CAT

Critically Appraised Topic

Titel: Models and applications of AI in hematological diagnostics

Author: Alexander Coulie

Supervisor: Christine Van Laer

Search/methodology verified by: Christine Van Laer

Date: 01/09/2025

CLINICAL BOTTOM LINE

Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL) models, shows strong potential to enhance diagnostic accuracy, efficiency, and reproducibility in hematology. Applications such as peripheral blood and bone marrow smear analysis, flow cytometry interpretation, and detection of dysplastic features in myelodysplastic syndromes (MDS) have demonstrated high sensitivity and specificity across multiple studies. These systems can support clinicians by reducing observer variability and accelerating workflows, especially in complex diagnostic settings.

However, challenges remain. Many models rely on single-center datasets with limited external validation, raising concerns about generalizability. Interpretability ("black box" effect), data quality, class imbalance, and regulatory hurdles further complicate clinical integration. Despite these barriers, the consistent high performance of AI-based models across proof-of-concept and early validation studies suggests that, with appropriate standardization, explainability measures, and multidisciplinary collaboration, AI will become an essential adjunct to hematological diagnostics rather than a replacement for expert clinical judgment.

CLINICAL/DIAGNOSTIC SCENARIO

The 5th edition of the World Health Organization (WHO) Classification of Haematolymphoid Tumours, 2022 introduced a new approach to diagnosing hematological malignancies (Alaggio et al., 2022). By combining clinical data, morphological findings, flow cytometry, and molecular and genetic biomarkers, WHO 2022 aims to achieve increasingly accurate and consistent diagnoses. Despite improved consensus, the diagnosis of hematological malignancies remains dependent on hematological expertise. In recent years the development of artificial intelligence (AI) took an unprecedented speed. This is no different in the field of medicine and particularly within (hematological) laboratory medicine.

AI models are now being applied for the analysis of complex or large sets of hematologic data. From the interpretation of digital images of blood smears and bone marrow aspirates using convolutional neural networks (CNNs), to the integration of laboratory, genomic, and electronic health record data for diagnosis and prognosis (Herman et al., 2021; Radakovich et al., 2020). Interest in AI-based models and their applications is growing which also can be noted at the annual meeting of the American Society of Hematology (ASH). In 2024, more than 175

abstracts reporting on AI-based systems or ML models for diagnosis, prognosis, and treatment decisions were accepted for publication representing a 14-fold increase compared to 2020 (Figure 1) (n.d, 2025)

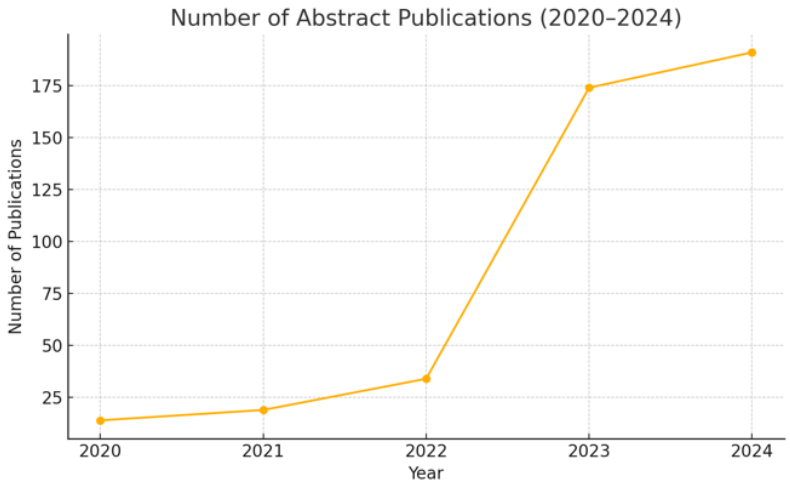

Figure 1, Increase in abstracts with AI, artificial intelligence as search term (ASH)

Due to the increased interest in AI, multidisciplinary collaboration between clinical biologists, laboratory staff, and engineers specialized in computer science will become essential. Although many laboratory staff and clinicians understand the importance of AI, many still get cold feet when it comes to putting it to practice (S.-X. Wang et al., 2024). A stakeholder survey found that while most in laboratory medicine believe AI will be valuable for improving efficiency and quality of care, poor AI knowledge, high costs, lack of clinical evidence, and implementation barriers remain major obstacles to its adoption (Paranjape et al., 2021).

This CAT aims to provide an overview of various AI models, with a particular focus on their role in hematological malignancies diagnostics. It is intended as a brief guide for clinical biologists interested in exploring the potential of AI, rather than a deep dive into the mathematical foundations of machine learning. By examining recent scientific literature and clinical applications, this CAT evaluates the extent to which artificial intelligence can enhance current diagnostic and treatment approaches in hematological disorders.

QUESTION(S)

1) What are the general principles of AI and ML in hematology diagnostics?
2) How can AI play a role in hematology focused on cytomorphology, flowcytometry and MDS?
3) What are the key barriers (technical, clinical, and regulatory) to the implementation of AI-driven diagnostic models in routine hematology laboratories?

SEARCH TERMS

During this literature review, the Medline (Pubmed) database was systematically searched for eligible articles concerning artificial intelligence applications in hematologic diagnostics. The following key words were used: “Artificial Intelligence”, “AI”, “deep learning”, “neural networks”, “machine learning”, “cytomorphology”, “peripheral blood smear”, “bone marrow aspirate”. In addition, the reference lists of the retrieved articles were searched.

