

PCM-Leukemia: A Large Scale Dataset for Detection and Classification of Acute Leukemia Subtypes

Lin Tun Naing¹ 

Apichat Photi-A² 

Kaung Htet Cho¹ 

Matthew N. Dailey¹ 

Mongkol Ekpanyapong^{*1} 

Piya Rujkijyanont^{*2} 

ST124225@AIT.ASIA

APICHAT.PHO@PMK.AC.TH

ST124092@AIT.ASIA

MDAILEY@AIT.ASIA

MONGKOL@AIT.ASIA

PIYA_RUJK@PCM.AC.TH

¹ *Artificial Intelligence Center, Asian Institute of Technology, Khlong Luang, Pathum Thani, Thailand*

² *Division of Hematology and Oncology, Department of Pediatrics, Phramongkutklao Hospital and Phramongkutklao College of Medicine, Ratchathewi, Bangkok, Thailand*

Editors: Under Review for MIDL 2026

Abstract

Acute leukemia, consisting of acute lymphoblastic leukemia and acute myeloid leukemia, is a common hematologic malignancy. While the survival rate of patients with acute leukemia in high-income countries has significantly improved with contemporary chemotherapy regimens, the outcomes of those patients in low- and middle-income countries are still poor given delayed diagnosis and treatment. Although bone marrow examination is the gold standard for diagnosis of leukemia, detection of leukemia cells in peripheral blood can urge healthcare providers to start initial supportive care and refer patients to tertiary hospitals for definite diagnosis and proper treatment. Recently, Artificial Intelligence (AI) has shown promise in automating this process, yet the efficacy of deep learning models is often limited by the scarcity of large-scale, annotated datasets. To address this gap, we introduce a large-scale novel dataset, named PCM-leukemia, collected from Phramongkutklao Hospital and Phramongkutklao College of Medicine, comprising 19,191 images annotated by two hematology specialists and a trained biomedical researcher. The dataset includes bounding box and cell type annotations for eight distinct classes, including lymphoblasts and myeloblasts, yielding a comprehensive collection of 40,103 extracted single-cell crops. To validate the dataset’s utility for developing robust diagnostic tools, we established baselines using state-of-the-art object detection (YOLO11, DEIM) and classification pipelines. Specifically, we compared a standard CNN baseline (ResNet50) against a foundation model pretrained on histopathological images (DinoBloom), utilizing both linear probing and fine-tuning. Experimental results on our hold-out test set demonstrate the dataset’s high quality, supporting a strong mAP of 87.6% for WBC only detection and a classification accuracy of 92.59% with Macro F1-Score of 92.11% using the fine-tuned DinoBloom model. Furthermore, to assess the dataset’s capacity to facilitate generalization, models trained on our data were evaluated on external benchmarks for both ALL and AML subtypes. On the ALL-IDB1 dataset—re-annotated by our experts to include bounding boxes—the fine-tuned model demonstrated strong direct transferability, achieving an accuracy of 88.02% and a Macro F1-Score of 74.27% without training on the external set. Conversely, evaluation on the Munich AML Morphology Dataset (LMU) revealed a more challenging transfer

* PR and ME are corresponding authors and contributed equally to the manuscript.

scenario, yielding a baseline accuracy of 39.67% and Macro F1-Score of 40.63% in the direct transfer setting. To address this domain shift issue, we employed a low-resource supervised adaptation strategy; by incorporating just 10% of the target data into the training process alongside our proposed dataset, accuracy on the remaining 90% hold-out set increased significantly to 81.94%, and the Macro F1-Score increased to 67.94%. These results confirm that the proposed dataset captures representative features necessary for training generalizable and adaptable AI systems.

Keywords: Acute Lymphoblastic Leukemia, Acute Myeloid Leukemia, Deep Learning, Object Detection, Classification, Domain Adaptation, Transfer Learning, Peripheral Blood Smear.

1. Introduction

Acute leukemia, including Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), remains a major global health burden. ALL is the most common childhood malignancy and achieves excellent survival rates when timely diagnosis and appropriate therapy are available. However, despite substantial improvements in chemotherapy and supportive care in high-income countries, outcomes in low- and middle-income regions remain disproportionately poor due to delayed diagnosis and limited access to specialized hematologic services (Rujkijyanont and Inaba, 2024). Bone marrow examination is the diagnostic gold standard, but it is invasive, painful, and often inaccessible in resource-limited settings. Consequently, the peripheral blood smear (PBS) is widely used as an initial screening tool for suspected leukemia. In many settings, PBS interpretation by general practitioners who are not hematologists is essential. Manual PBS assessment is labor-intensive, operator-dependent, and constrained by the scarcity of trained hematologists. These challenges underscore the urgent need for automated diagnostic tools capable of supporting clinicians, accelerating the recognition of leukemic blasts, and facilitating timely referral to advanced care centers.

In recent years, Deep Learning (DL) has revolutionized medical image analysis, with Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; Szegedy et al., 2015; He et al., 2016) and Vision Transformers (ViTs) (Dosovitskiy, 2020) achieving state-of-the-art results across various modalities. For object detection tasks, single-stage architectures such as YOLO (Redmon et al., 2016; Bochkovskiy et al., 2020; Jocher and Qiu, 2024) and Transformer-based models like DEIM-RT-DETR (Zhu et al., 2020; Huang et al., 2025) have enabled unified pipelines that simultaneously localize and classify cellular structures. In parallel, image classification has advanced through robust CNN backbones—such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and EfficientNet (Tan and Le, 2019)—as well as Vision Transformers (Dosovitskiy, 2020), which capture long-range dependencies in images. More recently, the paradigm has shifted toward Foundation Models adapted from Natural Language Processing (NLP). Large-scale self-supervised frameworks, such as DINOv2 (Oquab et al., 2023), have been successfully adapted to computational pathology; notably, DinoBloom (Koch et al., 2024) leverages this framework to extract robust features from histopathological images, offering a promising avenue for hematological analysis.

Despite these algorithmic advancements, the application of AI to leukemia diagnosis is hindered by a critical bottleneck: the scarcity of high-quality, large-scale, annotated datasets. While several open-source datasets exist, they often suffer from severe limi-

tations that impede the development of clinically deployable models. For instance, the widely used ALL-IDB1 (Labati et al., 2011) contains only 108 images with approximately 700 white blood cells, which is insufficient for training deep networks without overfitting. Other large-scale datasets, such as CNMC-Leukemia (Mourya et al., 2019), AML-Cytomorphology_LMU (Munich) (Matek et al., 2019), Acevedo (Acevedo et al., 2020) and Raabin-Leukemia (Kouzehkanan et al., 2022), present their own challenges: many consist only of single-cell crops rather than whole slides, lack bounding box annotations for detection tasks, or suffer from inconsistent resolution and labeling standards. Table 1 provides a detailed comparison of these existing benchmarks against our proposed contribution.

Furthermore, prior research often lacks rigorous external validation, raising concerns about model generalization. For example, Boldú et al. (2021) and Shaheen et al. (2021) developed CNN-based models for AML detection but evaluated them primarily on internal or self-collected data, leaving their performance on unseen domains uncertain. While Yan et al. (2025) recently proposed a large-scale framework for cell segmentation and classification, their dataset is limited to single-cell crops, removing the contextual information available in whole-slide images. Similarly, Syed et al. (2024) introduced a multi-stage classification pipeline for multiple leukemia subtypes but acknowledged limitations regarding data availability and code reproducibility. This lack of standardization and external validation underscores the need for a comprehensive, diverse, and well-annotated resource to bridge the domain gap in automated leukemia diagnosis.

