang J, Shen Q, et al. Development and evaluation of a leukemia diagnosis system using deep learning in real clinical scenarios. Front Pediatr. 2021;9:693676.
11. Guo L, Huang P, Huang D, Li Z, She C, Guo Q, et al. A classification method to classify bone marrow cells with class imbalance problem. Biomed Signal Process Control. 2022;72:103296. https://doi.org/10.1016/j.bspc.2021.103296.
12. Su J, Liu S, Song JA. A segmentation method based on HMRF for the aided diagnosis of acute myeloid leukemia. Comput Methods Programs Biomed. 2017;152:115–23. https://doi.org/10.1016/j.cmpb.2017.09.011.
13. Wu Y, Huang T, Ye R, Fang W, Lai S, Chang P, et al. A hematologist-level deep learning algorithm (BMSNet) for assessing the morphologies of single nuclear balls in bone marrow smears: algorithm development. J Mir Med Inform. 2020;8(4):e15963. https://doi.org/10.2196/15963.
14. 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. https://doi.org/10.1371/journal.pone.0205387.
15. Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. An expert-annotated dataset of bone marrow cytology in hematologic malignancies. Cancer Imaging Archive. 2021. https://doi.org/10.7937/TCIA.AXH3-T579.
16. 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):1917–27. https://doi.org/10.1182/blood.2020010568.
17. Cheng Z, Li Y. Improved YOLOv7 algorithm for detecting bone marrow cells. Sensors (Basel). 2023;23(17):7640. https://doi.org/10.3390/s23177640.
18. Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: a database and web-based tool for image annotation. Int J Comput Vis. 2008;77:157–73. https://doi.org/10.1007/s11263-007-0090-8.
19. Chen Y, Zhu Z, Zhu S, Qiu L, Zou B, Jia F, et al. Sckansformer: fine-grained classification of bone marrow cells via kansformer backbone and hierarchical attention mechanisms. IEEE J Biomed Health Inform. 2024. https://doi.org/10.1109/JBHI.2024.3471928.
20. Raina P, Dharamdasani S, Chinnam D, Sharma P, Gupta S. BoMBR: an annotated bone marrow biopsy dataset for segmentation of reticulin fibers. BioRxiv. 2024. https://doi.org/10.1101/2024.10.02.616389.
21. Singh A, Dharamdasani S, Sharma P, Gupta S. BaMBo: an annotated bone marrow biopsy dataset for segmentation task. BioRxiv. 2024. https://doi.org/10.1101/2024.10.02.616393.
22. Bodzas A, Kodytek P, Zidek J. A high-resolution large-scale dataset of pathological and normal white blood cells. Sci Data. 2023;10:466. https://doi.org/10.1038/s41597-023-02378-7.
23. Bodzas A, Kodytek P, Žídek J. A high-resolution large-scale dataset of pathological and normal white blood cells. figshare. Collection. 2023. https://doi.org/10.6084/m9.figshare.c.6612970.v1.
24. Lv Z, Cao X, Jin X, et al. High-accuracy morphological identification of bone marrow cells using deep learning-based Morphogo system. Sci Rep. 2023;13:13364. https://doi.org/10.1038/s41598-023-40424-x.
25. Bini L, Mojarrad FN, Liarou M, Matthes T, Marchand-Maillet S. FlowCyt: A Comparative Study of Deep Learning Approaches for Multi-Class Classification in Flow Cytometry Benchmarking. arXiv preprint arXiv:2403.00024. 2024.
26. Eckardt JN, Srivastava I, Wang Z, et al. Synthetic bone marrow images enhance real samples in the development of microscopy classification models for acute myeloid leukemia. npj Digit Med. 2025;8:173. https://doi.org/10.1038/s41746-025-01563-9.
27. Choi J, Ku Y, Yoo B, Kim J, Lee D, Chai Y, et al. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. PLoS ONE. 2017;12(12):e0189259. https://doi.org/10.1371/journal.pone.0189259.
28. Wang G, Zhao T, Fang Z, Lian H, Wang X, Li Z, et al. Experimental evaluation of deep learning method in reticulocyte enumeration in peripheral blood. Int J Lab Hematol. 2021;43:597–601. https://doi.org/10.1111/ijlh.13588.
29. Abdulrahman AA, Patel KH, Yang T, Koch DD, Sivers SM, Smith GH, et al. Is a 500-cell count necessary for bone marrow differentials? A proposed analytical method for validating a lower cutoff. Am J Clin Pathol. 2018;150:84–91. https://doi.org/10.1093/ajcp/aqy034.
30. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27.
31. Ludwig H, Baracaldo N, editors. Federated learning: a comprehensive overview of methods and applications. Springer; 2022.
32. Schoenpflug LA, Nie Y, Sheikhzadeh F, Koelzer VH. A review on federated learning in computational pathology. Comput Struct Biotechnol J. 2024;23:3938–45. https://doi.org/10.1016/j.csbj.2024.10.037.
33. Lian JW, Wei CH, Chen MY, Lin CC. Acute leukemia prediction and classification using convolutional neural network and generative adversarial network. Appl Soft Comput. 2024;163:111819.
34. Zhang Z, Huang X, Yan Q, Lin Y, Liu E, Mi Y, et al. The diagnosis of chronic myeloid leukemia with deep adversarial learning. Am J Pathol. 2022;192(7):1083–91.
35. Xie Y, Xing F, Shi X, Kong X, Su H, Yang L. Efficient and robust cell detection: a structured regression approach. Med Image Anal. 2018;44:245–54. https://doi.org/10.1016/j.media.2017.07.003.
36. Lee N, Jeong S, Park MJ, et al. Deep learning application of the discrimination of bone marrow aspiration cells in patients with myelodysplastic syndromes. Sci Rep. 2022;12:18677. https://doi.org/10.1038/s41598-022-21887-w.
37. D'Abbronzo G, D'Antonio A, De Chiara A, Panico L, Sparano L, Diluvio A, et al. Development of an artificial-intelligence-based tool for automated assessment of cellularity in bone marrow biopsies in Ph-negative myeloproliferative neoplasms. Cancers (Basel). 2024;16(9):1687. https://doi.org/10.3390/cancers16091687.
38. Otsu N. A threshold selection method from gray-level histograms. Automatica. 1975;11(285–296):23–7.
39. Liu H, Cao H, Song E. Bone marrow cells detection: a technique for the microscopic image analysis. J Med Syst. 2019;43:82. https://doi.org/10.1007/s10916-019-1185-9.
40. Osowski S, Markiewicz T, Marianska B, Moszczyński L. Feature generation for the cell image recognition of myelogenous leukemia. In: 2004 12th European Signal Processing Conference. IEEE; 2004. p. 753–756.
41. Shih FY. Image processing and mathematical morphology: fundamentals and applications. CRC Press; 2017.
42. Arslan S, Ozyurek E, Gunduz-Demir CA. Color and shape-based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images. Cytometry A. 2014;85:480–90. https://doi.org/10.1002/cyto.a.22457.
43. Theera-Umpon N. White blood cell segmentation and classification in microscopic bone marrow images. In: Proceedings of the international conference on fuzzy systems and knowledge discovery. Berlin: Springer; 2005. p. 787–96.
44. Song T, Sanchez V, EIDaly H, Rajpoot N. Simultaneous cell detection and classification in bone marrow histology images. IEEE J Biomed Health Inform. 2019;23:1469–76. https://doi.org/10.1109/JBHI.2018.2878945.

45. Conze PH, Andrade-Miranda G, Singh VK, Jaouen V, Visvikis D. Current and emerging trends in medical image segmentation with deep learning. IEEE Trans Radiat Plasma Med Sci. 2023;7(6):545–69.
46. Bendiabdallah MH, Settouti N. A comparison of U-net backbone architectures for the automatic white blood cells segmentation. WAS Science Nature (WASSN) ISSN: 2766–7715. 2021;4(1).
47. Du G, Cao X, Liang J, Chen X, Zhan Y. Medical image segmentation based on U-net: a review. J Imaging Sci Technol. 2020;64(2).
48. Qin J, Liu T, Wang Z, Liu L, Fang H. GCT-UNET: u-net image segmentation model for a small sample of adherent bone marrow cells based on a gated channel transform module. Electronics. 2022;11:3755. https://doi.org/10.3390/electronics11223755.
49. Zhang Z, Arabyarmohammadi S, Leo P, Meyerson H, Metheny L, Xu J, et al. Automatic myeloblast segmentation in acute myeloid leukemia images based on adversarial feature learning. Comput Methods Programs Biomed. 2024;243:107852. https://doi.org/10.1016/j.cmpb.2023.107852.
50. Fang T, Yuan P, Gong C, Jiang Y, Yu Y, Shang W, et al. Fast label-free recognition of NRBCs by deep-learning visual object detection and single-cell Raman spectroscopy. Analyst. 2022;147:1961–7. https://doi.org/10.1039/d2an00024e.