RELEVANT EVIDENCE/REFERENCES

Aboy, M., Minssen, T., & Vayena, E. (2024). Navigating the EU AI Act: implications for regulated digital medical products. *Npj Digital Medicine*, *7*(1), 237. https://doi.org/10.1038/s41746-024-01232-3

Acevedo, A., Alférez, S., Merino, A., Puigví, L., & Rodellar, J. (2019). Recognition of peripheral blood cell images using convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *180*, 105020. https://doi.org/10.1016/j.cmpb.2019.105020

Acevedo, A., Merino, A., Boldú, L., Molina, Á., Alférez, S., & Rodellar, J. (2021). A new convolutional neural network predictive model for the automatic recognition of hypogranulated neutrophils in myelodysplastic syndromes. *Computers in Biology and Medicine*, *134*, 104479. https://doi.org/10.1016/j.compbiomed.2021.104479

Alaggio, R., Amador, C., Anagnostopoulos, I., Attygalle, A. D., Araujo, I. B. de O., Berti, E., Bhagat, G., Borges, A. M., Boyer, D., Calaminici, M., Chadburn, A., Chan, J. K. C., Cheuk, W., Chng, W.-J., Choi, J. K., Chuang, S.-S., Coupland, S. E., Czader, M., Dave, S. S., … Xiao, W. (2022). The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia*, *36*(7), 1720–1748. https://doi.org/10.1038/s41375-022-01620-2

Alomari, Y. M., Sheikh Abdullah, S. N. H., Zaharatul Azma, R., & Omar, K. (2014). Automatic Detection and Quantification of WBCs and RBCs Using Iterative Structured Circle Detection Algorithm. *Computational and Mathematical Methods in Medicine*, *2014*(1), 979302. https://doi.org/https://doi.org/10.1155/2014/979302

Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA*, *319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Cheng, F.-M., Lo, S.-C., Lin, C.-C., Lo, W.-J., Chien, S.-Y., Sun, T.-H., & Hsu, K.-C. (2024). Deep learning assists in acute leukemia detection and cell classification via flow cytometry using the acute leukemia orientation tube. *Scientific Reports*, *14*(1), 8350. https://doi.org/10.1038/s41598-024-58580-z

Cheuque, C., Querales, M., León, R., Salas, R., & Torres, R. (2022). An Efficient Multi-Level Convolutional Neural Network Approach for White Blood Cells Classification. *Diagnostics (Basel, Switzerland)*, *12*(2). https://doi.org/10.3390/diagnostics12020248

China, C. R. (2023, December 20). *Five machine learning types to know*. https://www.ibm.com/think/topics/machine-learning-types

Dhar, T., Dey, N., Borra, S., & Sherratt, R. S. (2023). Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society*, *4*(1), 68–75. https://doi.org/10.1109/TTS.2023.3234203

Dilmegani, C. (2025). *Data Quality in AI: Challenges, Importance & Best Practices*. AIMultiple Research.

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Fein, J., & Shouval, R. (2021). A Reader's Guide to Machine Learning in Hematology. In *Hematopoiesis (ASH Trainee Council)*. American Society of Hematology.

Fürnkranz, J. (2010). Decision Tree. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 263–267). Springer US. https://doi.org/10.1007/978-0-387-30164-8_204

Ghaderzadeh, M., Aria, M., Hosseini, A., Asadi, F., Bashash, D., & Abolghasemi, H. (2022). A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood smear images. *International Journal of Intelligent Systems*, *37*(8), 5113–5133. https://doi.org/https://doi.org/10.1002/int.22753

Ghete, T., Kock, F., Pontones, M., Pfrang, D., Westphal, M., Höfener, H., & Metzler, M. (2024). Models for the marrow: A comprehensive review of AI-based cell classification methods and malignancy detection in bone marrow aspirate smears. *HemaSphere*, *8*(12), e70048. https://doi.org/10.1002/hem3.70048

Guo, L., Huang, P., Huang, D., Li, Z., She, C., Guo, Q., Zhang, Q., Li, J., Ma, Q., & Li, J. (2022). A classification method to classify bone marrow cells with class imbalance problem. *Biomedical Signal Processing and Control*, *72*, 103296. https://doi.org/https://doi.org/10.1016/j.bspc.2021.103296

Herman, D. S., Rhoads, D. D., Schulz, W. L., & Durant, T. J. S. (2021). Artificial Intelligence and Mapping a New Direction in Laboratory Medicine: A Review. *Clinical Chemistry*, *67*(11), 1466–1482. https://doi.org/10.1093/clinchem/hvab165

Hodes, A., Calvo, K. R., Dulau, A., Maric, I., Sun, J., & Braylan, R. (2019). The challenging task of enumerating blasts in the bone marrow. *Seminars in Hematology*, *56*(1), 58–64. https://doi.org/10.1053/J.SEMINHEMATOL.2018.07.001

Khan, M. B., Islam, T., Ahmad, M., Shahrior, R., & Riya, Z. N. (2021). A CNN Based Deep Learning Approach for Leukocytes Classification in Peripheral Blood from Microscopic Smear Blood Images. In M. S. Uddin & J. C. Bansal (Eds.), *Proceedings of International Joint Conference on Advances in Computational Intelligence* (pp. 67–76). Springer Singapore.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, N., Jeong, S., Park, M.-J., & Song, W. (2022). Deep learning application of the discrimination of bone marrow aspiration cells in patients with myelodysplastic syndromes. *Scientific Reports*, *12*(1), 18677. https://doi.org/10.1038/s41598-022-21887-w

Lewis, J. E., Shebelut, C. W., Drumheller, B. R., Zhang, X., Shanmugam, N., Attieh, M., Horwath, M. C., Khanna, A., Smith, G. H., Gutman, D. A., Aljudi, A., Cooper, L. A. D., & Jaye, D. L. (2023). An Automated Pipeline for Differential Cell Counts on Whole-Slide Bone Marrow Aspirate Smears. *Modern Pathology*, *36*(2), 100003. https://doi.org/10.1016/J.MODPAT.2022.100003

Li, H. (2024). *Introduction to Unsupervised Learning BT - Machine Learning Methods* (H. Li, Ed.; pp. 281–292). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3917-6_13

Lones, M. A. (2024). Avoiding common machine learning pitfalls. *Patterns*, *5*(10), 101046. https://doi.org/10.1016/J.PATTER.2024.101046

Monaghan, S. A., Li, J.-L., Liu, Y.-C., Ko, M.-Y., Boyiadzis, M., Chang, T.-Y., Wang, Y.-F., Lee, C.-C., Swerdlow, S. H., & Ko, B.-S. (2022). A Machine Learning Approach to the Classification of Acute Leukemias and Distinction From Nonneoplastic Cytopenias Using Flow Cytometry Data. *American Journal of Clinical Pathology*, *157*(4), 546–553. https://doi.org/10.1093/ajcp/aqab148