To address these challenges, we introduce **PCM-Leukemia**, a large-scale, novel dataset collected from Phramongkutklao Hospital and Phramongkutklao College of Medicine. Unlike previous datasets restricted to single-cell crops or classification-only labels, our dataset provides high-resolution whole images with expert-verified bounding boxes and cell type annotations for eight distinct classes. We establish robust baselines using state-of-the-art detection and classification models. Crucially, we move beyond internal validation by rigorously testing our models on distinct external benchmarks to assess generalization and adaptability.

The main contributions of this work can be summarized as follows:

1. **Large-Scale Annotated Dataset:** We introduce a novel, large-scale acute leukemia dataset comprising 19,191 whole-slide peripheral blood smear images and 40,103 extracted single-cell crops. Annotated by multiple experts, the dataset covers eight distinct white blood cell types, including critical lymphoblast and myeloblast classes, addressing the scarcity of high-quality, diverse data in hematology.
2. **Comprehensive Pipeline Benchmarking:** We design and compare one-stage (end-to-end detection) and two-stage (detection followed by classification) pipelines for leukemia cell identification, establishing strong baselines for future research using our proposed dataset.
3. **Foundation Model Integration:** We enhance the classification component of our pipeline by fine-tuning DinoBloom, a pathology-specific foundation model. We demonstrate that this approach yields superior accuracy compared to standard CNN baselines (e.g., ResNet50), validating the efficacy of self-supervised foundation models for precise leukemia cell discrimination.

4. **Cross-Domain Generalization & Adaptation:** We validate the utility of our dataset by evaluating our trained models on external benchmarks. We demonstrate strong direct transfer performance on ALL-IDB1. Furthermore, we address the challenge of domain shift on the AML Munich dataset by proposing a **low-resource supervised adaptation strategy**, proving that our models can be effectively adapted to new clinical environments with minimal target data.

2. Methodology

2.1. Data Collection

Peripheral blood smears were collected from multiple patients visiting Phramongkutklao Hospital and Phramongkutklao College of Medicine, Thailand, for routine screening or follow-up appointments. The data collection process and subsequent experiments were conducted under the approval of Ethical Committee, the Institutional Review Board Royal Thai Army for human data use (Reference No. IRBRTA 843/2563); consequently, patient consent was waived. The protocol strictly ensured the anonymization of patient identities to prevent backtracking.

For specimen preparation, Wright’s stain (Wright, 1902) was utilized, as it yields distinct cytoplasmic textures in white blood cells and provides high-fidelity nuclear details under microscopy. Images were acquired using an OLYMPUS BX53 microscope equipped with a high-end lens, capturing data at a resolution of 1920×1080 pixels under $100\times$ magnification (see Figure 1).

Table 1: Comparison of existing peripheral blood smear datasets for acute leukemia. The overview highlights differences in cell types, dataset scale, and annotation levels (e.g., full-frame bounding boxes versus pre-cropped single cells).

Dataset Source	Cell Types	Total Images	Types (Annotations)
ALL-IDB1 (Labati et al., 2011)	ALL	108	Full Frame (Centroids)
Munich AML (Matek et al., 2019)	AML	18,365	Cropped Single Cell
C-NMC 2019 (Mourya et al., 2019)	ALL	15,135	Cropped Single Cell (Segmentation)
Acevedo (Acevedo et al., 2020)	Normal	17,092	Cropped Single Cell
Raabin-WBC (Kouzehkanan et al., 2022)	Normal	$\sim 40,000$	Full Frame (Bboxes)
Raabin-Leukemia (Kouzehkanan et al., 2022)	ALL, AML	7,152	Full Frame (No Annotation)
LeukemiaAttri (Rehman et al., 2024)	ALL, AML, APL	$\sim 28,900$	Full Frame (Bboxes/Cell Attributes)
PCM-Leukemia (Ours)	Normal, ALL, AML, APL	19,191	Full Frame (Bboxes), Cropped Single Cell

2.2. Dataset Preparation

The collected images were annotated using the Computer Vision Annotation Tool (CVAT) (CVAT.ai, 2023) to generate bounding boxes for white blood cells. The dataset was categorized into eight primary classes: Lymphocyte, Neutrophil, Basophil, Eosinophil, Lymphoblast, Monocyte, Myeloblast_AML, and Myeloblast_APL. This stratification ensures a clear distinction between lymphoid and myeloid lineages. Notably, we specifically separated Acute Promyelocytic Leukemia (APL) from the general Acute Myeloid Leukemia (AML) class due to their divergent morphological profiles; while standard AML myeloblasts are characterized by fine chromatin and scarce granules, APL cells are distinctively marked by dense azurophilic granules and the frequent presence of Auer rods. Furthermore, to enhance the model’s robustness against negative examples, we introduced a ninth class labeled “Other” which includes artifacts, smudge cells, and cells located at the image edges with less than 50% visibility.

Table 2: Distribution of full-frame peripheral blood smear images across the training and testing sets. The split ensures a patient-level separation to prevent data leakage.

Dataset Split	Image Count
Train/Val	16,555
Test	2,636
Total	19,191

Table 3: Distribution of annotated white blood cell instances across the training and testing sets. The “Other” category comprises artifacts and smudge cells.

Cell Type	Train/Val	Test	Overall
Lymphocyte	2,383	320	2,703
Neutrophil	3,304	361	3,665
Basophil	1,576	169	1,745
Eosinophil	2,047	239	2,286
Lymphoblast	5,198	1,222	6,420
Monocyte	2,683	626	3,309
Myeloblast_AML	11,339	2,727	14,066
Myeloblast_APL	1,643	952	2,595
Other	2,197	1,117	3,314
Total	32,370	7,733	40,103

We provide converted binary class bounding boxes (“WBC” vs. “Other”) for the two-stage pipeline. In the classification stage of this pipeline, the “Other” class is explicitly excluded. The entire process of collection, annotation, and validation was performed by two hematology specialists and a trained biomedical researcher. To ensure the highest quality

of ground truth data, the annotation process was conducted through a rigorous three-stage workflow. First, an expert hematologist acquired the peripheral blood smear images directly from the microscope, establishing the primary slide-level diagnosis (Normal, ALL, AML, or APL). Subsequently, a biomedical researcher utilized the Computer Vision Annotation Tool (CVAT) to delineate bounding boxes around individual cells and assign initial class labels. Finally, a senior specialist reviewed and verified every annotated bounding box and cell category. This multi-tiered approach ensured that each cell underwent triple verification, effectively minimizing inter-observer variability and guaranteeing the consistency of the dataset. In total, we annotated 19,191 images (see Table 2), resulting in 36,789 white blood cell bounding boxes (excluding 3,314 instances classified as “Other”) (see Table 3). The dataset was split into training and hold-out testing sets at the patient level using an 85:15 ratio, yielding 16,555 training images and 2,636 testing images. To validate the generalization capability of our detection pipeline, the ALL-IDB1 dataset (Labati et al., 2011) was employed as an unseen benchmark. It was re-annotated using the identical protocol applied to our PCM-Leukemia dataset, yielding 701 white blood cell bounding boxes.

