Novel hierarchical deep learning models predict type of leukemia from whole slide microscopic images of peripheral blood

Novel hierarchical deep learning models predict type of leukemia from whole slide microscopic images of peripheral blood

Naveed Syed[1]^, Mohamed Eltag Salih Saeed[2], Shakir Hussain[3], Imran Mirza[3], Amira Mahmoud Abdalla[3], Eiman Ahmed Al Zaabi[3], Imrana Afrooz[4], Shahrukh Hashmi[2,5], Mohammad Yaqub[2]

[1]Hematology & Oncology Department, Sheikh Shakbout Medical City, Abu Dhabi, UAE; [2]Mohamed Bin Zayed University of Artificial Intelligence, UAE; [3]Hematopathology Department, Sheikh Shakbout Medical City, Abu Dhabi, UAE; [4]Clinical research department, Sheikh Shakbout Medical City, Abu Dhabi, UAE; [5]Division of Hematology, Department of Medicine, Mayo Clinic, Rochester, MN, USA

Contributions: (I) Conception and design: N Syed, S Hashmi, M Yaqub; (II) Administrative support: EA Al Zaabi, S Hussain, S Hashmi; (III) Provision of study materials or patients: N Syed, I Mirza, S Hussain; (IV) Collection and assembly of data: N Syed, AM Abdalla, I Afrooz, S Hussain; (V) Data analysis and interpretation: N Syed, MES Saeed, M Yaqub; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Naveed Syed, DNB. Hematology & Oncology Department, Shiekh Shakbout Medical City, 7th Floor, Tower D, PO Box: 11001 Abu Dhabi, United Arab Emirates. Email: naveed3642003@gmail.com.

Background: Diagnosis of leukemia relies on microscopic examination of suspected peripheral blood smears (PBS) by pathologists. However, this challenging and potentially error-prone process is further limited by the availability of pathologists. Developing a clinically applicable predictive artificial intelligence (AI) model offers a potential solution to these limitations. This study aimed to develop an AI model that predicts leukemia from microscopic images of peripheral blood smear.

Methods: We developed hierarchical deep learning (DL) pipelines that were trained and validated on 7,255 high-power field (HPF) microscopy images collected from pre-diagnosed cases at our centre between January 2021 to August 2023. The proposed pipelines employed a high performing binary DL model in step 1 and multiclass DL models in step 2 to predict the label of each image from a slide, and then aggregate the results to provide a patient-level diagnostic prediction. This approach mimics the examination of PBS by trained hematopathologists, who examine multiple HPF patches under a microscope to predict the type of leukemia. The pipelines' performance was assessed using F1 score, sensitivity, specificity, and accuracy, then compared against the three pathologists using Cohen's kappa score.

Results: The 3-class pipeline categorizes the slide as acute leukemia, chronic leukemia, or no leukemia. The F1 score, sensitivity, specificity, and accuracy were 94%, 94%, 95%, and 92% respectively. This pipeline outperformed the pathologists in all metrics. The agreement between this pipeline and pathologists ranged between 74–77%, while the agreement amongst the pathologists ranged from 82–87%. The 5-class pipeline categorizes the slide into acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), chronic myeloid leukemia (CML), chronic lymphoblastic leukemia (CLL), and no leukemia. It achieved an F1 score of 77%, sensitivity of 80%, specificity of 94%, and accuracy of 70%. The 5-class pipeline outperformed pathologists in the CML and CLL classes but was inferior in other classes. The agreement between this pipeline and pathologists ranged between 52–53%, while the agreement amongst the pathologists ranged from 72–79%.

Conclusions: Acute and chronic leukemia can be predicted with high sensitivity and specificity from whole-slide microscopic images of PBS by utilizing three deep-learning models in a hierarchical pattern and aggregating the results. This approach can efficiently predict and differentiate CLL from CML, but not ALL from AML. Applying cell segmentation in this level of the pipeline could improve the prediction in these classes. These pipelines can be applied in limited infrastructure settings to potentially identify types of leukemia with greater reliability.

^ ORCID: 0000-0002-8172-8129.

Keywords: Acute myeloid leukemia (AML); acute lymphoblastic leukemia (ALL); chronic myeloid leukemia (CML); chronic lymphocytic leukemia (CLL); deep learning (DL)

Received: 17 March 2024; Accepted: 21 August 2024; Published online: 14 October 2024.
doi: 10.21037/jmai-24-74
View this article at: https://dx.doi.org/10.21037/jmai-24-74

Introduction

Leukemia, a cancer of blood-forming cells, affected an estimated 490,875 people in the United States alone in 2020. The rate of new cases was 14.0 per 100,000 men and women per year (1). Leukemia is typically suspected when an abnormal white blood cell counts, or cancerous cells (blast cells) are detected in circulating blood. Modern haematology analysers with artificial intelligence (AI) capabilities efficiently flag samples with abnormal cell counts or unidentified cells. These flagged samples then undergo manual microscopic examination, which is critical for leukemia diagnosis (2). While present automated analysers help reduce the number of samples requiring manual examination, they have limitations. These include inconsistencies in identifying immature granulocytes, reactive lymphocytes, blast cells (cancer cells), and undefined cells. Additionally, their high cost and maintenance requirements restrict their use to high-volume laboratories in developed regions. In resource-limited settings, labs often use low-cost analysers with fewer features, leading to an increased reliance on manual microscopic exams. This, coupled with the globally rising workload for pathologists due to an aging population and increasing demand for complex diagnostic tests, creates a significant gap between demand and workforce (3). This shortage can lead to delayed diagnoses, especially in cases of leukemia, and impact access to critical pathology services, ultimately affecting patient care.

Highlight box

Key findings

- This study developed high-accuracy (>90%) deep learning models for leukemia diagnosis from blood smears. A 3-class model (acute, chronic, no leukemia) showed comparable performance to expert hematopathologists in all metrics (sensitivity, specificity, accuracy, F1 score). While agreement with pathologists was moderate, a 5-class model excelled at identifying specific chronic leukemia subtypes (chronic myeloid leukemia & chronic lymphocytic leukemia). Further research is needed to improve differentiation of acute subtypes (acute lymphoblastic leukemia *vs.* acute myeloid leukemia).

What is known and what is new?

- Currently, blood smear analysis by pathologists is the initial step for diagnosing leukemia subtypes. However, this method is limited by availability of trained pathologists and resources. This study presents deep learning pipelines as a promising alternative for high accuracy leukemia diagnosis from blood smears. These models perform comparably to pathologists for broad leukemia categories and show potential for identifying specific chronic subtypes.

What is the implication, and what should change now?

- Artificial intelligence (AI)-powered pipelines can be a valuable tool, especially in resource-limited settings. They have the potential to improve diagnostic accuracy and consistency. Future research will focus on enhancing differentiation of acute subtypes and validating the pipeline through clinical trials. Ultimately, these models could be integrated into AI-assisted or human-in-the-loop designs to benefit patients and physicians, particularly in resource-limited areas.

