Supplementary

This supplementary file provides comprehensive information on:

1. Challenges encountered during data collection, image processing and augmentation techniques employed.
2. Evaluation metrics used to evaluate the models and pipelines.
3. AI pipeline Predictions Compared to Pathologists (Numerical values and Tables)
4. Strengths and limitations of the proposed pipelines.
5. Recommendations for future research.

Appendix 1 Challenges during data collection, image preprocessing, and augmentations applied

The image capturing process for each slide was time-consuming, ranging from 30 to 40 minutes, as it involved meticulous adjustments of the slide on the microscope table, focusing the microscope, and capturing images for proper storage. While capturing images, high-power fields (HPFs) lacking nucleated WBCs were either avoided or excluded. Additionally, HPFs with poor stain quality or artifacts were not captured. The number of images collected from each peripheral blood smear (PBS) varied significantly, ranging from 3 to 150. This variation is likely due to the fluctuating white blood cell (WBC) count in peripheral blood, which can be influenced by both the type of leukemia and the impact of leukemic cells on the bone marrow. For chronic leukemia cases that transformed into acute leukemia, the corresponding PBS images were classified as acute leukemia.

All images were initially reshaped into squares by padding the shorter sides with zeros. Thereafter, they were down sampled to a size of 448x448 using bilinear interpolation. During training, data augmentation techniques were applied to prevent overfitting. These transformations included random flipping (horizontal and vertical), Gaussian blurring with a kernel size of 3, random scaling (between 60-100% of the image size), light colour jittering, and random sharpening.

Appendix 2 Evaluation metrics

Positive predictive value/precision:

$$PVV = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Sensitivity/recall:

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Specificity:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

F1 score:

$$F1 = \frac{True\ Positives}{True\ Positives + \frac{1}{2}\left(False\ Positives + False\ Negatives\right)}$$

Accuracy

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

Cohen's Kappa score: It measures the level of agreement between two raters or judges who each classify items into mutually exclusive categories

$$Cohen's\ Kappa\ score = \frac{p_o - p_e}{1 - p_e}$$

https://dx.doi.org/10.21037/jmai-24-74

where:

- p_o is the relative observed agreement among raters
- p_e is the hypothetical probability of chance agreement

  The value of Cohen's Kappa ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance, and negative values indicate disagreement.

Appendix 3 AI pipeline predictions compared to pathologists (numerical values and Tables)

While detailed performance metrics are visualized in figures within the main manuscript, Table S1 and S in the supplementary data provide the exact numerical values.

Three-class prediction compared to pathologists (Table S1)

Our AI pipeline achieved comparable performance to pathologists in a three-class prediction task, demonstrating high accuracy (92%). It is important to note that individual pathologists also achieved high accuracy, ranging from 91% to 96%. While the table provides specific values, further insights and visualizations of these results are presented within the main manuscript.

Table S1 Three-class prediction pipeline results: performance metrics of AI pipeline versus pathologists

Evaluator	Precision/PPV	Sensitivity/recall	F1	Accuracy	Specificity
AI	0.95	0.94	0.94	0.92	0.95
P1-1	0.86	0.91	0.88	0.92	0.96
P1-2	0.84	0.86	0.85	0.91	0.95
P2	0.88	0.93	0.90	0.91	0.95
P3	0.85	0.93	0.88	0.91	0.96

AI, artificial intelligence; P1-P3, pathologists 1, 2, 3, respectively; P1-1, pathologist 1, first evaluation; P1-2, pathologist 1, second evaluation.

Five-class prediction compared to pathologists (Table S2)

This table compares the performance of our AI model to individual pathologists in terms of precision (positive predictive value). While the table provides specific values, further insights and visualizations of these results are presented within the main manuscript.

Table S2 Five-class prediction pipeline results: performance metrics of AI pipeline versus pathologists

Evaluator	Precision/PPV	Sensitivity/recall	F1	Accuracy	Specificity
AI	0.80	0.80	0.77	0.70	0.93
P1-1	0.76	0.76	0.72	0.86	0.96
P1-2	0.73	0.71	0.71	0.84	0.96
P2	0.81	0.88	0.83	0.86	0.96
P3	0.75	0.84	0.78	0.86	0.97

AI, artificial intelligence; P1-P3, pathologists 1, 2, 3, respectively; P1-1, pathologist 1, first evaluation; P1-2, pathologist 1, second evaluation.

https://dx.doi.org/10.21037/jmai-24-74

Appendix 4 Strengths and limitations of the proposed AI pipelines

Our AI pipeline leveraged several strengths to achieve promising results. However, limitations remain, highlighting the need for further research, particularly in addressing the generalizability of the model to rarer leukemia types and diverse patient populations, as presented in Table S3.

Table S3 Strengths and limitations of the ai pipelines for leukemia classification

Strengths	Limitations
1. Largest leukemia dataset with the four most common clinically relevant types of leukemia (AML, ALL, CLL, and CML)	1. Uncertain criteria for reactive cell identification, which may have led to the misclassification of some images in the training data
2. De-identified images of peripheral blood smear used instead of bone marrow aspirate, reducing the invasiveness and increasing the reach	2. No testing on post-chemotherapy and relapsed cases, which may have altered morphological features
3. Consistent method for acquiring all images and reproducible dataset, facilitating the replication and validation of the method	3. No testing on patients with end-stage organ failure, which may affect the quality and interpretation of the peripheral blood smear images
4. Prediction based on holistic analysis of whole slides mimicking human experts, which improves accuracy and reliability of the diagnosis	4. Only trained and tested on four common types of leukemia, and may not be able to classify rare or complex cases, such as transformed leukemia or myelodysplastic syndrome
5. External validation of the methods using new dataset, demonstrating the generalizability and robustness of the method	5. Single-center training and testing, which may limit the diversity and representativeness of the data
6. Generalizable and scalable model, which can be applied to other types of leukemia or blood disorders with modifications	

Appendix 5 Future research directions

Beyond the current work, future research avenues should aim to further refine and enhance the development of clinically applicable ML models for leukemia classification. This includes:

1. Automation and standardization: implement automated image acquisition and loading pipelines to seamlessly integrate with the prediction model. This will enhance adaptability and promote standardization in image quality.
2. Scalability and disease coverage: investigate the inclusion of additional haematological disorders such as iron deficiency anaemia, megaloblastic anaemia, beta thalassemia, sickle cell disease, and haemolytic anaemia.
3. Multimodal data integration: explore the integration of multi-modal data, including clinical information, molecular profiles, and genetic test results, to augment model performance and build trust in its predictions.
4. Cell segmentation and enumeration: incorporation of cell segmentation algorithms to quantify specific cell populations of interest. This can refine model performance in challenging cases and facilitate the inclusion of diseases like myelodysplastic syndromes, where cell counts are crucial for diagnosis.
5. Generalizability: conduct multicentre studies to evaluate the AI's performance on external datasets from diverse geographic and clinical settings. This will ensure real-world applicability across various healthcare systems.
6. Model evolution and refinement: design systems where ML models can continuously learn and evolve alongside the emergence of new leukemia subtypes or variants.