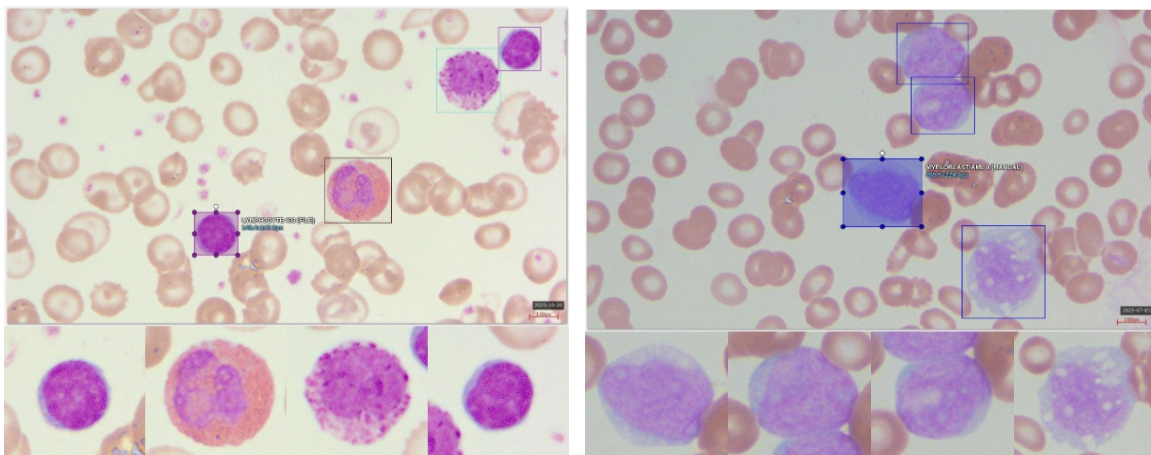


Figure 1: Qualitative visualization of annotations. Both panels display full-frame peripheral blood smear images with annotated bounding boxes (top) alongside the corresponding single-cell crops derived from these annotations (bottom). The left panel illustrates a variety of cell types, specifically (from left to right): a Lymphocyte, an Eosinophil, a Basophil, and a Lymphocyte. The right panel displays multiple Myeloblasts annotated within a single frame.

2.3. Training Pipelines

We investigated two distinct training strategies to evaluate our dataset: a one-stage end-to-end detection pipeline and a two-stage pipeline consisting of detection followed by clas-

sification. All experiments were conducted on a workstation equipped with an Intel Core i9 (14th Gen) processor, 64GB of RAM, and an NVIDIA RTX 4090 (24GB) GPU.

2.3.1. ONE-STAGE PIPELINE (END-TO-END DETECTION)

In the one-stage approach, we employed two state-of-the-art detection models: YOLO11m (Jocher and Qiu, 2024), representing a CNN-based architecture, and DEIM-RT-DETRv2 (Huang et al., 2025), a Vision Transformer (Dosovitskiy, 2020), ViT-based model. Both models were trained on the full 9-class annotated PCM-Leukemia dataset.

Preprocessing and Augmentation: As both models utilize backbones pretrained on ImageNet (Deng et al., 2009), input images were pixel-normalized using ImageNet statistics. Standard online augmentation techniques were applied, including resizing to 640×640 , random rotation, and random horizontal/vertical flipping.

Training Configuration: For model validation, the training data was randomly split into an 80:20 ratio, with a separate hold-out test set used for final evaluation.

- **YOLO11m:** Trained for 100 epochs with early stopping (converged at 85 epochs). The learning rate was initialized at 0.01 and followed a scheduler with a decay factor of 0.01.
- **DEIM-RT-DETRv2:** Trained for 60 epochs. The learning rate followed a schedule starting at 2×10^{-5} and ending at 1×10^{-5} .

To assess generalization, we also evaluated the trained models on our re-annotated ALL-IDB1 dataset as an unseen benchmark.

2.3.2. TWO-STAGE PIPELINE (DETECTION + CLASSIFICATION)

This pipeline decouples localization and classification. First, a binary detection model was trained to identify regions of interest (ROI) using only two classes: “WBC” and “Other”. The training configuration and evaluation protocols for this detector were identical to the one-stage pipeline.

Classification Refinement: Detected single-cell images were cropped from the bounding boxes and padded with a 50-pixel margin on all sides to preserve context. We compared a CNN baseline (ResNet50, pretrained on ImageNet) against a ViT-based foundation model (Dinobloom (Koch et al., 2024), pretrained on large-scale histopathology data).

For Dinobloom, we evaluated two fine-tuning strategies, resulting in a comparison of three total models (ResNet50, Dinobloom-LP+FT, and Dinobloom-Direct):

1. **Linear Probing + Fine-tuning (LP+FT):** The backbone was frozen, and the linear head was trained for 10 epochs. Subsequently, the backbone was unfrozen for an additional 50 epochs of fine-tuning with a reduced learning rate.
2. **Direct Fine-tuning:** The entire model was unfrozen from the start and fine-tuned directly.

The input image size we used to train both models is 224×224 . Performance was reported on the hold-out test set and the unseen ALL-IDB1 cropped single cells.

2.3.3. CROSS-DOMAIN GENERALIZATION AND ADAPTATION

While ALL-IDB1 serves as a benchmark for lymphoblasts, it lacks sufficient myeloblast samples. To address this, we incorporated the Munich AML Morphology Dataset (LMU) (Matek et al., 2019) to evaluate performance on Acute Myeloid Leukemia. However, initial experiments revealed a significant domain shift between the PCM-Leukemia training data and the LMU dataset, leading to poor adaptation of the baseline models.

To mitigate this, we conducted a low-resource adaptation experiment using only 10% of the Munich AML dataset for training. This experiment utilized the best-performing configuration from the classification stage (Dinobloom with LP+FT) to assess the model’s ability to adapt to new domains with limited data.

2.4. Evaluation Metrics

To evaluate the performance of the proposed models, we adopt Accuracy Score and Macro F1-Score for classification and Mean Average Precision (mAP) for object detection.

2.4.1. IMAGE CLASSIFICATION METRICS

We utilize **Accuracy** to measure the overall proportion of correct predictions. To address potential class imbalance, we also report the **Macro F1-Score**. These metrics are derived from the confusion matrix elements:

- **True Positives (TP)**: Samples correctly identified as belonging to the positive class.
- **False Positives (FP)**: Negative samples incorrectly labeled as positive (Type I error).
- **False Negatives (FN)**: Positive samples incorrectly labeled as negative (Type II error).

The F1-Score is the harmonic mean of **Precision** (P) and **Recall** (R), calculated as follows:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (1)$$

The Macro F1-score treats all classes uniformly by averaging the per-class F1 scores:

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^N F1_i \quad (2)$$

where N is the total number of classes.

2.4.2. OBJECT DETECTION METRICS

Detection performance is evaluated using **mean Average Precision (mAP)**. First, we determine the spatial correctness of a predicted box (B_p) relative to the ground truth (B_{gt}) using Intersection over Union (IoU):

$$\text{IoU} = \frac{\text{Area}(B_p \cap B_{gt})}{\text{Area}(B_p \cup B_{gt})} \quad (3)$$

A prediction is classified as a **True Positive (TP)** if the IoU exceeds a predefined threshold (typically 0.5). If the IoU is below this threshold, the prediction is a **False Positive (FP)**. Ground truth objects that are not detected are counted as **False Negatives (FN)**.