Machine learning models have shown significant potential in analysing images to predict the type of leukemia. The majority of the machine learning models reported were trained on binary classification tasks to identify leukemia versus no leukemia, achieving an accuracy greater than 95% (4-7). These models had limited clinical application because they were trained and tested on small, publicly available, homogenous datasets, which are not representative of the clinical experience (6,8). The publicly available datasets used for training included acute lymphoblastic leukemia-international database (ALL-IDB) leukemia image repository (4), American Society of Haematology (ASH) (5), and Google images (9). No external validation was done in the majority of studies, which renders the reliability and generalizability of these models questionable (10,11). Image augmentation methods partially address the issue of small datasets by generating new images, but they do not capture the wide range of variability observed in leukemia types (12-15). Multiple disease-related and approach-related challenges hampered the application of ML models (16).

In this paper, we address the following research question: Can clinically applicable AI models be developed that can accurately predict the type of leukemia from peripheral blood smear (PBS) images? To answer this, we manually collected images and developed models that mimic the pathologist's evaluation of smears. We adopted the hierarchical application of multiple models from our earlier work, which is an image-based seven-class prediction model (17). To authenticate the present strategy, we performed external validation using a different dataset and compared the AI performance with three trained pathologists. We also present the inter-observer agreement on the predictions between AI versus pathologists. We present this article in accordance with the TRIPOD reporting checklist (available at https://jmai.amegroups.com/article/view/10.21037/jmai-24-74/rc).

Methods

Dataset acquisition, training/validation set composition, image processing

The data acquired consisted of three image datasets collected over a period of two and a half years, from January 2021 till August 2023. High power field images (HPF: 1,000× magnification, oil immersion) were captured using a digital camera (Model DP25/73/74, Olympus, Japan) attached to a microscope (Model, BX53, Olympus, Japan). The PBS were prepared using May-Grünwald Giemsa stain by automated slide preparation system (DI60, Sysmex Corporation, Japan). All leukemia PBS utilized in this study belonged to adult patients diagnosed and treated at our hospital. The PBS for the 'No Leukemia' class were selected from random normal samples and prepared similarly. All HPF images were collected manually by moving the slide in a 'Z' pattern under the microscope, capturing almost all the fields which contained nucleated cells [white blood cells (WBCs)]. This study was conducted in accordance with the principles of the Declaration of Helsinki (as revised in 2013). Approval was granted by the Ethics Committee of Sheikh Shakbout Medical City (Date. 23-03-2022/No. MAFREC-273). Informed consent was not applicable due to the retrospective nature of the study using anonymized data and minimal risk to subjects.

Training set 1

It consisted of 3,422 images obtained from 70 patients. These were sorted into seven groups, each representing a class: acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), acute promyelocytic leukemia (APML), chronic myeloid leukemia (CML), chronic lymphocytic leukemia (CLL), reactive cells, and no leukemia. All HPF images were collected and sorted into the specific class category. Images with no leukemic cells were sorted into the 'No Leukemia' class, and images with reactive cells were sorted into the 'Reactive Cells' class. Moreover, images of poor quality were removed. Two pathologists conducted this sorting. In patients with profound leukopenia, PBS with less than three images were excluded. Any duplicates were removed. The images varied in resolution, with the majority having a resolution of 1,200×1,600 pixels. Training and internal validation sets were created by splitting the data per slide to prevent any leakage, i.e., images from the same patient were only used in training or internal validation, but not both.

Training set 2

This training set was introduced in the later stages of experimentation to decrease the gap between the training and external validation data. This gap arose due to the changes in capture devices and protocols, leading to data drift. Images in this dataset were collected using DP25 and DP74 camera, unlike DP73 in training set 1. Some images were captured by a new technician, with altered light conditions and resolution. A new pathologist joined here who helped in sorting and guiding the trainee. This dataset consisted of 860 new images from 28 patients, along with the original 3,422 images from 70 patients. The final dataset, therefore, consisted of 4,282 images from 98 patients. It was split in an 80:20 ratio per patient for the training and testing of the AI models. *Table 1* presents the precise distribution of images across classes for the training, internal validation, and external validation sets.

External validation set

This dataset consists of 2,973 images from 64 patients (AML: 22, ALL: 7, CML: 4, CLL: 3, no leukemia: 28). The images are further categorized by class and detailed in *Table 1*: AML (1,252 images), ALL (346 images), CML (226 images), CLL (225 images), and normal (924 images). This was not included in model training or shared with the team involved in AI pipeline development. This meticulous process and multiple calibrations ensured the integrity of our image classifications. Subsequently, we implemented a stringent anonymization protocol to protect patient privacy. This protocol involved removing all identifiable patient information, thereby providing a clean, accurate, and confidential dataset.

Table 1 Distribution of images and patients across classes in training, internal validation, and external validation sets

Split	Type	ALL	AML	CLL	CML	No leukemia
Training	Images	471	1,149	440	404	938
	Patients	9	27	7	8	27
Internal validation	Images	20	174	191	151	344
	Patients	2	7	2	2	7
External validation	Images	346	1,252	225	226	924
	Patients	7	22	3	4	28

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphoblastic leukemia; CML, chronic myeloid leukemia.

The classes in this set were reduced to five (ALL, AML, CLL, CML, and no leukemia), given that APML is a subset of AML. In addition, reactive cells were merged with normal cells since they are not a leukemia subtype and can be found more commonly in normal than in leukemia smears. The AI pipelines were evaluated on this data at two levels: a 3-class setting (acute, chronic, and no leukemia) and a 5-class setting (ALL, AML, CLL, CML, and no leukemia).

Image preprocessing, augmentations and experimental setup

All images were initially reshaped into squares by padding the shorter sides with zeros. Thereafter, they were down sampled to a size of 448×448 using bilinear interpolation. During training, data augmentation techniques were applied to prevent overfitting. For further details on image processing, addressing unique challenges during data collection process and augmentation techniques applied, please refer to the Supplementary file (Appendix 1).

Statistical analysis

The performance of the pipelines and models was evaluated using positive predictive value (PPV), F1 score, sensitivity, specificity, precision, and accuracy. These metrics were aggregated on a macro level to treat all classes equally despite their imbalance. The key metric used to select the best model during validation was the F1-score. This metric was chosen because it considers class imbalance and provides a good general indication of the models' performance. Cohen's Kappa score was used to evaluate the agreement between AI pipelines and predictions. The value of Cohen's Kappa ranged from −1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance, and negative values indicate disagreement. For detailed description of each evaluation metric, please refer to the Supplementary file (Appendix 2).

DL models, classification pipelines, confusion matrices and validation

We used a deep learning (DL)-based approach to classify leukemia types from blood smear images. To do this, the DenseNet-121 (18) architecture was chosen, as it has shown relatively better results compared to other models in our experiments. It builds upon the well-known ResNet (7) architecture by adding dense connections. These connections feed inputs within a block to all subsequent layers within that block, providing more information to each layer and allowing the network to become more efficient with fewer channels needed.