Morell, A. (2025, August 19). *AI & Innovation at CellaVision: 30 Years of Intelligent Microscopy*. https://www.cellavision.com/about/news/ai-innovation-cellavision-30-years-intelligent-microscopy

Mori, J., Kaji, S., Kawai, H., Kida, S., Tsubokura, M., Fukatsu, M., Harada, K., Noji, H., Ikezoe, T., Maeda, T., & Matsuda, A. (2020). Assessment of dysplasia in bone marrow smear with convolutional neural network. *Scientific Reports*, *10*(1), 14734. https://doi.org/10.1038/s41598-020-71752-x

n.d. (2025). *American Society of Hematology*. https://ashpublications.org/search-results?page=1&q=

Ng, D. P., Simonson, P. D., Tarnok, A., Lucas, F., Kern, W., Rolf, N., Bogdanoski, G., Green, C., Brinkman, R. R., & Czechowska, K. (2024). Recommendations for using artificial intelligence in clinical flow cytometry. *Cytometry. Part B, Clinical Cytometry*, *106*(4), 228–238. https://doi.org/10.1002/cyto.b.22166

Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., Moayedi, Y., Brudno, M., Hope, A., Ross, H., & McIntosh, C. (2024). Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *Npj Digital Medicine*, *7*(1), 124. https://doi.org/10.1038/s41746-024-01118-4

Paranjape, K., Schinkel, M., Hammer, R. D., Schouten, B., Nannan Panday, R. S., Elbers, P. W. G., Kramer, M. H. H., & Nanayakkara, P. (2021). The Value of Artificial Intelligence in Laboratory Medicine. *American Journal of Clinical Pathology*, *155*(6), 823–831. https://doi.org/10.1093/AJCP/AQAA170,

Radakovich, N., Nagy, M., & Nazha, A. (2020). Artificial Intelligence in Hematology: Current Challenges and Opportunities. *Current Hematologic Malignancy Reports*, *15*(3), 203–210. https://doi.org/10.1007/s11899-020-00575-4

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sasada, K., Yamamoto, N., Masuda, H., Tanaka, Y., Ishihara, A., Takamatsu, Y., Yatomi, Y., Katsuda, W., Sato, I., & Matsui, H. (2018). Inter-observer variance and the need for standardization in the morphological classification of myelodysplastic syndrome. *Leukemia Research*, *69*, 54–59. https://doi.org/https://doi.org/10.1016/j.leukres.2018.04.003

Shahzad, M., Ali, F., Shirazi, S. H., Rasheed, A., Ahmad, A., Shah, B., & Kwak, D. (2024). Blood cell image segmentation and classification: a systematic review. *PeerJ. Computer Science*, *10*, e1813. https://doi.org/10.7717/peerj-cs.1813

Shams, U. A., Javed, I., Fizan, M., Shah, A. R., Mustafa, G., Zubair, M., Massoud, Y., Mehmood, M. Q., & Naveed, M. A. (2024). Bio-net dataset: AI-based diagnostic solutions using peripheral blood smear images. *Blood Cells, Molecules, and Diseases*, *105*, 102823. https://doi.org/10.1016/J.BCMD.2024.102823

Su, J., Wang, Y., Zhang, J., Niu, S., Han, J., Xing, L., & Song, J. (2023). ROI-BMC-DNNet: An efficient automatic analysis model of whole-slide scanned bone marrow aspirate images for the diagnosis of hematological disorders. *Biomedical Signal Processing and Control*, *86*, 105243. https://doi.org/https://doi.org/10.1016/j.bspc.2023.105243

Supervisor, E. D. P., Attoresi, M., Bernardo, V., Lareo, X., & Velasco, L. (2023). *TechDispatch – Explainable artificial intelligence. #2/2023* (M. Attoresi, X. Lareo, & L. Velasco, Eds.). Publications Office of the European Union. https://doi.org/doi/10.2804/802043

Tayebi, R. M., Mu, Y., Dehkharghanian, T., Ross, C., Sur, M., Foley, R., Tizhoosh, H. R., & Campbell, C. J. V. (2022). Automated bone marrow cytology using deep learning to generate a histogram of cell types. *Communications Medicine*, *2*, 45. https://doi.org/10.1038/s43856-022-00107-6

Têtu, B., & Hassell, L. A. (2016). *Digital Pathology* (K. J. Kaplan & L. K. F. Rao, Eds.; 1st ed.). Springer Charm.

Wang, M., Dong, C., Gao, Y., Li, J., Han, M., & Wang, L. (2022). A Deep Learning Model for the Automatic Recognition of Aplastic Anemia, Myelodysplastic Syndromes, and Acute Myeloid Leukemia Based on Bone Marrow Smear. *Frontiers in Oncology*, *12*, 844978. https://doi.org/10.3389/fonc.2022.844978

Wang, S.-X., Huang, Z.-F., Li, J., Wu, Y., Du, J., & Li, T. (2024). Optimization of diagnosis and treatment of hematological diseases via artificial intelligence. *Frontiers in Medicine*, *11*, 1487234. https://doi.org/10.3389/fmed.2024.1487234

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Zhao, Y., Diao, Y., Zheng, J., Li, X., & Luan, H. (2024). Performance evaluation of the digital morphology analyser Sysmex DI-60 for white blood cell differentials in abnormal samples. *Scientific Reports*, *14*(1), 14344. https://doi.org/10.1038/s41598-024-65427-0

Machine learning and Models of Artificial Intelligence

The terms machine learning (ML) and artificial intelligence (AI) are often used interchangeably. ML is a subset of AI in which computers can learn from data; it can discover relationships without predefined rules. ML can extract complex relationships out of data which surpass human analytical capacity. Strictly speaking, ML is a component of AI, with AI imitating human cognitive abilities, such as learning, planning, reasoning, understanding language, or problem-solving (S.-X. Wang et al., 2024).