The **Average Precision (AP)** for a single class is defined as the area under the Precision-Recall curve $p(r)$. In practice, this is computed as the weighted mean of precisions at each threshold of recall:

$$AP = \int_0^1 p(r) dr \quad (4)$$

Finally, the **mean Average Precision (mAP)** is computed by averaging the AP scores across all object classes to provide a single holistic performance metric:

$$mAP = \frac{1}{C} \sum_{j=1}^C AP_j \quad (5)$$

where C is the total number of object classes.

3. Experiments and Results

3.1. Performance Evaluation of One-Stage Detector

Our two baseline models for 9-class detection demonstrated strong convergence on the internal hold-out set, achieving mAPs of 80.8% and 81.8% for YOLO11m and DEIM-RT-DETRv2, respectively. However, when evaluated on the unseen external benchmark (ALL-IDB1), performance dropped significantly, yielding mAPs of 52.4% and 60.3%, respectively (see Table 4). In this challenging scenario, the Vision Transformer-based DEIM architecture exhibited superior generalization capabilities compared to the CNN-based YOLO11m. As illustrated in Figure 10, the primary source of error was the misclassification of lymphoblasts as myeloblasts and promyelocytes. Crucially, however, we observed that localization performance remained robust: 691 out of 701 total WBCs in the external set were successfully detected, regardless of classification accuracy. This discrepancy suggests that while domain shift impacts fine-grained feature recognition, it does not hinder the model’s ability to distinguish WBCs from the background. Motivated by this high detection recall, we propose a two-stage pipeline: utilizing a binary detector to localize WBCs followed by a dedicated classification model to predict cell subtypes.

3.2. Performance Evaluation on Two Stage Pipeline

3.2.1. WBC DETECTOR

Table 5 presents the quantitative results of the first stage of our pipeline: binary object detection. At first glance, the overall mAP values—ranging from 74.6% to 75.7% on the internal set and 68.6% on the external benchmark—appear lower than those achieved in the single-stage 9-class experiment. However, this metric is heavily penalized by the “Other” class, which comprises highly heterogeneous features such as smudge cells, staining artifacts, and partial cells at the image edges. The semantic ambiguity and high intra-class variance of these artifacts make them difficult to bound precisely, thereby suppressing the mean Average Precision.

Table 4: Performance comparison of one-stage detection models. We evaluated both YOLO11m and DEIM-RT-DETRv2 on our internal hold-out test set (PCM-Leukemia) and an unseen external benchmark (ALL-IDB1). The primary metric is mean Average Precision (mAP) at IoU thresholds from 0.50 to 0.95.

Model	Test Dataset	mAP ₅₀₋₉₅
YOLO11m	PCM-Leukemia (Internal)	80.8
	ALL-IDB1 (External)	52.4
DEIM-RT-DETRv2	PCM-Leukemia (Internal)	81.8
	ALL-IDB1 (External)	60.3

This disparity is clearly illustrated in the Precision-Recall curves shown in Figure 2. While the “Other” class struggles with APs of 63.8% (internal) and 55.1% (external), the target “WBC” class maintains high robustness, achieving APs of 87.6% and 82.1%, respectively. This confirms that the detector is highly effective at localizing clinically relevant cells despite the noise introduced by artifacts. Crucially, since the “Other” class is discarded prior to the classification stage, its lower detection score does not negatively impact the final diagnostic pipeline. Given this strong WBC localization capability, we proceed to the second stage: cropping the detected white blood cells and feeding them into a specialized classification model to resolve the fine-grained subtypes.

Table 5: Performance of the binary detection models (Stage 1 of the two-stage pipeline). Both YOLO11m and DEIM-RT-DETRv2 were trained to detect only two classes: “WBC” and “Other”. Note that the overall mAP is impacted by the semantic diversity of the “Other” class.

Model	Test Dataset	mAP ₅₀₋₉₅
YOLO11m	PCM-Leukemia (Internal)	74.6
	ALL-IDB1 (External)	58.1
DEIM-RT-DETRv2	PCM-Leukemia (Internal)	75.7
	ALL-IDB1 (External)	68.6

3.2.2. SINGLE-CELL CROP CLASSIFICATION

Following the localization of WBCs in the first stage, we evaluated the classification performance on the cropped single-cell images. Table 6 presents the comparison between the ResNet50 baseline and the DinoBloom foundation model on both the internal PCM-Leukemia set and the external ALL-IDB1 benchmark.

On the internal test set, the ResNet50 baseline achieved an F1-score of 82.95% and an accuracy of 87.94%. The DinoBloom model, when fine-tuned using the Linear Probing plus

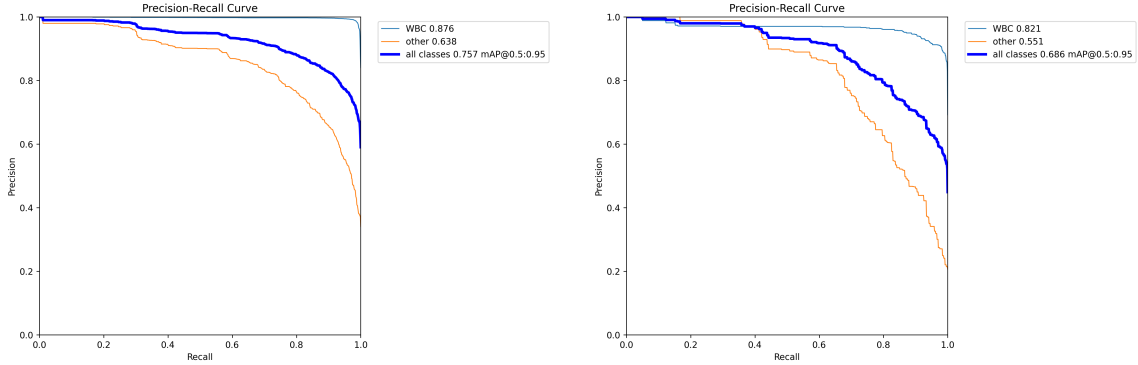


Figure 2: Precision-Recall curves for the binary detection stage (WBC vs. Other). The left panel evaluates performance on the internal PCM-Leukemia hold-out set, while the right panel shows results on the external ALL-IDB1 benchmark. Note the consistently high Average Precision (AP) for the target “WBC” class (87.6% and 82.1%) compared to the significantly lower performance on the “Other” class (63.8% and 55.1%). This performance gap illustrates that the overall mAP is suppressed by artifacts, while the detector remains highly reliable for white blood cells.

Fine-Tuning (LP+FT) strategy, outperformed the baseline with an F1-score of 92.11% and an accuracy of 92.59%.

When evaluated on the external ALL-IDB1 dataset, the ResNet50 model showed a decline in performance, recording an F1-score of 56.08%. In contrast, the DinoBloom (LP+FT) model maintained a higher generalization capability, achieving an F1-score of 74.27% and an accuracy of 88.02%. Additionally, comparing the fine-tuning strategies for DinoBloom, the two-step LP+FT approach yielded higher internal metrics (92.11% F1) compared to direct fine-tuning (87.96% F1).

Table 6: Performance comparison of Stage 2 classification models. We benchmarked a CNN baseline (ResNet50) against the DinoBloom foundation model using different fine-tuning strategies. The models were evaluated on the internal PCM-Leukemia holdout test set and the external ALL-IDB1 benchmark.