51. Statkevych R, Stirenko S, Gordienko Y. Human kidney tissue image segmentation by U-Net models. In: IEEE EUROCON 2021–19th international conference on smart technologies. IEEE; 2021. p. 129–34.
52. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Cham: Springer Nature Switzerland; 2022. p. 205–218.
53. Fu F, Guenther A, Sakhdari A, McKee TD, Xia D. Deep learning accurately quantifies plasma cell percentages on CD138-stained bone marrow samples. J Pathol Inform. 2022;13:100011. https://doi.org/10.1016/j.jpi.2022.100011.
54. Song T, Sanchez V, EIDaly H, Rajpoot N. Dual-channel active contour model for megakaryocytic cell segmentation in bone marrow trephine histology images. IEEE Trans Biomed Eng. 2017;64:2913–23. https://doi.org/10.1109/TBME.2017.2690863.
55. Moshavash Z, Danyali H, Helfroush M. An automatic and robust decision support system for accurate acute leukemia diagnosis from blood microscopic images. J Digit Imaging. 2018;31:702–17. https://doi.org/10.1007/s10278-018-0074-y.
56. Duan J, Yu L. A WBC segmentation method based on HSI color space. In: Proceedings of the 2011 4th IEEE international conference on broadband network and multimedia technology. Shenzhen; 2011. p. 629–32. https://doi.org/10.1109/ICBNMT.2011.6156011
57. Chen J, Wu X, Wang C, Yang Z, Wu X, Ran L, Liu Y. Texture-Unet: A Texture-Aware Network for Bone Marrow Smear Whole-Slide Image Region of Interest Segmentation. 2024;2006–2010. https://doi.org/10.1109/ICASSP48485.2024.10447261.
58. Ratwal J, Bekri D, Boussema C, Sarkis R, Kunz N, Koliqi T, et al. Marrowquant across aging and aplasia: a digital pathology workflow for quantification of bone marrow compartments in histological sections. Front Endocrinol (Lausanne). 2020;11:480. https://doi.org/10.3389/fendo.2020.00480.
59. Sarkis R, Burri O, Royer-Chardon C, Schyrr F, Blum S, Costanza M, et al. MarrowQuant 2.0: a digital pathology workflow assisting bone marrow evaluation in experimental and clinical hematology. Mod Pathol. 2023;36(4):100088. https://doi.org/10.1016/j.modpat.2022.100088.
60. 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):318–27. https://doi.org/10.1016/j.pathol.2021.07.011.
61. Xu Y, Quan R, Xu W, Huang Y, Chen X, Liu F. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. Bioengineering. 2024;11(10):1034. https://doi.org/10.3390/bioengineering11010034.
62. Wang X, Wang Y, Qi C, Qiao S, Yang S, Wang R, 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;22:15330338221150068. https://doi.org/10.1177/15330338221150069.
63. Kutlu H, Avci E, Özyurt F. White blood cells detection and classification based on regional convolutional neural networks. Med Hypotheses. 2020;135:109472. https://doi.org/10.1016/j.mehy.2019.109472.
64. Tang G, Fu X, Wang Z, Chen M. A machine learning tool using digital microscopy (morphogo) for the identification of abnormal lymphocytes in the bone marrow. Acta Cytol. 2021;65(4):354–7. https://doi.org/10.1159/000518382.
65. Moura P, Dobbe J, Streekstra G, Rab M, Veldthuis M, Fermo E, et al. Rapid diagnosis of hereditary haemolytic anaemias using automated rheoscopy and supervised machine learning. Br J Haematol. 2020;190(4):e250–5. https://doi.org/10.1111/bjh.16868.
66. Lewis JE, Shebelut CW, Drumheller BR, Zhang X, Shanmugam N, Attieh M, et al. An automated pipeline for differential cell counts on whole-slide bone marrow aspirate smears. Mod Pathol. 2023;36(2):100003.
67. 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):538–44. https://doi.org/10.1038/s42256-019-0101-9.
68. Goldgof GM, Sun S, Van Cleave J, Wang L, Lucas F, Brown L, Spector JD, Boiocchi L, Baik J, Zhu M, Ardon O. DeepHeme: A generalizable, bone marrow classifier with hematopathologist-level performance. bioRxiv. 2023.
69. 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. https://doi.org/10.1177/1533033818802789.
70. 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.
71. Müller D, Soto-Rey I, Kramer F. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. IEEE Access. 2022;10:66467–80. https://doi.org/10.1109/ACCESS.2022.3182399.