It is often assumed that ML is a new concept, but this is not the case. ML has been around for several decades and is being used in wide range of applications into everyday life (Fein & Shouval, 2021). A well-known example and one of the pioneers of ML is A.L. Samuel. In 1962, Samuel built a checkers playing program that developed its own algorithm based on the moves of past positions and outcomes. He demonstrated that computers could "learn" and adapt, laying the foundations for today's ML and AI (Samuel, 1959).

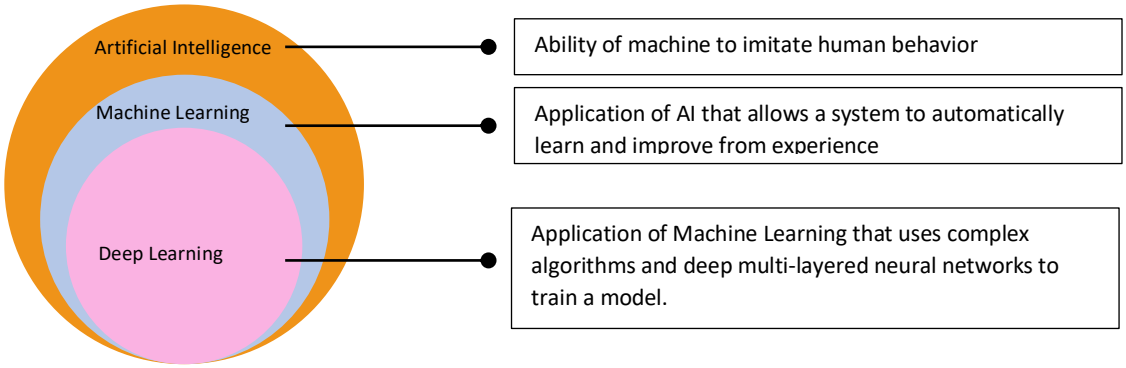

Figure 2, the difference between artificial intelligence, machine learning and deep learning (source: adapted from common AI/ML/DL taxonomy diagrams).

ML should not be seen merely as a computer-driven process, but rather as existing along a "machine learning spectrum," as proposed by Beam and Kohane. This spectrum illustrates a range of models, from those predominantly shaped by human input, like traditional clinical risk scores, to those that rely almost entirely on data, such as deep learning systems. The position of an algorithm on this spectrum depends on the extent of human assumptions encoded before learning begins: the more an algorithm learns independently from data, the further it lies along the data-driven end of the spectrum (Beam & Kohane, 2018). The most advanced type of ML is deep learning. A powerful branch of ML, but it typically requires massive amounts of data to produce reliable predictions. It starts with a single neuron, where inputs x_i are multiplied by corresponding weights w_i summed with a bias term, and then passed through a non-linear transformation function (activation function). This transformation enables the model to capture highly complex patterns. The "deep" in deep learning comes from stacking multiple layers of neurons, where the outputs of one layer serve as inputs for the next. This layered architecture makes deep learning exceptionally powerful, if given enough data, but also difficult to interpret. Weights in hidden layers, assigned to outputs from earlier neuron, generally lack direct meaning. For this reason, deep learning models are often described as black boxes, with their inner workings largely obscure (Fein & Shouval, 2021).

ML can be divided into three types of learning: (I) unsupervised learning, (II) supervised learning, and (III) reinforced learning (China, 2023).

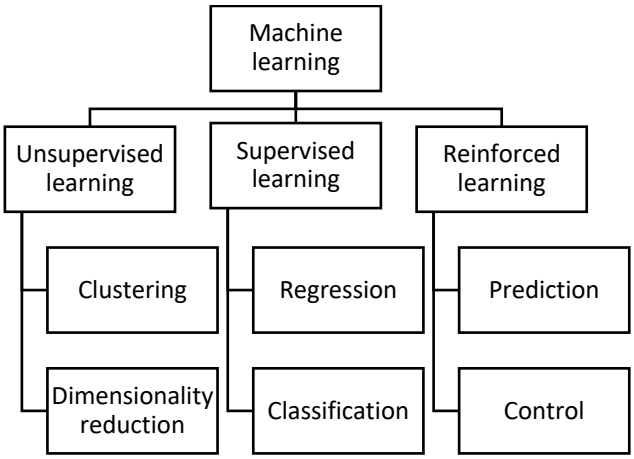

Figure 3, overview of the different learning models of machine learning.

Unsupervised and supervised learning

Unsupervised learning algorithms

Unsupervised learning is a form of data driven ML where patterns are recognized in unlabeled data. This is done by extracting useful patterns from the data through clustering, dimensional reduction, and association tasks (Li, 2024). For example, Monaghan et al. used unsupervised ML models to identify patterns in flow cytometry data by extracting phenotype-level features without the need for manual gating or dimensionality reduction, thereby preserving the full complexity of the original data. These features were then used in a supervised learning model (support vector machine classifier) to distinguish between acute promyelocytic leukemia and other forms of acute leukemia (Monaghan et al., 2022)

Supervised learning algorithms

In supervised learning, a model is trained on a labeled dataset by learning the relationship between input features and corresponding output labels. During training, the algorithm processes input-output pairs to generate a predictive function, optimizing it by minimizing error. Once trained, the model can generalize from this data to accurately predict outputs from previously unseen inputs. This method is often used in classification (e.g. support vector machines, decision trees and random forests) and regression algorithms (e.g. linear and logistic regression). An example of supervised learning in hematology is image recognition, in which the algorithm is trained on thousands of pictures of labeled peripheral blood smears (“training set”), after which the model can make the correct diagnosis based on unseen data (Fein & Shouval, 2021).

There are different supervised learning approaches used in supervised learning algorithms. Commonly used approaches in the medical domain are decision trees, random forests, gradient boosting and convolutional neural networks.

Reinforced learning

In Reinforced learning, an agent learns optimal behaviors through trial-and-error interactions with its environment, guided by rewards or penalties. It is particularly useful for sequential decision-making problems under uncertainty (Tadepalli, Givan, & Driessens, 2004). However, a detailed exploration of reinforcement learning is outside the scope of this critical appraisal.

A. Decision trees

A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It models decisions and their possible consequences as a tree-like structure composed of nodes and branches. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a predicted output (class label or numerical value) (Fürnkranz, 2010).