Initially, the models were used to classify individual images (patches) independently. However, this was ineffective and counterintuitive to the diagnostic process, where a patient is given a single diagnosis based on a holistic examination of their images. Hence, a patient-level approach was developed. This was a multi-model approach, with the input still being individual ungrouped images. The outcomes of the different models were then aggregated to generate a patient-level diagnosis.

First, a model was trained for binary classification (leukemia *vs.* normal) at the first stage of the pipeline. Thereafter, two models were trained to further classify the type of leukemia in a multi-class setting, as shown in *Figures 1,2*. The first of these separated the leukemia class into acute and chronic, while the latter further separated the leukemia classes into ALL, AML, CLL & CML. The reason for this multi-stage approach is that the binary

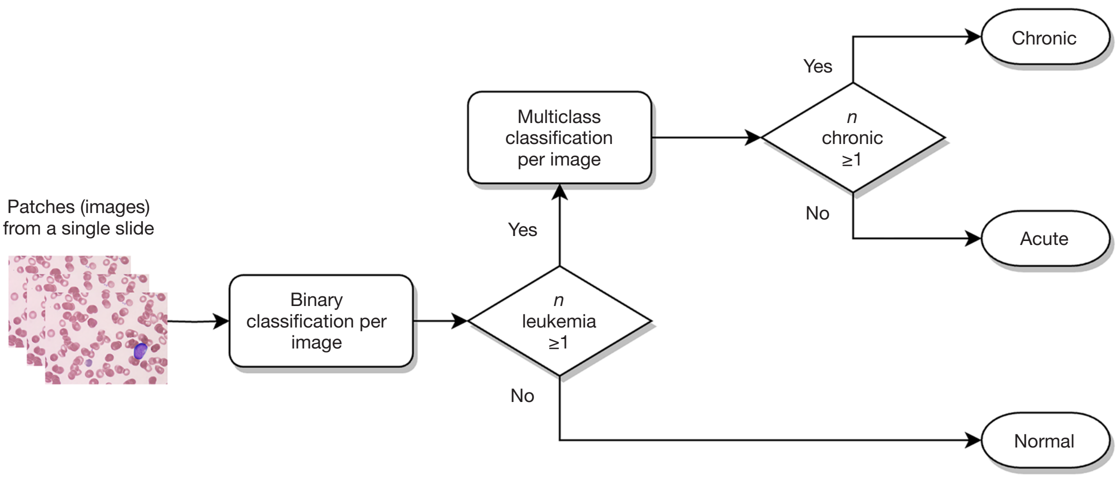


Figure 1 Flow diagram of a hierarchical three-class prediction pipeline with binary classification at step 1 and multiclass classification at step 2. The pipeline takes in a set of patches from a single slide. If one or more patches show leukemia, multiclass classification, performed on every patch, is used to further classify whether the leukemia is acute or chronic and then generate a patient-level diagnosis.

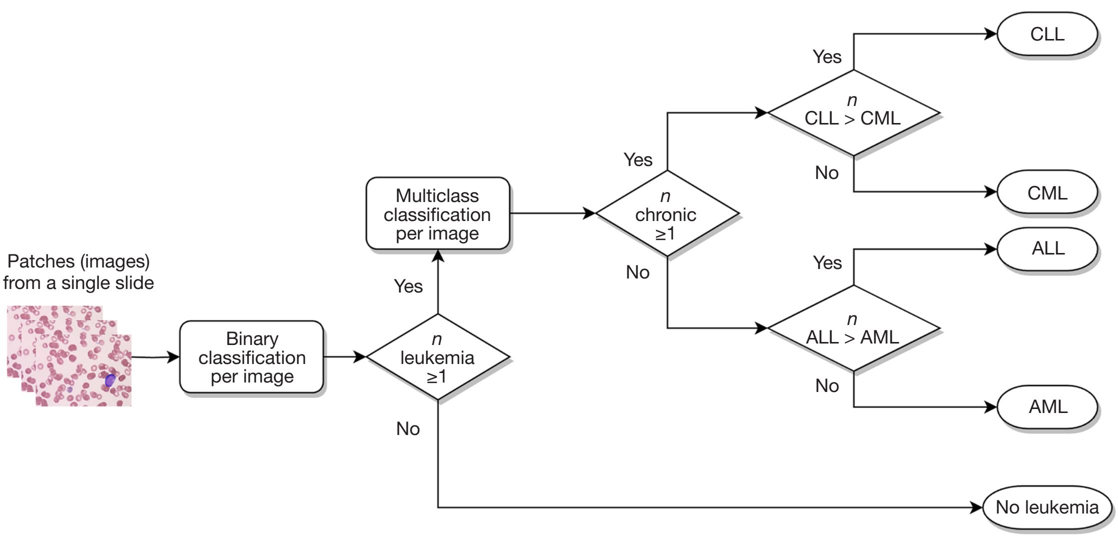


Figure 2 Flow diagram of a hierarchical five-class prediction pipeline with binary classification at step 1, multiclass classification at step 2, and majority logic at step 3. The pipeline takes in as input a set of patches from a single slide. If one or more patches show leukemia, multiclass classification is performed using two models (3-class and 5-class), at different levels of the hierarchy, generating two labels per image. These are then used to generate a patient-level diagnosis using the voting steps shown above. CLL, chronic lymphoblastic leukemia; CML, chronic myeloid leukemia; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia.

classification task was relatively easy for the models, and the outcomes were reliable. Directly classifying the subtype of leukemia was more challenging and led to more mistakes between the normal and leukemia classes.

The final part of the pipeline was an aggregation method to generate a patient-level diagnosis based on the models' outputs. A rule-based method was developed to formulate the final decision, as shown in *Figure 1*. The method shown uses two models for 3-class leukemia classification, with the three classes being Acute, Chronic, and No Leukemia. We followed a similar strategy for 5-class classification of ALL, AML, CLL, CML, and No Leukemia, as shown in *Figure 2*,

Table 2 Summary of the deep learning models deployed in AI pipelines

Model name	Task	Classes
Model 1	Binary	Leukemia, no leukemia
Model 2	Multi-class	Acute, chronic, no leukemia, reactive cells
Model 3	Multi-class	ALL, AML, CLL, CML, no leukemia

AI, artificial intelligence; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphoblastic leukemia; CML, chronic myeloid leukemia.

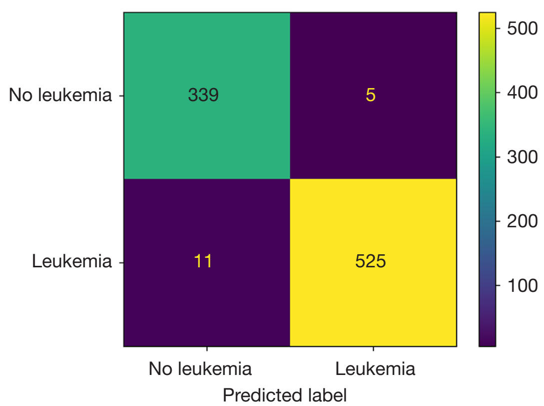


Figure 3 Confusion matrix for binary model. This confusion matrix summarizes the performance of the binary model in classifying the two classes: leukemia, no leukemia. It shows the number of correctly and incorrectly classified images for each class. TP: 525; TN: 339; FP: 5 (type I error); FN: 11 (type II error). TP, true positive; TN, true negative; FP, false positive; FN, false negative.

wherein three models were used (one binary model and two multi-class models).