Model	Test Dataset	F1-Score (Macro)	Accuracy
ResNet50 (Baseline)	PCM-Leukemia (Internal)	82.95	87.94
	ALL-IDB1 (External)	56.08	66.62
DinoBloom (Direct FT)	PCM-Leukemia (Internal)	87.96	91.29
DinoBloom (LP+FT)	PCM-Leukemia (Internal)	92.11	92.59
	ALL-IDB1 (External)	74.27	88.02

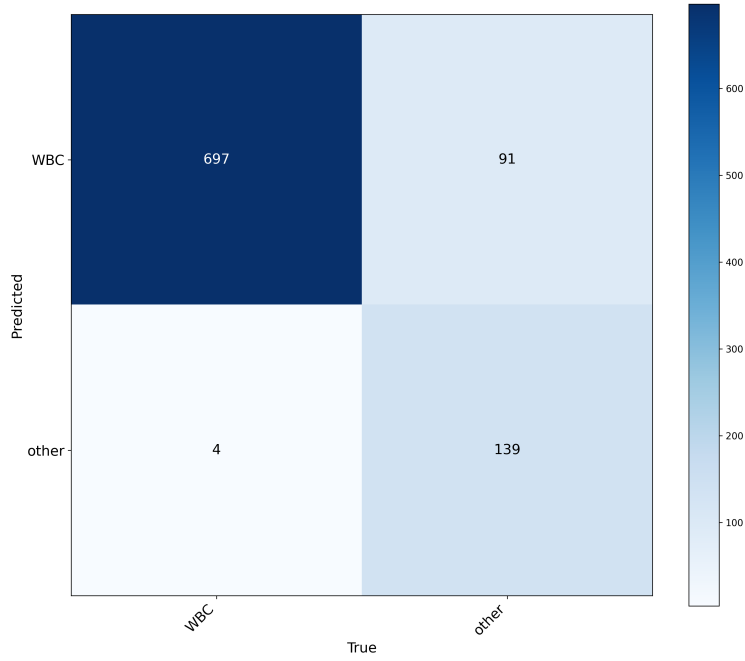


Figure 3: External validation results: Confusion matrix of the two-stage, WBC detector trained on PCM-Leukemia and evaluated on the independent ALL-IDB1 dataset

3.2.3. CROSS-DOMAIN ADAPTATION ON MUNICH AML MORPHOLOGY DATASET

We further evaluated the model’s adaptability on the Munich AML Morphology Dataset (LMU). As shown in Table 7, the direct application of the PCM-trained model (Zero-Shot) on the full LMU dataset resulted in an accuracy of 39.67% and an F1-score of 40.63%.

To address this severe domain shift, we applied partial fine-tuning using a random 10% subset of the Munich dataset. This adaptation strategy significantly improved performance on the remaining 90% hold-out set, increasing the accuracy to 81.94% and the F1-score to 67.94%.

Table 7: Evaluation of domain adaptation on the Munich AML Morphology Dataset (LMU). We compared the zero-shot performance (model trained solely on PCM-Leukemia) against a partial fine-tuning approach, where the model was adapted using only a random 10% subset of the LMU dataset.

Training Strategy	Test Dataset	F1-Score (Macro)	Accuracy
Zero-Shot (PCM-Only)	LMU (Full)	40.63	39.67
Partial FT (PCM+10% LMU)	LMU (Hold-out 90%)	67.94	81.94

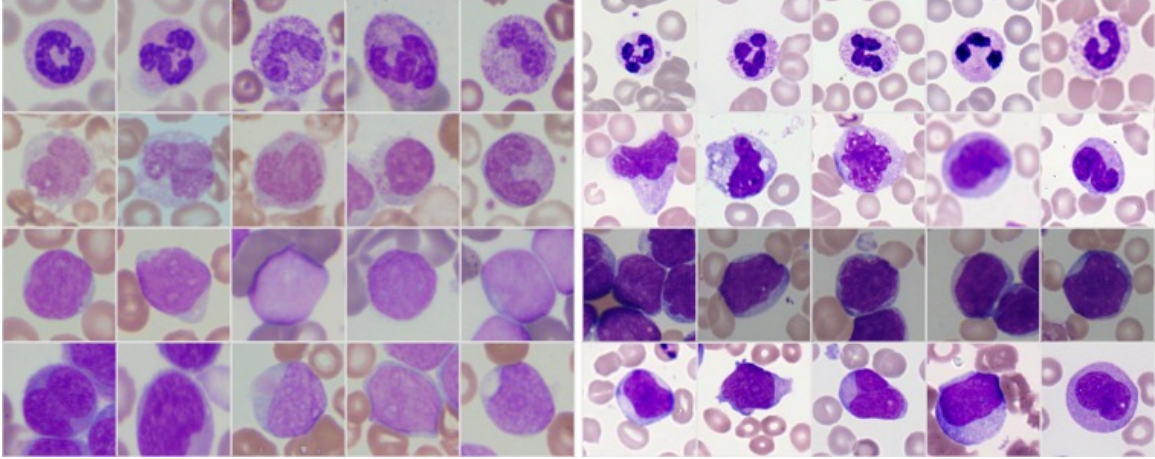


Figure 4: Representative comparison of morphological variability between the internal and external datasets. The left panel displays samples from our PCM-Leukemia dataset, while the right panel shows corresponding cell types from the external benchmarks: LMU (Rows 1, 3, and 4) and ALL-IDB1 (Row 2). The grid is organized by cell type: (Row 1) Neutrophils, (Row 2) Monocytes, (Row 3) Lymphoblasts, and (Row 4) Myeloblasts. Note the visible differences in staining characteristics, background noise, and illumination between the internal and external sources, illustrating the domain shift challenges addressed in this study.

4. Discussion

4.1. Dataset Robustness in Detection Tasks

A primary objective of this study was to benchmark the utility of the PCM-Leukemia dataset for real-world detection tasks. Our experiments revealed a critical insight regarding the dataset’s annotation quality: while single-stage detectors (YOLO11m, DEIM) experienced drops in fine-grained classification accuracy on external data, they maintained high recall in distinguishing White Blood Cells (WBCs) from the background. This confirms that the bounding box annotations in PCM-Leukemia capture robust, domain-invariant “objectness” features. This reliability allows future researchers to use our dataset as a trusted source for training high-recall WBC locators, which is the foundational step for any automated hematology pipeline.

4.2. Enabling Cross-Domain Generalization

The true value of a dataset lies in its ability to train models that generalize beyond the training distribution. Our benchmarking demonstrated that PCM-Leukemia is sufficiently diverse to effectively fine-tune large-scale foundation models. When we utilized the dataset to train the DinoBloom architecture, the resulting model achieved significant performance gains on the unseen ALL-IDB1 benchmark (F1-score of 74.27% vs. 56.08% for the baseline).

This result validates PCM-Leukemia as a high-quality downstream task for pathology foundation models, proving that it contains the necessary morphological diversity to bridge the domain gap between different staining protocols and scanners.