72. Wang C, Huang S, Lee Y, Shen Y, Meng S, Gaol J. Deep learning for bone marrow cell detection and classification on whole-slide images. Med Image Anal. 2022;75:102270. https://doi.org/10.1016/j.media.2021.102270.
73. Chandradevan R, Aljudi A, Drumheller B, Kunananthaseelan N, Amgad M, Gutman D, 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. https://doi.org/10.1038/s41374-019-0325-7.
74. Manescu P, Narayanan P, Bendkowski C, Elmi M, Claveau R, Pawar V, et al. Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning. Sci Rep. 2023;13(1):2562.
75. 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.

76. Tayebi RM, Mu Y, Dehkharghanian T, Ross C, Sur M, Foley R, et al. Automated bone marrow cytology using deep learning to generate a histogram of cell types. Commun Med. 2022;2(1):45.
77. 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.
78. Fu X, Fu M, Li Q, Peng X, Lu J, Fang F, et al. Morphogo: an automatic bone marrow cell classification system on digital images analyzed by artificial intelligence. Acta Cytol. 2020;64(6):588–96.
79. Romould RV, Payne P, Chatterjee R, Gourisaria MK. EfficientNetB0: Comparing Transfer Learning and Scratch Training on Benchmark Datasets. In2025 Seventh International Conference on Computational Intelligence andCommunication Technologies (CCICT). IEEE; 2025. p. 416–421.
80. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC Med Imaging. 2022;22(1):69.
81. Nakamura I, Ida H, Yabuta M, Kashiwa W, Tsukamoto M, Sato S, et al. Evaluation of two semi-supervised learning methods and their combination for automatic classification of bone marrow cells. Sci Rep. 2022;12:16736. https://doi.org/10.1038/s41598-022-20651-4.
82. Alshahrani H, Sharma G, Anand V, Gupta S, Sulaiman A, Elmagzoub MA, 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.
83. 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.
84. 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. https://doi.org/10.1016/j.cmpb.2022.107019.
85. Ananthakrishnan 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. 2022;13(1):112.
86. 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):575–89. https://doi.org/10.1007/s11517-023-02955-3.
87. Gorman C, Punzo D, Octaviano I, Pieper S, Longabaugh WJ, Clunie DA, et al. Interoperable slide microscopy viewer and annotation tool for imaging data science and computational pathology. Nat Commun. 2023;14(1):1572.
88. Clunie DA. DICOM format and protocol standardization—a core requirement for digital pathology success. Toxicol Pathol. 2021;49(4):738–49. https://doi.org/10.1177/0192623320965893.
89. Wagner SJ, Matek C, Boushehri SS, Boxberg M, Lamm L, Sadafi A, et al. Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology. Mod Pathol. 2024;37(1):100350.
90. Asif A, Rajpoot K, Graham S, Snead D, Minhas F, Rajpoot N. Unleashing the potential of AI for pathology: challenges and recommendations. J Pathol. 2023;260(5):564–77. https://doi.org/10.1002/path.6168.
91. Kaczmarzyk JR, O'Callaghan A, Inglis F, Gat S, Kurc T, Gupta R, et al. Open and reusable deep learning for pathology with WSInfer and QuPath. npj Precision Oncology. 2024;8(1):9.
92. Angeloni M, Rizzi D, Schoen S, Caputo A, Merolla F, Hartmann A, et al. Closing the gap in the clinical adoption of computational pathology: a standardized, open-source framework to integrate deep-learning models into the laboratory information system. Genome Med. 2025;17(1):60.
93. Katz BZ, Feldman MD, Tessema M, Benisty D, Toles GS, Andre A, et al. Evaluation of Scopio labs X100 full field PBS: the first high-resolution full field viewing of peripheral blood specimens combined with artificial intelligence-based morphological analysis. Int J Lab Hematol. 2021;43(6):1408–16. https://doi.org/10.1111/ijlh.13681.
94. Scopio Labs Receives FDA Clearance For AI-Driven Bone Marrow Analysis Application | Inside Precision Medicine. [Internet]. Retrieved October 30, 2025, from https://www.insideprecisionmedicine.com/topics/molecular-dx/scopio-labs-receives-fda-clearance-for-ai-driven-bone-marrow-analysis-application/
95. Scopio Labs Launches World's First Digital Application for Bone Marrow Aspirate Imaging and Review | Scopio Labs. [Internet]. Retrieved October 30, 2025, from https://scopiolabs.com/news/scopio-labs-launches-worlds-first-digital-application-for-bone-marrow-aspirate-imaging-and-review/
96. Siemens to distribute Scopio Labs' digital imaging platforms. [Internet]. Retrieved October 31, 2025, from https://www.medicaldevice-network.com/news/siemens-healthineers-scopio-labs-digital-imaging/
97. Saft L, Vaara E, Ljung E, Kwiecinska A, Kumar D, Timar B. A deep-learning algorithm (AIFORIA) for classification of hematopoietic cells in bone marrow aspirate smears based on nine cell classes—a feasible approach for routine screening? J Hematopathol. 2025;18(1):12. https://doi.org/10.1007/s12308-025-00625-x.