All three models used in the pipelines were trained for 300 epochs on a single NVIDIA Quadro RTX 6000 GPU with 24 GB of memory, using a batch size of 32 and the AdamW optimizer with a learning rate of 1e−3 (0.001). The categorical cross-entropy loss function was used for optimization. All three models were then tested on a held-out set of 880 (20%) images from the training set.

Model 1: the binary model deployed at step 1 of both pipelines

This model was trained to differentiate leukemia from no leukemia, as shown in *Table 2*. The confusion matrix shown in *Figure 3* summarizes the number of correctly and incorrectly classified cases for the two classes. The confusion matrix revealed 525 true positives, 339 true negatives, 5 false positives, and 11 false negatives. This model achieved an impressive accuracy of approximately 98%, with sensitivity (recall) and specificity around 97% and 98%, respectively. The precision (PPV) stood at 99%, and the F1 score at approximately 98%. Hence, this model was employed at step 1 of the 3-class and 5-class pipelines due to its high reliability, as detailed in *Table 3*.

Model 2: the 4-class model deployed at step 2 of 3-class pipeline

This model was developed to predict four classes: acute leukemia, chronic leukemia, no leukemia and reactive cells, as shown in *Table 2*. The confusion matrix summarizes the number of correctly and incorrectly classified cases for the four classes as shown in *Figure 4*. This model achieved an accuracy of approximately 81%, with sensitivity (recall) and specificity around 81% and 94%, respectively. The precision (PPV) stood at 81%, and the F1 score at approximately 81%, as detailed in *Table 3*.

Model 3: the 5-class model at step 2 of 5-class pipeline

This model was developed to predict five classes ALL, AML, CLL, CML and No Leukemia, as shown in *Table 2*. The confusion matrix summarizes the number of correctly and incorrectly classified cases for the five classes as shown in *Figure 5*. This model achieved an accuracy of approximately 81%, with sensitivity (recall) and specificity around 81% and 95%, respectively. The precision (PPV) stood at 83%, and the F1 score at approximately 81%, as detailed in *Table 3*.

Results

Three-class prediction pipeline for the external validation set

This pipeline was designed to predict one of three classes: acute, chronic, and no leukemia (which includes reactive cells). On the external validation set (2,973 images from 64 patients), its sensitivity was 94%, and its specificity was 96%. All the performance metrics, the PPV, sensitivity, specificity, F1 score, and accuracy of pipeline were above 90%, and they outperformed the three pathologists, as

Table 3 Performance metrics of deep learning models deployed in AI pipelines

Evaluator	Precision/PPV	Sensitivity/recall	F1	Accuracy	Specificity
Model 1	0.99	0.97	0.98	0.98	0.98
Model 2	0.81	0.81	0.81	0.81	0.94
Model 3	0.83	0.81	0.81	0.81	0.95

AI, artificial intelligence; PPV, positive predictive value.

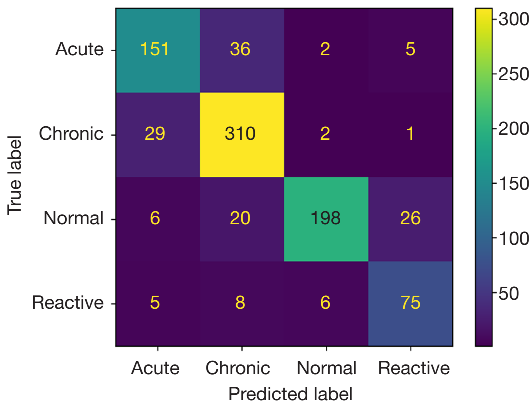


Figure 4 Confusion matrix for 4-class model. This confusion matrix summarizes the performance of the model in classifying the images into four classes: acute leukemia (acute), chronic leukemia (chronic), no leukemia (normal), and reactive cells (reactive). Each cell shows the number of images where the model predicted a certain class (horizontal axis) compared to the actual true class (vertical axis). Diagonal cells (shaded green): represent correct predictions. Off-diagonal cells: represent misclassifications.

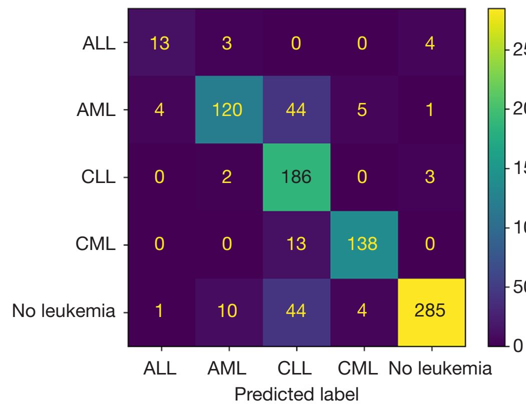


Figure 5 Confusion matrix for 5-class model. This confusion matrix summarizes the performance of the model in classifying the images into five classes: ALL, AML, CLL, CML, no leukemia. Each cell shows the number of images where the model predicted a certain class (horizontal axis) compared to the actual true class (vertical axis). Diagonal cells (shaded green): represent correct predictions. Off-diagonal cells: represent misclassifications. ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphoblastic leukemia; CML, chronic myeloid leukemia.

shown in *Figure 6*. As shown in *Figure 6A*, the F1 score of the AI models was greater than 0.9 in all three classes and achieved an almost perfect F1 score of 1 in predicting chronic leukemia. As noted in *Figure 6B*, a few patients with no leukemia were labelled as acute leukemia, but none of the acute or chronic leukemia cases were labelled as no leukemia. In *Figure 6C*, the average performance of the AI models excelled in all five parameters tested. In addition, when compared to a single-model approach, the hierarchical models perform better in terms of precision and F1-score (*Figure 7*). Moreover, the Cohen's Kappa agreement between the AI models and pathologists ranged between 74–77%, while the agreement amongst the three pathologists ranged from 82–87%, as shown in *Figure 8*.