4.3. Facilitating Low-Resource Adaptation

We further validated the utility of PCM-Leukemia by attempting to address the severe domain shift present in the Munich AML dataset. Our experiments demonstrated that a model pretrained on PCM-Leukemia could recover over 80% accuracy on the LMU dataset using only 10% of the target data for adaptation. While this highlights the transfer learning potential of our dataset to significantly reduce the annotation burden for new clinical sites, we acknowledge that the domain gap is not fully resolved. Consequently, PCM-Leukemia serves as a robust “source domain” foundation, offering a promising baseline for future research to develop more advanced techniques for tackling heterogeneity in laboratory environments.

5. Conclusions

In this work, we introduce PCM-Leukemia, a novel, large-scale dataset designed to address the scarcity of diverse, expert-annotated data in hematology. By benchmarking various pipeline strategies, we demonstrated that our dataset enables the training of models capable of robust WBC localization and effective cross-domain generalization. Specifically, we showed that training on PCM-Leukemia allows for strong direct transfer to external ALL datasets and facilitates rapid, low-resource adaptation for AML diagnostics. While fully automated cross-domain diagnosis remains a challenge, PCM-Leukemia establishes a critical, publicly available baseline, lowering the barrier for future research and accelerating the development of reliable, AI-driven diagnostic tools.

6. Future Work

While our current pipeline effectively addresses many challenges in leukemia identification, significant hurdles remain due to major domain shifts caused by varied collection protocols, technician variances, and diverse equipment reliability. Although the fundamental morphology of blood cells remains consistent, these extrinsic variations necessitate more advanced mitigation strategies. In future work, we aim to enhance the granularity and information density of our dataset through labor-intensive annotation methods. Specifically, we plan to incorporate full-cell segmentation—similar to the protocols in (Mourya et al., 2019)—along with detailed morphological attribute tagging for the nucleus and cytoplasm (Rehman et al., 2024).

Furthermore, to fundamentally address the domain shift, we plan to expand our data collection to include heterogeneous sources from multiple institutions, thereby increasing the intrinsic diversity of the training distribution. Finally, we will investigate advanced online and offline augmentation strategies, such as class-level color normalization and aggressive color jittering, to further improve the model’s robustness against staining variability.

Acknowledgments

This project is funded by the Health Systems Research Institute (HSRI) under funding number 68-086.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Data Availability

The official implementation code is publicly available on GitHub at <https://github.com/l-kuo/pcm-leukemia>. The dataset is hosted at <https://qnap-2.aicenter.dynu.com/share.cgi?ssid=bd169009b6d048c6bfa802043baa6601>

References

- Andrea Acevedo, Anna Merino, Santiago Alf  rez,   ngel Molina, Laura Bold  , and Jos   Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020. ISSN 2352-3409. doi: <https://doi.org/10.1016/j.dib.2020.105474>. URL <https://www.sciencedirect.com/science/article/pii/S2352340920303681>.
- Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
- Laura Bold  , Anna Merino, Andrea Acevedo, Angel Molina, and Jos   Rodellar. A deep learning model (alnet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images. *Computer Methods and Programs in Biomedicine*, 202:105999, 2021.
- Corporation CVAT.ai. Computer vision annotation tool (cvat), June 2023. URL <https://doi.org/10.5281/zenodo.8070041>.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer visions and [pattern recognition]*, pages 770–778, 2016.

- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.
- Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. Deim: Detr with improved matching for fast convergence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15162–15171, 2025.
- Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. URL <https://github.com/ultralytics/ultralytics>.
- Valentin Koch, Sophia J. Wagner, Salome Kazemina, Ece Sancar, Matthias Hehr, Julia A. Schnabel, Tingying Peng, and Carsten Marr. DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology . In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.
- Zahra Mousavi Kouzehkanan, Sepehr Saghari, Sajad Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmail Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, et al. A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. *Scientific reports*, 12(1):1123, 2022.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.
- Ruggero Donida Labati, Vincenzo Piuri, and Fabio Scotti. All-idb: The acute lymphoblastic leukemia image database for image processing. In *2011 18th IEEE International Conference on Image Processing*, pages 2045–2048, 2011. doi: 10.1109/ICIP.2011.6115881.
- Christian Matek, Simone Schwarz, Karsten Spiekermann, and Carsten Marr. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*, 1(11):538–544, 2019.
- Simmi Mourya, Sonaal Kant, Pulkit Kumar, Anubha Gupta, and Rita Gupta. All challenge dataset of isbi 2019 (c-nmc 2019), 2019. URL <https://www.cancerimagingarchive.net/collection/c-nmc-2019/>.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

- Abdul Rehman, Talha Meraj, Aiman Mahmood Minhas, Ayisha Imran, Mohsen Ali, and Waqas Sultani. A large-scale multi domain leukemia dataset for the white blood cells detection with morphological attributes for explainability. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer, 2024.
- Piya Rujkijyanont and Hiroto Inaba. Diagnostic and treatment strategies for pediatric acute lymphoblastic leukemia in low-and middle-income countries. *Leukemia*, 38(8):1649–1662, 2024.
- Maneela Shaheen, Rafiullah Khan, Rajesh Roshan Biswal, Mohib Ullah, Atif Khan, M Irfan Uddin, Mahdi Zareei, and Abdul Waheed. Acute myeloid leukemia (aml) detection using alexnet model. *Complexity*, 2021(1):6658192, 2021.
- Naveed Syed, Mohamed Eltag Salih Saeed, Shakir Hussain, Imran Mirza, Amira Mahmoud Abdalla, Eiman Ahmed Al Zaabi, Imrana Afroz, Shahrukh Hashmi, and Mohammad Yaqub. Novel hierarchical deep learning models predict type of leukemia from whole slide microscopic images of peripheral blood. *Journal of Medical Artificial Intelligence*, 8(0), 2024. ISSN 2617-2496. URL <https://jmai.amegroups.org/article/view/9379>.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- James Homer Wright. A rapid method for the differential staining of blood films and malarial parasites. *Journal of Medical Research*, 7(1):138–144, 1902.
- Geng Yan, Gao Mingyang, Shi Wei, Liang Hongping, Qin Liyuan, Liu Ailan, Kong Xiaomei, Zhao Huilan, Zhao Juanjuan, and Qiang Yan. Diagnosis and typing of leukemia using a single peripheral blood cell through deep learning. *Cancer Science*, 116(2):533–543, 2025.
- Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Abbreviations

Abbreviation	Full Term
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
APL	Acute Promyelocytic Leukemia
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CNN	Convolutional Neural Network
DL	Deep Learning
FT	Finetuning
IRB	Institutional Review Board
LP	Linear Probing
PBS	Peripheral Blood Smear
SOTA	State of the Art
ViT	Vision Transformer
WBC	White Blood Cell

Appendix A. Blood Cells Descriptions from PCM-Leukemia

Figure 5 outlines the specific characteristics used to label cell images. It includes representative examples of normal white blood cells (Lymphocyte, Neutrophil, Basophil, Eosinophil, Monocyte) and leukemic cells (Lymphoblast, Myeloblast/AML, Promyelocyte/APL). Key diagnostic features regarding nuclear structure, cytoplasm appearance, and granularity are detailed for each class.