B. Random forest

A random forest is considered as an ensemble of decision trees. Leo Breiman and Adele Cutler published the first article on random forests in 2001. They developed an algorithm that combined multiple decision trees to generate an average prediction. Each tree is built using a random subset of the data, and at each decision point within a tree, a random subset of the available variables is considered for splitting. This had the major advantage of reducing overfitting (Fein & Shouval, 2021).

C. Gradient boosting

Gradient boosting typically begins with a single decision tree whose predictive accuracy is evaluated. Based on its errors, adjustments are made to create a new tree aimed at reducing those errors. This cycle repeats, with each subsequent tree improving upon the predictions of the previous one. The final output is a weighted average of all trees, where weights reflect each tree's accuracy. Unlike random forests, gradient boosting carries a greater risk of overfitting (Fein & Shouval, 2021).

D. Convolutional neural networks

Convolutional neural networks (CNNs) are a class of deep learning models designed to process data with a grid-like structure, such as images. CNNs consist of multiple layers, including convolutional layers, activation functions, pooling layers, and fully connected layers. In convolutional layers, small filters (or kernels) slide across the input data to detect local patterns such as edges, textures, or shapes. These filters are applied across the entire input using shared weights, producing feature maps that highlight the presence of specific features in various regions.

Following convolutions, non-linear activation functions like ReLU (Rectified Linear Unit) introduce non-linearity into the model, enabling it to learn more complex functions. Pooling layers, typically max pooling, reduce the spatial dimensions of the data, improving computational efficiency and robustness to minor shifts and distortions. After several layers of convolutions and pooling, fully connected layers perform high-level reasoning, such as classification or regression.

One of the key advantages of CNNs over traditional image processing methods is their ability to automatically learn useful features from raw data through training. This is achieved using the backpropagation algorithm, which adjusts the network's weights based on the error between the predicted and actual outputs. This end-to-end learning process enables CNNs to discover hierarchical representations, from low-level patterns to high-level concepts, making them powerful tools in computer vision (LeCun et al., 2015).

In peripheral blood image processing, there are several steps to follow before the data can be fed to the neural network. First there is image acquisition, where the images of blood smears are collected. These images contain various blood cells under different staining and light conditions and hence, need to be preprocessed. Preprocessing is an important step in machine vision applications, directly affecting the performance of ML-based models by emphasizing the useful features of the image. Preprocessing exists of noise reduction (e.g. applying Gaussian filters), decoding and resizing, normalization (e.g. scaling the pixel intensities), segmentation (isolating the cells from the background) and augmentation (increasing the variant of training samples by applying random image transformations). In a second step the CNN architecture will be set up. There is the input of a 2D color image of a blood smear (e.g. 32 x 32 x 3 / length, width, RGB). The convolutional layers detect cell edges, nucleus shape and cytoplasm textures. As stated in previous alinea the pooling layers will down sample the feature maps (Shahzad et al., 2024).

Input data for machine learning

Datasets are split into a training, validation, and test set. The training data set is used to build the model and improve the accuracy of the model by adjusting parameters. Later the validation and test set are used to evaluate the model. The validation test set finetunes the model hyperparameters and the test set evaluates the final model. It is important that the dataset contains as much biologic diversity as possible and hence, reduce the chances of overfitting (Xu & Goodacre, 2018).

Pitfalls of AI

Lack of interpretability

A major challenge in the deployment of artificial intelligence systems is their lack of interpretability, often described as the "black box" effect. Many AI models, particularly those based on machine learning and deep learning, operate in ways that are opaque to their developers, making it difficult to explain how specific decisions or predictions are reached. This opacity can hide underlying issues such as bias, inaccuracies, or discriminatory

outcomes, with serious consequences, especially in medical diagnoses. When users and those affected by AI systems cannot understand the reasoning behind automated outcomes, it risks losing trust, undermining accountability, and limiting the ability to correct unfair decisions (Supervisor et al., 2023).

Overfitting

Overfitting is a common issue in machine learning where a model learns the training data too well, including its noise and random fluctuations, rather than capturing the underlying general patterns. As a result, the model performs very well on the training dataset but fails to generalize to unseen (test or real-world) data. This typically happens when the model is too complex relative to the amount of data, such as having too many parameters, deep trees, or too many training epochs (Domingos, 2012). To overcome overfitting large training data sets and data augmentation can be implemented. Data augmentation involves modifying, for example, the orientation, brightness, or contrast of the training images to create slight variations and thus enhance the variability of the dataset (Ghete et al., 2024).

Data quality

Data quality is essential for AI, as it directly influences the performance, accuracy, and reliability of AI models. High-quality data allows models to make better predictions and yield more reliable outcomes. Addressing biases in data is crucial for ensuring data quality. This prevents the amplification of biases in AI-generated outputs, helping minimize unfair treatment of specific groups or individuals. Furthermore, a diverse and representative dataset enhances an AI model's ability to generalize well across different situations and inputs, ensuring its performance and relevance across various contexts and user groups (Dilmegani, 2025).

Data quantity

ML models often need a lot of data. With more data, models are less likely to overfit to small, unrepresentative samples and can instead learn robust features that improve accuracy and reliability (Lones, 2024).

Class imbalance

In many real-world datasets, some classes (categories) are represented by a large number of examples, while others appear only rarely. If this imbalance is not addressed, a model may become biased toward the majority class, achieving high overall accuracy while performing poorly on the minority class. In medical imaging, such as bone marrow cell classification, this can be critical because minority classes may represent rare but important disease indicators. Without proper handling, these rare but clinically significant cells may be misclassified or even ignored by the model, leading to unreliable diagnostic outcomes (Guo et al., 2022).