Five-class prediction pipeline for the external validation set

This pipeline was designed to predict one of five classes: ALL, AML, CML, CLL, and no leukemia. On the external validation set (2,973 images from 64 patients), its sensitivity was 80%, and its specificity was 94%. The F1 scores of the AI pipeline were 0.57 for the AML class and 0.36 for the ALL class, while the F1 score was 1 for the CLL and CML classes. Like the AI pipeline, the F1 score for the three pathologists was lower, ranging between 0.22 to 0.66 for ALL and 0.82 to 0.86 for AML. However, the F1 score of the AI pipeline excelled in the CML and CLL classes. The F1 score of the AI pipeline for the no leukemia class was near 0.9, approaching the score of the other pathologists *Figure 9A*. As shown in

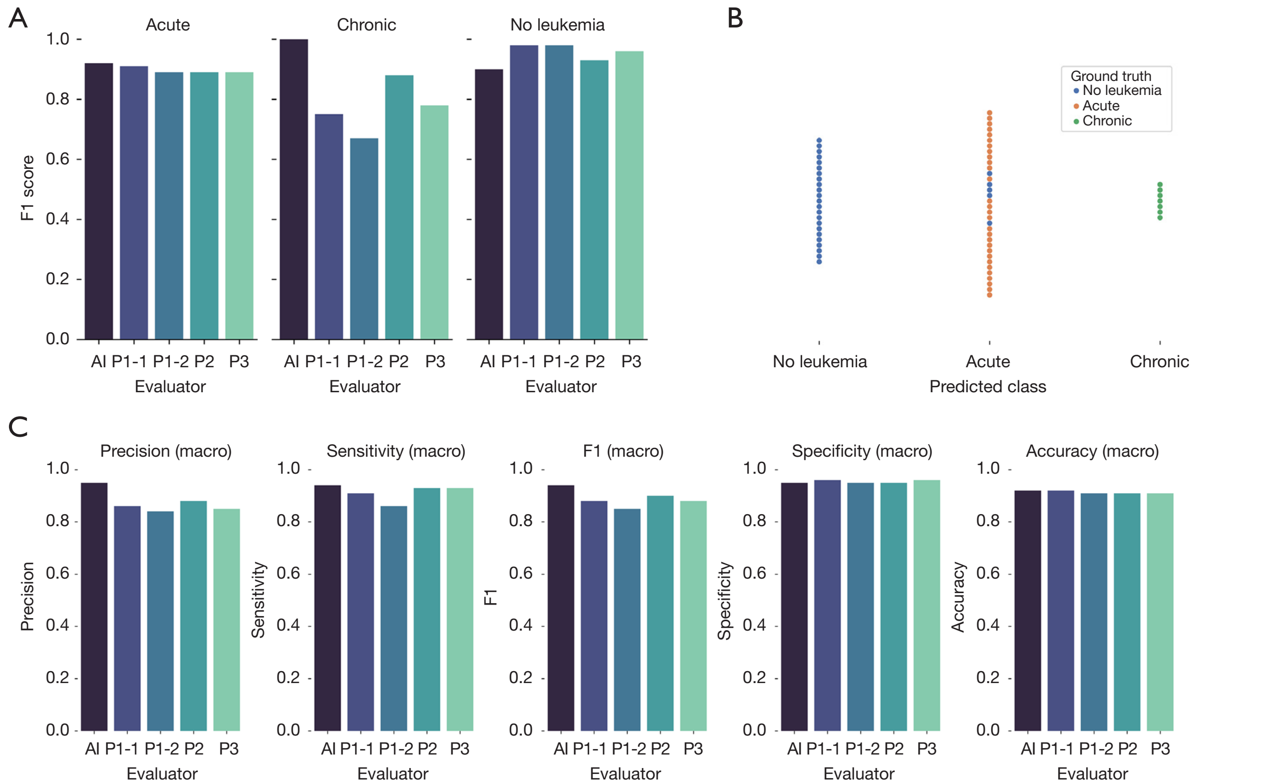


Figure 6 Comparative performance metrics of three-class patient level prediction pipeline and three pathologists in identifying and diagnosing two classes of leukemia, in addition to the no leukemia class. (A) The F1 scores for acute, chronic, and no leukemia cases are displayed, highlighting the AI pipelines' superior F1 score in acute and chronic leukemia classification. (B) The number of patients AI pipelines correctly identified, with a notable correct classification of all leukemia cases as leukemia. (C) Provides a detailed comparison of positive predictive value, sensitivity, F1 score, specificity, and accuracy between the AI pipeline and pathologists. Note: P1-1 and P1-2 represent two different evaluations six weeks apart by the same pathologist, for intra-observer consistency. AI, artificial intelligence.

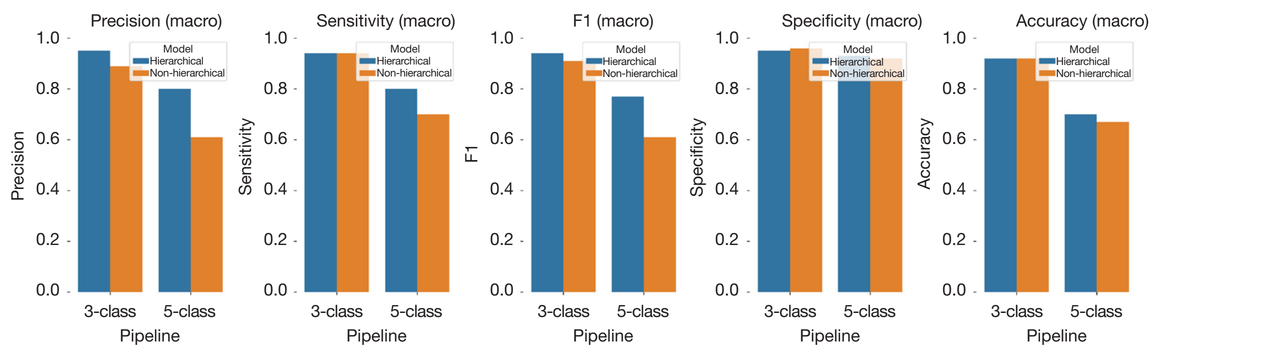
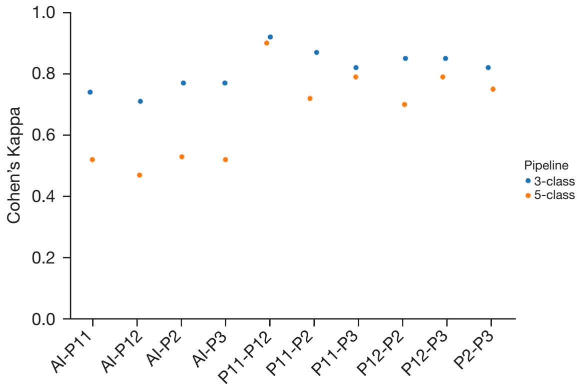


Figure 7 Comparison of patient-level performance between the hierarchical approach and a non-hierarchical single-model approach with the 3-class and 5-class pipelines. The hierarchical approach performs better overall, with the only exception being the specificity of the 3-class pipeline.