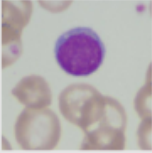
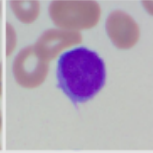
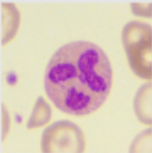
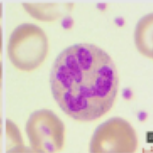
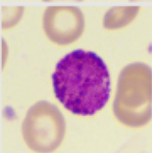
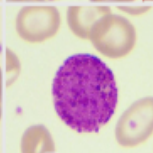
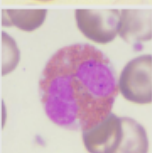
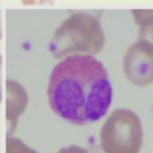
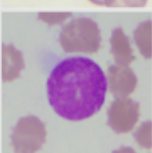
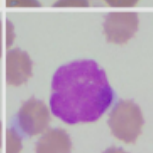
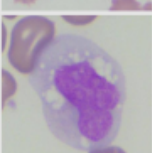
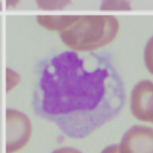
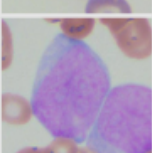
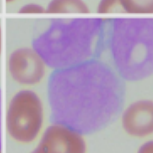
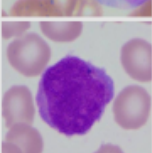
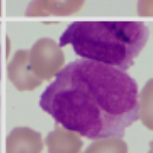
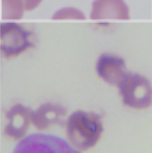
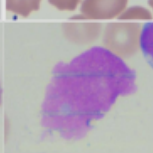
		<p>Class Name: Lymphocyte</p> <p>Nucleus: Round or slightly indented; dense, clumped (dark purple) chromatin.</p> <p>Cytoplasm: Scanty, rim-like, pale blue.</p> <p>Size: Small (similar to red blood cell size)</p>
		<p>Class Name: Neutrophil</p> <p>Nucleus: Multi-lobed (2–5 lobes) connected by thin filaments.</p> <p>Cytoplasm: Pale pink or salmon-colored.</p> <p>Granules: Fine, faint pink/purple granules (often indistinct).</p>
		<p>Class Name: Basophil</p> <p>Nucleus: Bi-lobed or S-shaped, but often obscured by granules.</p> <p>Cytoplasm: Clear to pinkish (usually hidden).</p> <p>Granules: Large, coarse, deep blue-black granules that overlay the nucleus.</p>
		<p>Class Name: Eosinophil</p> <p>Nucleus: Bi-lobed (spectacle-shaped).</p> <p>Cytoplasm: Pale blue or colorless background.</p> <p>Granules: Large, uniform, bright orange-red (refractile) granules; do not obscure the nucleus.</p>
		<p>Class Name: Lymphoblast</p> <p>Nucleus: Round or oval; coarse or slightly clumped chromatin; indistinct or small nucleoli (0–2).</p> <p>Cytoplasm: Scanty, deep blue, agranular.</p> <p>Size: Generally smaller than myeloblasts; high nuclear-to-cytoplasmic (N:C) ratio.</p>
		<p>Class Name: Monocyte</p> <p>Nucleus: Large, irregular, kidney-bean or horseshoe shaped; "lacy" or delicate chromatin.</p> <p>Cytoplasm: Abundant, dull gray-blue (ground-glass appearance).</p> <p>Features: Often contains vacuoles (white holes) and fine, dust-like granules.</p>
		<p>Class Name: Myeloblast(AML)</p> <p>Nucleus: Round or oval; fine, delicate (lacy) chromatin; prominent nucleoli (2–5).</p> <p>Cytoplasm: Scanty to moderate, blue.</p> <p>Features: May contain fine azurophilic granules or single Auer rods (needle-like inclusions).</p>
		<p>Class Name: Promyelocyte (APL)</p> <p>Nucleus: Often bilobed, folded, or butterfly-shaped (reniform).</p> <p>Cytoplasm: Densely packed with heavy, dark azurophilic granules (hypergranular).</p> <p>Features: Frequently contains bundles of Auer rods ("faggot cells"); nucleus often obscured by heavy granulation.</p>
		<p>Class Name: Other</p> <ul style="list-style-type: none"> - Cells at the margin > 50% not visible - Smudge Cells - Artifacts

Figure 5: Classification guide and morphological criteria for the proposed PCM-Leukemia dataset.

Appendix B. Class Distribution of PCM-Leukemia Training Data

Figure 6 illustrates the total number of annotated instances for each of the nine categories, detailing the dataset’s composition across normal leukocytes and leukemic blasts.

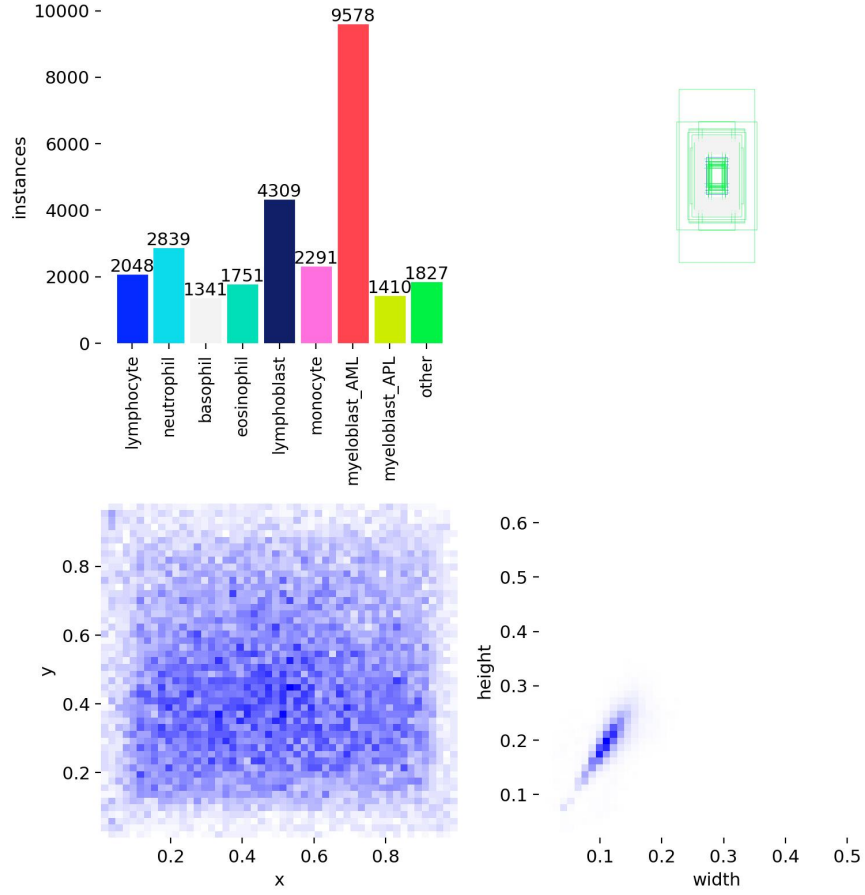


Figure 6: Class Distribution of PCM-Leukemia

Appendix C. One-stage Object Detector (YOLO11m) Training

Figure 7 depicts the training curves for the PCM-Leukemia Dataset, showing convergence at the 85th epoch. Additionally, Figure 8 reveals a mAP@0.5 of 0.973 on the validation split.