Applications of AI models in hematology

AI in Cytomorphology

For many decades, microscopic examination of peripheral blood smears and bone marrow aspirate smears has formed the cornerstone of diagnosing both malignant and non-malignant hematological disorders. Although highly informative, these manual methods depend heavily on the expertise of trained laboratory hematologists and are subject to interobserver variability (Hodes et al., 2019; Lewis et al., 2023). Since the late 1990s and early 2000s, advances in digital microscopy and automated image analysis have aimed to complement traditional light microscopy by enabling high-resolution digital imaging, standardized workflows, and the potential for computer-assisted interpretation (Têtu & Hassell, 2016)

Peripheral blood smears

In the early years basic image processing methods such as preprocessing, segmentation, and object identification were used to extract the desired objects (e.g., blood cells) and feed them into a classification system like a random forest to determine the correct cell type. Initial methods in peripheral blood smears (PBS) were based on narrow tasks such as counting red and white blood cells (Alomari et al., 2014). More modern systems combine deep learning for detection/segmentation with classifiers that produce differentials, flag atypical cells, quantify dysplasia, and even suggest diagnoses, while keeping a human in the loop for verification where needed (Ghaderzadeh et al., 2022; Lewis et al., 2023; Shams et al., 2024).

In clinical laboratories, analyzers such as Sysmex DI-60 and other CellaVision platforms using digital imaging preclassify 100-200 WBCs per peripheral blood smear and present images for verification by laboratory technician or specialist. Recent evaluation of DI-60 on abnormal samples (including acute leukemia and MDS) showed high correlations with manual counts for common categories after verification, very strong agreement for blast detection ($\kappa \approx 0.96$) and weaker performance for plasma cells, underscoring why human review remains essential for certain rarer classes (Zhao et al., 2024). While these analyzers used more traditional computer vision methods, new methods known as deep learning have emerged. Compared with the traditional computer vision methods, in DL algorithms engineers are not required to hard code the feature extraction algorithms and hence, making segmentation algorithms redundant (Khan et al., 2021; Morell, 2025). Acevedo et al. trained a CNN to classify eight cell types in peripheral blood and achieved accuracies over 97% (Acevedo et al., 2019).

One study proposed a multi-level convolutional neural network (ML-CNN) in which in a first level a rapid R-CNN detected regions of interest, namely individual WBCs, and separated them into mononuclear (lymphocytes, monocytes) and polymorphonuclear (neutrophils, eosinophils) cells. In a second level two parallel CNN's classified the subtypes in each group. They achieved 98.4% average performance across the different metrics whereby mononuclear cells were classified with near-perfect accuracy (>99.9%) (Cheuque et al., 2022).

Bone Marrow smears

In contrast to PBS, whole slide bone marrow smears (BMS) are more complex for performing DL algorithms. Heterogenous cell density throughout the slide, overlapping cells and different maturation stages of the cells, make it challenging to apply automatization. When manually assessing a whole slide image (WSI), we identify the appropriate region in the WSI. This region should be free of overlapping cells and artefacts, and preferably located near a marrow fragment to ensure a representative differential count and assessment. The process of manually scanning for the appropriate region needs to be translated into the automation workflow. This brings us to three major challenges for automatization of digital BMS analyzes. The first challenge is region of interest (ROI) detection, where suitable regions or tiles must be selected from WSI before analyzing cells. The second challenge is object detection, which involves identifying individual bone marrow cells or non-cellular structures as distinct from the background. Deep learning models such as R-CNN, Fast R-CNN, and Faster R-CNN have been used for this, but they rely on region proposals and separate classification steps, making them complex and computationally inefficient. The third challenge is object classification, where detected cells or objects must be categorized into specific classes based on subtle cytological features, a task made even more difficult in conditions like MDS due to morphological dysplasia (Tayebi et al., 2022).

Su et al. developed a ROI-BMC-DNNet, a deep learning model that automatically identifies high-quality regions in whole-slide bone marrow aspirate images and accurately performs nucleated cell differential counting. The segmentation model achieved over 92% precision in patch retrieval and about 87–91% recall/precision in cell

detection (Su et al., 2023). Tayebi et al. used another approach to identify ROI's. They identified ROIs by splitting WSIs into tiles, having experts label them, then training a DenseNet121 CNN to automatically classify tiles as ROI or non-ROI. The model achieved 97% accuracy, 90% precision, 99% specificity and 78% recall. Identifying suitable regions within the BMS for differential cell counting is a critical step, as it ensures accurate cytologic and morphologic evaluation while minimizing diagnostic variability. It is expected that the evolution of artificial intelligence will play an increasingly vital role in diagnostics in the coming years. Future research will determine how the applications of AI will further develop within laboratory diagnostics.

AI in flow cytometry

Flow cytometry plays a crucial role in the diagnostic workup and follow-up of various hematological malignancies.

Diagnosis

Cheng et al. explores the use of deep learning to support acute leukemia detection and cell classification in flow cytometry using the EuroFlow acute leukemia orientation tube (ALOT). Researchers retrospectively analyzed flow cytometry data from 241 patients (2017–2022) and trained AI models, including ResNet-50 and a custom architecture called EverFlow. Instead of using raw data, the researchers use image analyzing models to achieve their goal. The model achieved high sensitivity in detecting acute myeloid leukemia (94.6%) and B-lymphoblastic leukemia (98.2%), while also showing strong performance in classifying physiological cells. Although challenges remain in identifying CD34-negative pathological cells and complex cases like myelodysplasia, the AI significantly reduced analysis time compared to manual gating. Overall, the findings demonstrate that deep learning can enhance the accuracy, speed, and reproducibility of leukemia diagnosis via flow cytometry, though larger datasets and further refinement are needed for broader clinical application (Cheng et al., 2024).

Limitations

Flow cytometry AI-models face significant limitations in reproducibility across laboratories because of technical and procedural variability. Differences in cytometer platforms, antibody panels, gating strategies, and specimen preparation introduce inconsistencies, while batch effects from reagent lots or machine calibration affect results. In addition, the lack of universally adopted data standardization methods makes it difficult to harmonize outputs, and AI models trained on local datasets often fail to generalize to external cohorts. laboratories (Ng et al., 2024).

MDS as prototype disease in the AI evolution

According to the World Health Organization MDS is defined as a clonal hematopoietic stem cell neoplasm characterized by cytopenia, morphological dysplasia and by progressively ineffective hematopoiesis with an increased risk of acute myeloid leukemia. MDS can be classified into two major entities, each with specific subtypes: those with genetic defining abnormalities and those with morphological defining abnormalities (Alaggio et al., 2022). Since initial diagnosis of MDS still relies on morphological features, the establishment of a uniform diagnostic criterium of cell morphology is essential. A study by Sasada et al. (Sasada et al., 2018) found significant variability among observers in diagnosing MDS based on cell morphology, highlighting the need for standardization and machine learning based diagnostic support systems.