	C Kappa 3-class	C Kappa 5-class
AI-P11	0.74	0.52
AI-P12	0.71	0.47
AI-P2	0.77	0.53
AI-P3	0.77	0.52
P11-P12	0.92	0.9
P11-P2	0.87	0.72
P11-P3	0.82	0.79
P12-P2	0.85	0.7
P12-P3	0.85	0.79
P2-P3	0.82	0.75

Figure 8 Comparison of the agreement between AI pipelines and pathologists in leukemia diagnosis using Cohen's Kappa score. The scores for 3-class and 5-class patient level prediction pipelines are shown by blue and red dots respectively. The graph shows that the three-class pipeline has higher agreement scores than the 5-class pipeline. The table on the right gives the exact Kappa scores for each evaluator pair. The AI pipeline agrees less with pathologists than pathologists do amongst themselves. AI, artificial intelligence.

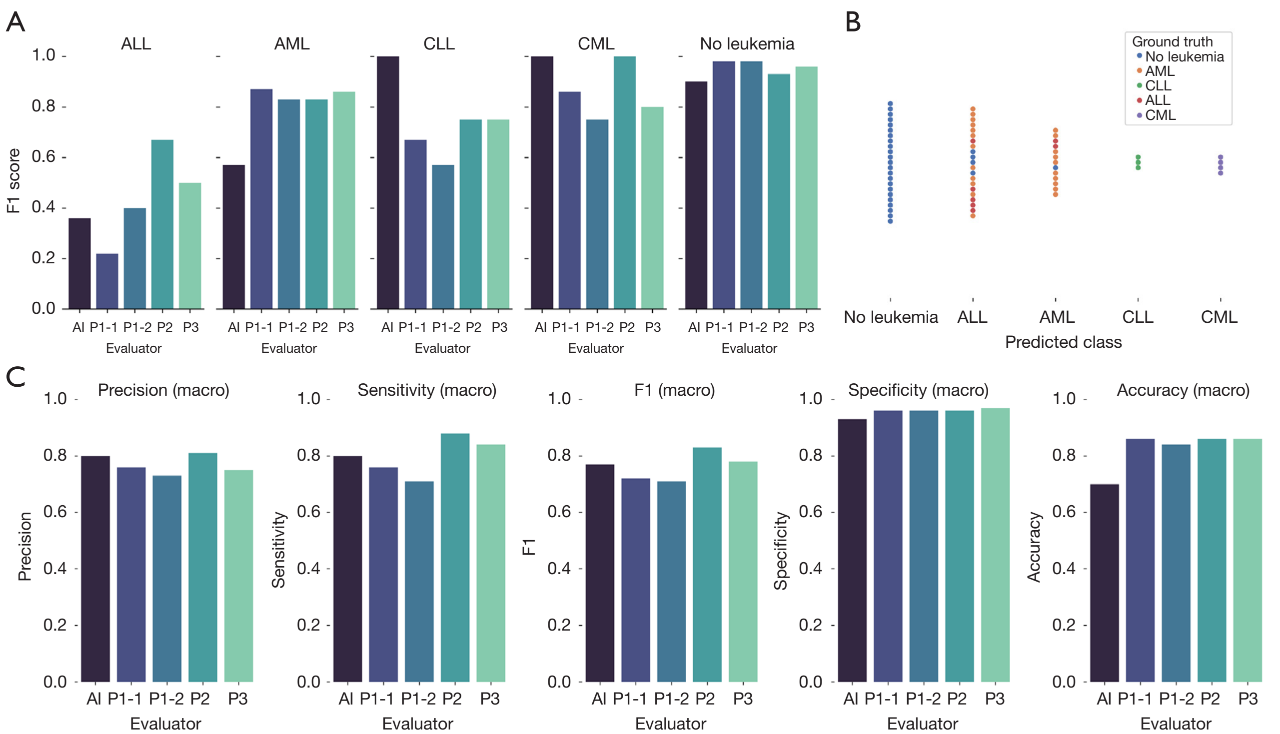


Figure 9 Comparison of the 5-class patient level prediction pipeline and three pathologists in diagnosing leukemia types (AML, ALL, CML, CLL) in addition to no leukemia. (A) The AI pipelines' higher F1 scores for CML, CLL, and no leukemia, but lower for ALL and AML. (B) The AI pipelines' classification of patients. (C) Presents a detailed comparison of various metrics between the pipeline and pathologists, with P1-1 and P1-2 denoting two evaluations by same pathologist six weeks apart for intra-observer consistency. AI, artificial intelligence; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphoblastic leukemia; CML, chronic myeloid leukemia.

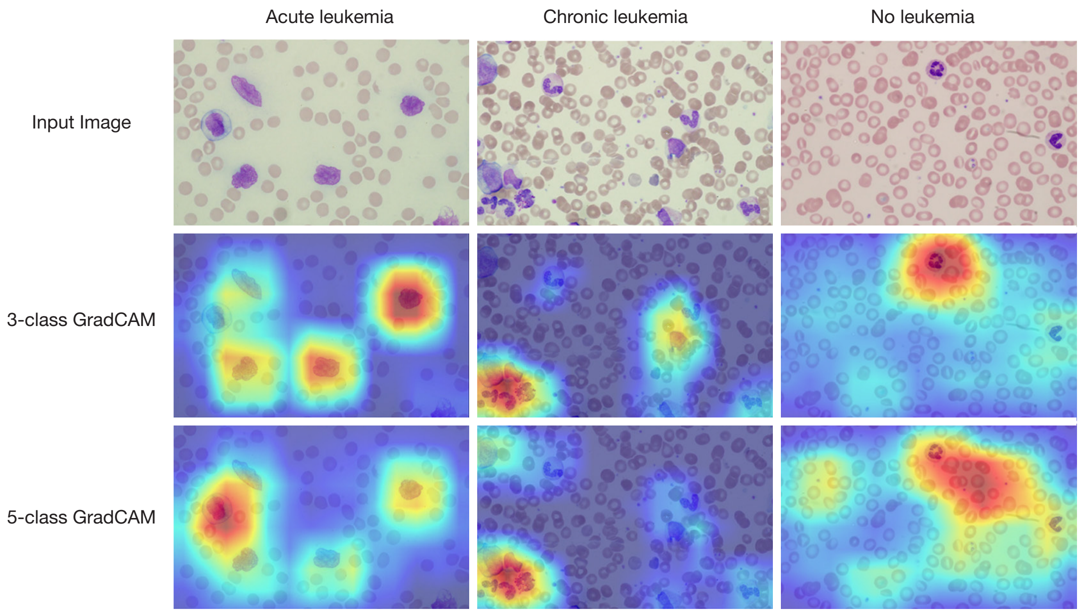


Figure 10 Gradient-weighted Class Activation Maps (GradCAM) for samples from the external validation set. The model predictions seem to correctly focus on white blood cells in general and on blast/leukemia cells in particular when classifying leukemia samples. Pappenheim method, 1,000× magnification.

Figure 9B, none of the leukemia cases were classified as no leukemia, and all the CML and CLL cases were predicted correctly by the AI pipeline. The overall accuracy of the pipeline was 0.77, compared to a range of 0.84–0.88 for the other pathologists, as shown in *Figure 9C*. Moreover, the hierarchical pipeline outperformed a single-model approach in all metrics (*Figure* 7). For detailed numerical values of each pipeline, please refer to the Supplementary file (Appendix 3), Tables S1,S2.