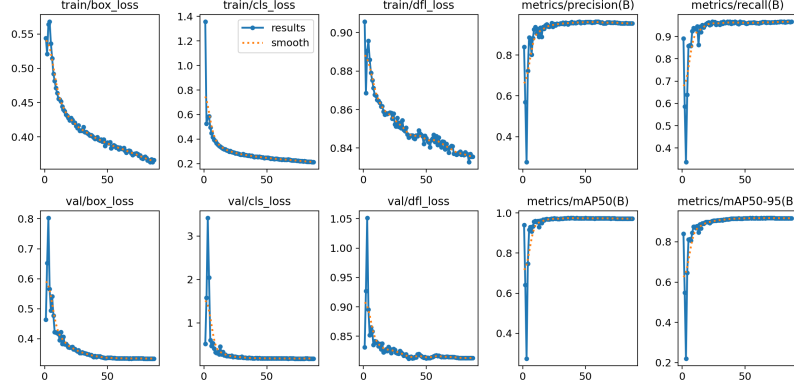


Figure 7: One-stage detector training graph.

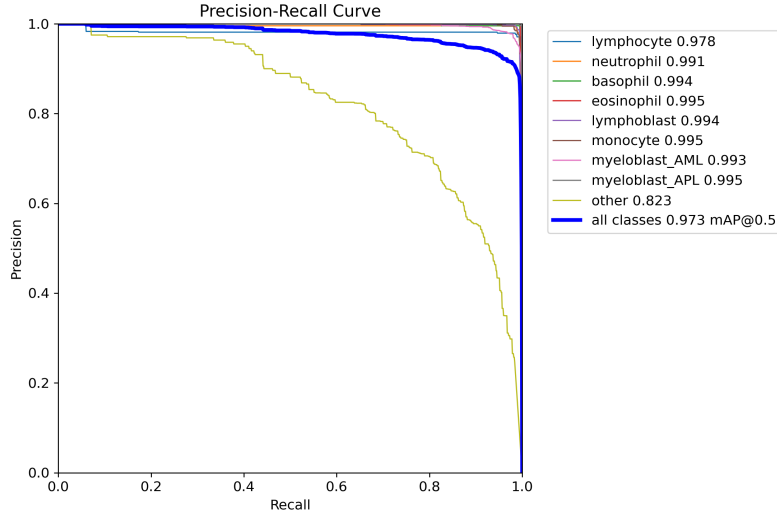


Figure 8: Precision-Recall curve of one-stage detector on PCM-Leukemia

Appendix D. Classification Model (ResNet50) from Two-stage pipeline

As shown in Figure 9, the ResNet50 model stabilizes at 83 epochs but is limited to an accuracy of 66%.

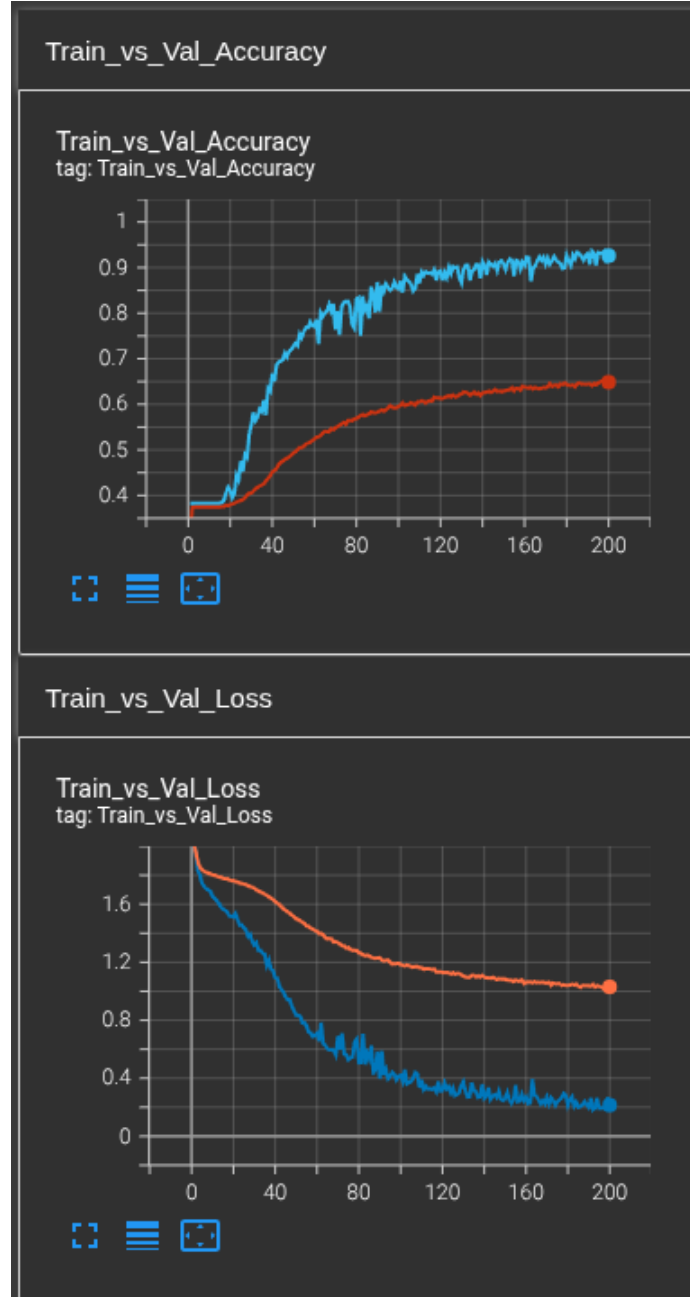


Figure 9: Training and validation graphs for two-stage classification model (ResNet50).

Appendix E. Detection Model (DEIM) from one-stage pipeline

Figure 10 reveals a higher error rate for lymphoblasts, which are predominantly misclassified as AML and APL.

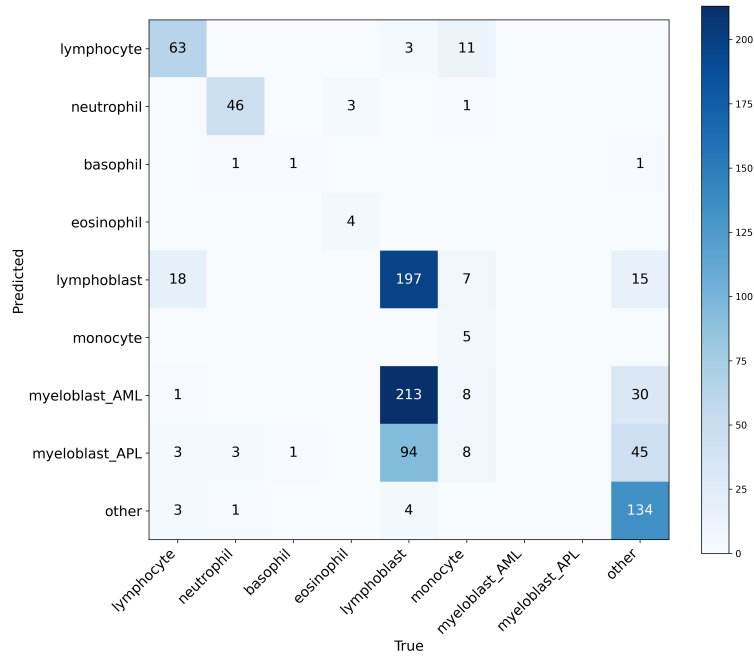


Figure 10: External validation results: Confusion matrix of the one-stage detector, DEIM, trained on PCM-Leukemia and evaluated on the independent ALL-IDB1 dataset.