In this CAT I performed a comprehensive review to find an answer to the following research question: "What is the diagnostic accuracy of AI-based image analysis tools in identifying myelodysplastic syndromes compared to conventional morphological evaluation?"

Pubmed and Embase were searched. Search terms were ("Myelodysplastic Syndromes"[MeSH Terms] OR "Myelodysplastic Syndromes"[Title/Abstract] OR "myelodysplastic syndrome") AND ("Artificial Intelligence"[MeSH Terms] OR "Artificial Intelligence"[Title/Abstract] OR "Machine Learning"[Title/Abstract] OR "Deep Learning"[Title/Abstract]). Publication date has been set for the last ten years.

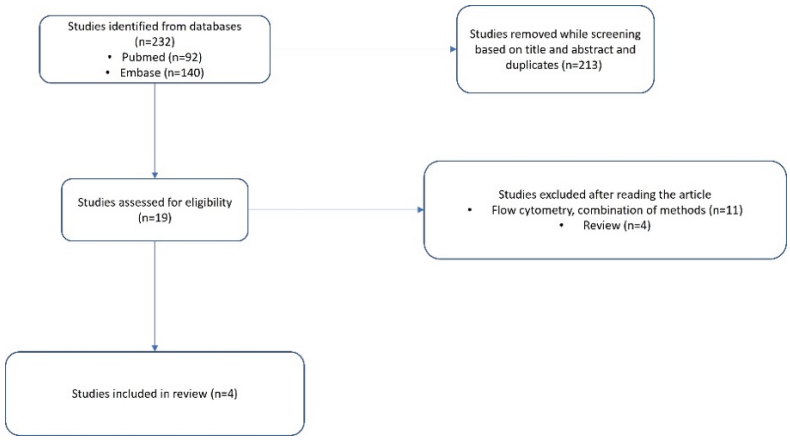

Figure 4, The search strategy followed the "PIRT" framework, focusing on patients undergoing morphological evaluation for MDS, with the intervention being the use of AI-driven morphology tools, assessing outcomes in terms of diagnostic performance metrics such as sensitivity, specificity, and accuracy, within the timeframe of studies published between 2015 and 2025.

After exclusion, four studies were included in the review and described in table Attachment I.

Results

Wang et al. developed a deep learning model using a ResNet-50 convolutional neural network pretrained on ImageNet to automatically classify bone marrow smear images into aplastic anemia (AA),MDS, or acute myeloid leukemia (AML). They trained the model on augmented images from the ASH Image Bank and validated it on clinical data from a hospital. The model was optimized with an outcome weight of 1:9 (favoring three-classification) and 200 training epochs, achieving high performance in both binary (MDS vs. others) and three-way classification tasks, with AUCs of 0.985 and 0.968, and testing accuracies of 91.4% and 92.9%, respectively. The system was able to successfully differentiate between MDS/non-MDS and aplastic anemia, MDS and AML (M. Wang et al., 2022).

Mori et al. developed an artificial intelligence system to detect decreased granules (DG) in neutrophils from bone marrow smears. Using deep learning with transfer learning from ImageNet on a ResNet-152 architecture, the system was trained on 1,797 labelled cell images from patients with myelodysplastic syndromes (MDS) and other hematologic diseases. To improve accuracy and handle imbalanced data, they implemented data augmentation and a “Doctor-in-the-loop” iterative correction process. The final model achieved high diagnostic performance with 97.2% accuracy and 0.944 AUC when including all DG grades, and even higher accuracy (98.2%) when excluding ambiguous DG1 cases, showing its potential to support clinical diagnosis of MDS (Mori et al., 2020).

Lee et al. developed a deep learning-based algorithm to automatically classify hematopoietic cells and detect dysplastic features in bone marrow aspiration smears from patients with myelodysplastic syndromes (MDS). Using whole-slide images from 34 MDS patients and 24 controls, they manually labeled and segmented 8065 cells into eight categories across three major lineages. Training a convolutional neural network (InceptionV3), they achieved high classification performance, with area under the curve (AUC) values ranging from 0.945 to 0.996 and accuracy between 91.2% and 99.3%. While the model showed excellent specificity, sensitivity varied across cell types (Lee et al., 2022).

The study by Acevedo et al. developed a convolutional neural network model, named DysplasiaNet, to automatically identify hypogranulated neutrophils, an important dysplastic feature in peripheral blood smears of patients with MDS. They trained and evaluated eight CNN architectures on a dataset of over 20,000 images and selected the best-performing model based on diagnostic accuracy. The final model achieved 95.5% sensitivity, 94.3% specificity, and 94.85% overall accuracy in a proof-of-concept test using previously unseen patient data (Acevedo et al., 2021).

In conclusion, these studies show that deep learning models can be very useful for diagnosing blood disorders, especially myelodysplastic syndromes. Using different CNN architectures, the systems were able to classify cells and detect abnormal features with high accuracy. Methods like transfer learning, data augmentation, and expert feedback made the models more reliable. Although some challenges remain, such as differences in performance between cell types and unclear cases, the results suggest that AI can support doctors by making bone marrow and blood smear analysis faster and more accurate in clinical practice.

Challenges of Artificial Intelligence in today's hematology laboratories

The implementation of ML models will assist clinicians in interpreting data, increasing objectivity, and improving diagnostic accuracy. Although it is becoming clear that AI will become an integral component of laboratory diagnostics, significant challenges remain. A first challenge is the limited flexibility of AI in interpreting laboratory data. Unlike human intelligence, which can make intuitive decisions in combination with a contextual framework, this is not the case within AI systems. They rely solely on learned patterns, which can pose risks when contextual information is lacking. These risks may be reduced by designing ML systems to support laboratorians through recommendations, rather than replacing the laboratorian or clinician (Herman et al., 2021).