The Cohen's Kappa agreement between the AI pipeline and the pathologists ranged between 52–53%, while the agreement among the pathologists ranged from 72–79%, as seen in *Figure 8*. AI struggled particularly with distinguishing between ALL and AML. In six out of ten cases, all three pathologists correctly identified ALL or AML, but the AI model mislabelled them. In four other instances, the model labelled patients with no leukemia as ALL or AML. Conversely, there were two cases where the AI correctly identified the leukemia type, while all three pathologists mislabelled them. In one case, a pathologist agreed on CLL when the patient had ALL, and in the other, a pathologist agreed on AML, but the case was ALL. The remaining cases involved disagreement among one or two pathologists about the AI's prediction. This included situations where one pathologist agreed with the AI's correct or incorrect prediction while the other two disagreed, and cases where two pathologists agreed with the AI's prediction while the other one disagreed.

Discussion

In this study, we developed two clinically applicable AI pipelines using hierarchically deployed DL models to detect and predict the leukemia types from whole slide PBS HPF images. Mimicking pathologists' holistic examination, these pipelines determine the type of leukemia from PBS images. These are trained and tested on the large, real-world, annotated leukemia HPF image datasets, which represent the majority of leukemia types. The 3-class pipeline classifies the slide into acute, chronic, or no leukemia with a sensitivity of 94%, specificity of 96%, F1 score of 94%, and accuracy of 92%. This pipeline outperformed pathologists and has the potential to identify leukemia with greater reliability. Pipeline 2, a 5-class prediction pipeline that delves deeper into four types of leukemia in addition to no leukemia, achieved a sensitivity of 80%, specificity of 94%, F1 score of 77%, and accuracy of 70%. This 5-class pipeline also performed well within the range of pathologists and has the potential to speed up the clinical decision process. Qualitatively, the individual models have shown logical outcomes in terms of gradient-weighted class activation maps as can be seen in *Figure 10*. Both the 3-class and 5-class models correctly tend to focus on the abnormal white blood cells when making a decision. Our study shows the potential of deploying multiple innovative DL approaches in hierarchical pattern along with computer logic for improving the speed and efficiency of leukemia diagnosis. The performance gain achieved by the introduction of the hierarchy of models has also been demonstrated, particularly so with the 5-class pipeline, as shown in *Figure* 7. We expect our models to be scalable and generalizable and can be employed at other centre with minimal modifications.

Performance and challenges of 3-class prediction pipeline

This 3-class pipeline prediction excelled in comparison to pathologists, effectively identifying and classifying leukemia into two clinically relevant classes. The higher F1 scores of this pipeline, compared to the 5-class pipeline, affirm that the performance of DL models improves with a decreasing number of classes exhibiting optimal performance in binary classification. Similar higher accuracies (>95%) are well reported in other binary classification studies of leukemia, even though the datasets were small and homogeneous (4,5). In our AI pipeline, leveraging the binary model at step 1 contributed to achieving heightened specificity (96%), ensuring accurate identification of leukemia cases without mislabelling any as 'No Leukemia'. The addition of a multi-class models and aggregative computer logic at step 2 also led to better performance. A few cases from 'No Leukemia' were misclassified as leukemia, which can be attributed to the difficult-to-classify cells, the reactive cells. These cells may be difficult to distinguish from the leukemia cells, even for the pathologists, and for the modern AI-enabled advanced haematology analysers (19). The distinguishing features of these cells are not established, and the agreement among pathologists on these is limited, which may have led to some images in training set with inappropriate ground truth label (20). Higher performance in the chronic leukemia class may be attributable to their distinguished morphological features, and these classes are relatively easy to identify using manual microscopy (21,22). Even though ML models trained on microscopic images of bone marrow are more desirable for diagnosis, bone marrow tests are invasive and limited to tertiary healthcare centres (23). Overall, this 3-class prediction pipeline achieved higher sensitivity (94%) and specificity, making it suitable for clinically applicable medical applications, with the potential for deployment in primary and mid-level healthcare facilities (24,25).

Performance and challenges of the 5-class prediction pipeline

Our initial work, published elsewhere, was based on a seven-class AI classification model (AML, ALL, CML, CLL, no leukemia, APML, reactive cells) using training set 1. While it achieved high performance on the test set, it exhibited low performance on a new dataset. This outcome was counterintuitive for a single image-based prediction (17). This led us to improve the method and develop the present AI pipelines. The 5-class AI pipeline outperformed (F1 score) pathologists in the CML and CLL classes, but its performance was lower than pathologists in the ALL, AML, and no leukemia classes.

The AI pipeline faced greater challenge in distinguishing ALL from AML, as these are known to be the most difficult classes to predict and it is evident from the drop in the pathologists performance in these classes, as shown in *Figure 9*. These are the classes where present diagnostic practice depends on surface molecule characterization and genetic abnormalities to confirm the leukemia type. Due to morphological similarity between reactive, atypical, and neoplastic lymphocytes, the difficulty level for ALL identification is much higher than AML (20,26).

There are no published reports of 5-class leukemia AI prediction model to compare our results head-on. However, our 5-class prediction AI pipeline achieved 77% overall accuracy compared to 81.7% accuracy in a 4-class classification model by Ahmed *et al.* (27). This is the first study to our knowledge that uses a more intuitive patient-level leukemia classification using microscopic HPF images as it is clinically relevant (28). Higher performance in the CLL and CML classes may be due to multiple factors such as the clearly defined morphological characteristics and adequate training data (21,23). But a lower number of cases in the validation dataset may also provide false reassurance. The performance may fall when tested on a greater number of cases, which was not possible at this stage due to the number of cases available (29). The agreement among this pipeline and pathologist on predictions was lower than the 3-class approach. Agreement between this AI pipeline and pathologists was lower than the agreement amongst the pathologists reflects the differences in AI models way of learning from the images compared to trained pathologists. While this 5-class pipeline shows promise for identifying

CLL and CML, further research with larger datasets is necessary to improve performance across all classes and enhance agreement with pathologists, particularly for challenging classes.

Strengths and limitations of the proposed AI pipelines

Our AI pipelines, leveraging a large, diverse dataset, a whole-slide image analysis approach, and innovative use of multiple DL models, achieved promising performance. However, further research is needed to address limitations, such as including rarer leukemia types and diverse patient populations. For detailed description of strengths and limitations, please refer to Supplementary file (Appendix 4) and Table S3.

Conclusions

Our study demonstrates the feasibility of developing a comprehensive and clinically relevant DL-based pipeline for identifying leukemia and predicting common types from PBS images. We achieved this by constructing a large, annotated real-world dataset of HPF microscopy images. Our approach mimics pathologist evaluation methods by utilizing whole slide images and combining multiple models hierarchically. By aggregating image-level outputs, we provide patient-level predictions.