98. Scopio Full-field Digital Cell Morphology - Siemens Healthineers USA. [Internet]. Retrieved October 31, 2025, from https://www.siemens-healthineers.com/en-us/hematology/systems/scopio
99. Hehr M, Sadafi A, Matek C, Lienemann P, Pohlkamp C, Haferlach T, et al. Explainable AI identifies diagnostic cells of genetic AML subtypes. PLoS Digit Health. 2023;2(15):e0000187.
100. Zanca F, Brusasco C, Pesapane F, Kwade Z, Beckers R, Avanzo M. Regulatory aspects of the use of artificial intelligence medical software. In: Seminars in radiation oncology, vol. 32, No. 4. WB Saunders; 2022. p. 432–41.
101. Zhiwei Info&Tech - Revolutionize medical examination with artificial intelligence. [Internet]. Retrieved October 31, 2025, from https://www.morphogo.com/en/
102. Kim H, Hur M, d'Onofrio G, Zini G. Real-world application of digital morphology analyzers: practical issues and challenges in clinical laboratories. Diagnostics. 2025;15(6):677.
103. Lapić I, Miloš M, Dorotić M, Drenški V, Coen Herak D, Rogić D. Analytical validation of white blood cell differential and platelet assessment on the Sysmex DI-60 digital morphology analyzer. Int J Lab Hematol. 2023;45(5):668–77. https://doi.org/10.1111/ijlh.14101.
104. Khongjaroensakun N, Chinudomwong P, Chaothai N, Chamchomdao L, Suriyachand K, Paisooksantivatana K. Retracted: White blood cell differentials performance of a new automated digital cell morphology analyzer: Mindray MC-80. Int J Lab Hematol. 2023;45(2):260.
105. Zini G. How i investigate difficult cells at the optical microscope. Int J Lab Hematol. 2021;43(3):346–53. https://doi.org/10.1111/ijlh.13437.
106. Romano NH, Habringer S, Schläpfer P, Ruiz C, Widmer CC. AI-based detection of myelodysplastic syndrome from neutrophil morphology in peripheral blood. Blood. 2024;144:7507.
107. Chen P, Zhang L, Cao X, et al. Detection of circulating plasma cells in peripheral blood using deep learning-based morphological analysis. Cancer. 2024;130(10):1884–93. https://doi.org/10.1002/cncr.35202.
108. Seoni S, Shahini A, Meiburger KM, Marzola F, Rotunno G, Acharya UR, et al. All you need is data preparation: a systematic review of image harmonization techniques in Multi-center/device

studies for medical support systems. Comput Methods Programs Biomed. 2024;23:108200.
109. 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):16971–81. https://doi.org/10.1007/s00432-023-05384-9.
110. Tripathi S, Augustin AI, Sukumaran R, Dheer S, Kim E. HematoNet: expert level classification of bone marrow cytology morphology in hematological malignancy with deep learning. Artificial Intelligence in the Life Sciences. 2022;2(1):100043.
111. Wang J. Deep learning in hematology: from molecules to patients. Clin Hematol Int. 2024;6(8):19.
112. Zeng J, Li W, Zheng B, Xiao L, Zhang X, Zhong Q, et al. ACMTR: attention-guided, combined multi-scale, transformer reasoning-based network for 3D CT pelvic functional bone marrow segmentation. Biomed Signal Process Control. 2023;82(1):104522.
113. Qin C, Zheng B, Li W, Chen H, Zeng J, Wu C, et al. MAD-Net: multi-attention dense network for functional bone marrow segmentation. Comput Biol Med. 2023;154(1):106428.
114. Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal. 2022;81(1):102559.
115. Deshpande NM, Gite S, Pradhan B. Unlocking the potential: machine learning and deep learning in leukemia diagnosis with explainable AI. IoT Sensors, ML, AI and XAI: Empowering A Smarter World. 2024:201–58
116. Eckardt JN, Middeke JM, Riechert S, Schmittmann T, Sulaiman AS, Kramer M, et al. Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. Leukemia. 2022;36(1):111–8. https://doi.org/10.1038/s41375-021-01408-w.