Secondly, many DL algorithms require a large amount of data but this sensitive clinical patient data is of not publicly available or health systems are not eager to share this data due to privacy concerns (Herman et al., 2021). This can be noted by the numerous published studies (cfr. MDS paragraph) that only rely on single-centered datasets or lack external validation datasets, which severely limits the generalizability and reproducibility of the model (Ong Ly et al., 2024). Future collaboration between different institutions and the availability of large amounts of "clean" or annotated data could help resolve this issue.

A third challenge is explainability. As mentioned earlier, many algorithms (e.g., CNNs) produce mathematically very complex yet accurate computations and generate an output, while no one truly understands how or why the algorithm arrived at this prediction. Although these systems often surpass physicians in diagnostic accuracy, the lack of explainability can give rise to ethical dilemmas. A patient may lose autonomy when the clinical reasoning behind a medical decision disappears (Dhar et al., 2023)

A fourth challenge is the regulation of AI systems. The success of AI within laboratory diagnostics will largely depend on the approval and regulation of these systems. AI models can be considered medical devices and must therefore meet the highest quality standards before they may be used in healthcare. In Europe, the EU AI Act, MDR/IVDR for AI applications, and GDPR legislation exist, with the future showing whether this legal framework will facilitate or hinder the use of AI in healthcare (Aboy et al., 2024).

TO DO/ACTIONS

1) Establish multi-center, standardized datasets: future implementation of AI in hematology requires the creation of large, multi-center datasets that are well-annotated and standardized. This will reduce bias, improve external validation, and enhance the generalizability of AI models across different laboratories and patient populations
2) Improve interpretability and explainability of AI models: AI tools should be developed with built-in explainability features such as Grad-CAM, heatmaps, or feature importance visualization. Increasing transparency will reduce the "black box" effect, increase clinical trust, and support AI as a decision-support tool rather than a replacement for hematologists.
3) Develop regulatory and clinical integration frameworks: clear frameworks for validation, regulatory approval, and integration into clinical workflows are essential. Collaboration with regulators, clinical biologists, and IT specialists will ensure compliance with the EU AI Act, MDR/IVDR, and GDPR, while pilot studies in routine labs can evaluate practical implementation.
4) Structured education programs should be developed for laboratory staff, clinicians, and medical students to improve AI literacy. Workshops, e-learning modules, and integration into professional training curricula will help reduce adoption barriers, ensure correct use, and prepare the laboratory for AI-supported diagnostics.

ATTACHMENTS

Attachment 1

Study	Year	BM dataset	Model	Outcome	Key results	Advantages	Disadvantages
Wang M. et al.	2022	115 BMS from the ASH Image Bank (32 MDS, 26 AA, 57 AML) 432 BMS from clinical data for external validation (214 MDS, 115 AA, 103 AML)	ResNet-50 (CNN)	Distinguishing MDS from AA and AML	**MDS vs. non-MDS** AUC 0.985 (testing) AUC 0.942 (validation) Accuracy 91.4% (testing), Accuracy 92.1% (validation) Sens 99.2% (testing) Spec 88.1% (testing) Spec 93.8% (validation) **Three-class (AA vs. MDS vs. AML)** AUC 0.968 (testing) AUC 0.948 (validation) Accuracy 92.9% (testing), Accuracy 91.5% (validation) Sens 85.7% (testing) Sens 88.7% (validation)	Excellent performance metrics Internally and externally validated	Requires clinician assistance

Mori J. et al.	2020	35 BMS with MDS 1797 labels single-cell images	Faster R-CNN with ResNet-101 ResNet-152	Decreased granules in neutrophils (representative for dysplasia) for recognizing MDS	**Mild – moderate/severe dysplasia** AUC 0.944 - 0.921 Accuracy 97.2% - 98.2% Sens 91.0% - 85.2% Spec 97.7% - 98.9% PPV 76.3% - 80.6% NPV 99.3% - 99.2%	High accuracy Efficient design Open-source code	Limited dataset (only DG). Not generalizable to all dysplasia. Small sample size (two morphologists labeled data: bias/subjectivity) Imbalanced Data: severe dysplasia had n = 11 samples Misclassification DG versus IG No external validation (independent morphologists)
Lee N. et al.	2022	34 BMS with MDS 24 normal BM slides 8065 single cells in 8 categories	U-Net segmentation InceptionV3 CNN classification	Automatically detecting and classifying dysplastic cells in bone marrow aspiration smears from patients with myelodysplastic syndromes (MDS).	**Dysplastic granulocytes (GD)** AUC 0.996 Sens 90% Spec 99.9% **Erytroid dysplasia** AUC ? Sens 79% Spec 99.2% **Megakaryocytic dysplasia** AUC ? Sens 89.9% Spec 94.8%	High classification accuracy Eight distinct categories, more robust algorithm Grad-CAM heatmap to visualize the algorithm's focus area, increasing transparency Open-source code	Imbalanced data which can bias performance metrics and reduce generalizability Limited dataset size

Acevedo A. et al.	2021	20670 images of normal and dysplastic neutrophils	8 own CNN models were trained and best performing model was selected: DysplasiaNet	Automatically recognize hypogranulated neutrophils in peripheral blood smears as key marker of MDS	**Unseen test set** AUC 0.982 Sens 95.5% Spec 94.3% PPV 94% Accuracy 94.85%	High diagnostic accuracy Limited number of parameters (+/- 73000) It's resource-efficient and doesn't require GPU. Suitable for clinical integration. Built from scratch for the specific task using a custom dataset. This enhances the performance of the model Dataset of 20670 annotated images which improves the robustness of the model. Different methods were used to show how the model distinguishes between dysplastic and normal neutrophils, increasing clinical trust.	Limited scope of dysplasia features (only highlights hypogranulation) Limited to retrospective data. Lack of multi-center data which may limit generalizability to different lab settings or populations. Binary classification only. No grading of severity was implemented or addressing other possible causes of hypogranulation

Abbreviations

BMS: Bone marrow smear
ASH: American Society of Hematology
MDS Myelodysplastic syndrome
AA: Aplastic Anemia
AML: Acute Myeloid Leukemia
AUC: Area Under the Curve

PPV: Positive Predictive Value
Grad-CAM: Gradient-weighted Class Activation Mapping