Remarkably, our model outperformed pathologists in the 3-class classification and achieved comparable performance in the 5-class classification, despite the inherent challenge of differentiating between ALL and AML. Our proposed method holds promise for expediting and enhancing the accuracy of leukemia diagnosis, potentially serving as an effective diagnostic tool for improved patient care.

Future research

We aim to expand disease coverage to include additional blood disorders, incorporate cell segmentation for refined model performance, automate the process for seamless integration and explore multi-modal data integration to enhance accuracy and trust in predictions. For further details on future research recommendations, please refer to the Supplementary file (Appendix 5).

Acknowledgments

We are grateful to our medical laboratory technologists, Saleh Bin Amro, Haider Rafik, Maria Christina Fajarito, Ma. Jospehine Tantoco, Rommel Ramiscal, and Sara Sebait Al Katheer, for their invaluable assistance in providing and meticulously sorting the slides used in this study. We also extend our special thanks to Haneym Bakil Manisan, the laboratory supervisor, for providing essential support by ensuring access to the microscopes and laboratory space required for our image data collection process. The contributions of all these dedicated staff members were instrumental in building the robust dataset that formed the foundation of this research.
Funding: None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://jmai.amegroups.com/article/view/10.21037/jmai-24-74/rc

Data Sharing Statement: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-24-74/dss

Peer Review File: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-24-74/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-24-74/coif). N.S., M.E.S.S., I.A. and M.Y. are planning to submit a patent application for Mohamed bin Zayed University of AI and Sheikh Shakhbout Medical City. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the principles of the Declaration of Helsinki (as revised in 2013). Approval was granted by the Ethics Committee of Sheikh Shakbout Medical City (Date. 23-03-2022/No. MAFREC-273). Informed consent was not applicable due to the retrospective nature of the study using anonymized data and minimal risk to subjects.

References

1. Cancer Stat Facts: Leukemia. Available online: https://seer.cancer.gov/statfacts/html/leuks.html [Cited 2024 Feb 27].
2. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. Leukemia 2022;36:1703-19.
3. Andrey Bychkov, Michael Schubert. the Pathologist. 2023. Constant Demand, Patchy Supply The first comprehensive study on the global pathologist workforce both identifies key disparities in pathologist supply and prepares the ground for correcting these imbalances. Available online: https://thepathologist.com/outside-the-lab/constant-demand-patchy-supply
4. Patel N, Mishra A. Automated Leukaemia Detection Using Microscopic Images. Procedia Comput Sci 2015;58:635-42.
5. Thanh TTP, Vununu C, Atoev S, et al. Leukemia Blood Cell Image Classification Using Convolutional Neural Network. International Journal of Computer Theory and Engineering 2018;10:54-8.
6. Ghaderzadeh M, Asadi F, Hosseini A, et al. Machine Learning in Detection and Classification of Leukemia Using Smear Blood Images: A Systematic Review. In: Wang P. editor. Sci Program 2021;2021:1-14.
7. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:770-8.
8. El Alaoui Y, Elomri A, Qaraqe M, et al. A Review of Artificial Intelligence Applications in Hematology Management: Current Practices and Future Prospects. J Med Internet Res 2022;24:e36490.
9. Salah HT, Muhsen IN, Salama ME, et al. Machine learning applications in the diagnosis of leukemia: Current trends and future directions. Int J Lab Hematol 2019;41:717-25.
10. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how, when and where? Clin Kidney J 2021;14:49-58.
11. Manz CR, Chen J, Liu M, et al. Validation of a Machine Learning Algorithm to Predict 180-Day Mortality for Outpatients With Cancer. JAMA Oncol 2020;6:1723-30.
12. Chola C, Muaad AY, Bin Heyat MB, et al. BCNet: A Deep Learning Computer-Aided Diagnosis Framework for Human Peripheral Blood Cell Identification. Diagnostics (Basel) 2022;12:2815.
13. Qin F, Gao N, Peng Y, et al. Fine-grained leukocyte classification with deep residual learning for microscopic images. Comput Methods Programs Biomed 2018;162:243-52.
14. Çnar A, Tuncer SA. Classification of lymphocytes, monocytes, eosinophils, and neutrophils on white blood cells using hybrid Alexnet-GoogleNet-SVM. SN Appl Sci 2021;3:503.
15. Li Y, Zhu R, Mi L, et al. Segmentation of White Blood Cell from Acute Lymphoblastic Leukemia Images Using Dual-Threshold Method. Comput Math Methods Med 2016;2016:9514707.
16. Eckardt JN, Middeke JM, Riechert S, et al. Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. Leukemia 2022;36:111-8.
17. Hamdi I, El-Gendy H, Sharshar A, et al. Breaking down the Hierarchy: A New Approach to Leukemia Classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2023:104-13.
18. Huang G, Liu Z, Van Der Maaten L, et al. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017:2261-9.
19. Kim HN, Hur M, Kim H, et al. Performance of automated digital cell imaging analyzer Sysmex DI-60. Clin Chem Lab Med 2017;56:94-102.
20. van der Meer W, van Gelder W, de Keijzer R, et al. The divergent morphological classification of variant lymphocytes in blood smears. J Clin Pathol 2007;60:838-9.
21. Thompson PA, Kantarjian HM, Cortes JE. Diagnosis and Treatment of Chronic Myeloid Leukemia in 2015. Mayo Clin Proc 2015;90:1440-54.
22. Strati P, Jain N, O'Brien S. Chronic Lymphocytic Leukemia: Diagnosis and Treatment. Mayo Clin Proc 2018;93:651-64.
23. Huang F, Guang P, Li F, et al. 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:e23154.
24. Bain BJ. Diagnosis from the blood smear. N Engl J Med 2005;353:498-507.
25. Glaros AG, Kline RB. Understanding the accuracy of tests with cutting scores: the sensitivity, specificity, and predictive value model. J Clin Psychol 1988;44:1013-23.
26. Efebera YA, Caligiuri MA. Classification and Clinical Manifestations of Lymphocyte and Plasma Cell Disorders. In: Kaushansky K, Prchal JT, Burns LJ, et al. editors. Williams Hematology, 10e. New York, NY: McGraw-Hill Education; 2021. Available online: accessmedicine.mhmedical.com/content.aspx?aid=1178757691
27. Ahmed N, Yigit A, Isik Z, et al. Identification of Leukemia Subtypes from Microscopic Images Using Convolutional Neural Network. Diagnostics (Basel) 2019;9:104.
28. Khened M, Kori A, Rajkumar H, et al. A generalized deep learning framework for whole-slide image segmentation and analysis. Sci Rep 2021;11:11579.
29. Siegel RL, Miller KD, Fuchs HE, et al. Cancer Statistics, 2021. CA Cancer J Clin 2021;71:7-33.

doi: 10.21037/jmai-24-74
Cite this article as: Syed N, Saeed MES, Hussain S, Mirza I, Abdalla AM, Al Zaabi EA, Afrooz I, Hashmi S, Yaqub M. Novel hierarchical deep learning models predict type of leukemia from whole slide microscopic images of peripheral blood. J Med Artif Intell 2025;8:5.

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.

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