scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Large-Scale Peripheral Blood Cell Dataset for Automated Hematological Analysis

Atıf Eren Yarıkan[1,2], Can Örer[3], Volkan Akyıldız[3], Zeki Kuş[4] ✉, Musa Aydin[5], Kerim Erhan Palaoğlu[6], Said İncir[3,7], Kemal Baysal[7], Cemal Özçelik[2], Berna Kiraz[2,8] & Alper Kiraz[2,9,10]

White blood cell classification is fundamental to hematological diagnosis, yet existing datasets are limited in scale and class diversity. We present a comprehensive peripheral blood cell dataset comprising 31,489 high-resolution microscopic images across 13 distinct cell classes, representing the largest publicly available collection for automated blood cell analysis. Images are acquired using the Sysmex DI-60 system from May-Grünwald-Giemsa-stained blood smears at 100 × magnification under standardized laboratory conditions. Expert hematologists with over 10 years of experience performed manual annotation with high inter-rater agreement (Cohen's kappa >0.85 for all classes). The dataset includes common cell types such as segmented neutrophils and lymphocytes, alongside diagnostically critical but rare subtypes, including myelocytes, blasts, and reactive lymphocytes. Images are organized into training, validation, and test splits (70:10:20 ratio) with consistent 368 × 368 pixel resolution. Baseline experiments using 14 deep learning architectures demonstrate the dataset's utility, with DenseNet-121 achieving 95.23% accuracy. KU-Optofil PBC Dataset addresses critical gaps in medical image analysis datasets and supports the development of robust automated hematology systems for clinical applications.

Background & Summary

Medical Context. Peripheral blood cells (PBCs) are a cornerstone of the immune system, and their counts and types provide critical insights into a patient's health[1]. The "differential count" of PBCs, which measures the proportions of leukocyte subtypes, is a standard laboratory test used in clinical practice that helps diagnose a variety of illnesses. Changes in PBC distribution often signal underlying issues: for example, bacterial infections typically raise neutrophil counts, allergic reactions elevate basophils, and certain leukemias present with an excess of immature "blast" cells[2]. Accurate classification of PBC subtypes is therefore vital for early diagnosis, guiding treatment decisions, and even monitoring disease progression[1]. In clinical hematology, PBC analysis is a fundamental component of routine diagnostic evaluation. Identifying leukocyte abnormalities—whether morphological or quantitative—is critical for diagnosing hematological malignancies, anemias, and immune-mediated disorders. For example, the detection of circulating leukemic blast cells or pronounced lymphocytosis can serve as critical diagnostic markers for specific leukemia subtypes. Given the substantial diagnostic implications, differential PBC counts are routinely included in nearly all complete blood count (CBC) analyses performed in hospital settings[1,2]. Traditionally, this analysis has been conducted either manually by qualified healthcare professionals or through the use of automated hematology analyzers. However, manual review is time-consuming and subject to inter-observer variability, while basic automated counters can flag quantitative abnormalities but may miss subtle morphological cues. Errors or delays in PBC classification can lead to misdiagnosis or inappropriate therapy, directly impacting patient outcomes.

[1]Computer Engineering, Fatih Sultan Mehmet Vakıf University, İstanbul, Türkiye. [2]Optofil A.Ş., İstanbul, Türkiye. [3]Clinical Laboratory, Koç University Hospital, Istanbul, Türkiye. [4]AI and Data Engineering, Fatih Sultan Mehmet Vakıf University, İstanbul, Türkiye. [5]AI and Data Engineering, Samsun University, Samsun, Türkiye. [6]Clinical Laboratory, VKV American Hospital, Istanbul, Türkiye. [7]Department of Medical Biochemistry, Koç University School of Medicine, Istanbul, Türkiye. [8]AI and Data Engineering, Istanbul Technical University, İstanbul, Türkiye. [9]Department of Physics, Koç University, Istanbul, Türkiye. [10]Department of Electrical and Electronics Engineering, Koç University, Istanbul, Türkiye. ✉e-mail: zkus@fsm.edu.tr

Dataset	Year	Task(s)	No. of. distinct PBC's	PBC Classes
IDB[3]	2005	Classification	510	2
IDB2[4]	2011	Segmentation, Classification	260	2
LISC[5]	2011	Segmentation, Classification	250	5
Munich[6]	2020	Classification	18,365	15
Raabin[7]	2022	Segmentation, Classification, Detection	17,965	5
HRLS[8]	2023	Classification	16,027	9
WBCAtt[9]	2023	Classification	10,298	5
LeukemiaAttri[13]	2024	Detection	7,857	14
Ours	2025	Classification	**31,489**	13

Table 1. Comparison of publicly available peripheral blood cell (PBC) datasets for machine learning tasks.

This has driven interest in advanced image-based analysis: indeed, leveraging artificial intelligence (AI) for reliable PBC identification promises to enhance diagnostic accuracy and efficiency in hematology[2].

Current Dataset Limitations. Table 1 presents a comprehensive comparison of publicly available peripheral blood cell datasets used for automated analysis tasks. Early datasets such as IDB (2005)[3] and IDB2 (2011)[4] are relatively small, containing 510 and 260 samples respectively, with binary classification tasks. The LISC dataset (2011)[5], although similar in size with 250 cells, marked an important advancement by expanding to a 5-class classification problem. The field has evolved significantly over the past two decades, with recent datasets experiencing substantial increases in both size and complexity. A significant leap in scale and complexity occurred starting in 2020. The Munich dataset[6] dramatically increases the number of available cells to 18,365 and expands the classification task to 15 distinct classes. This provided a much richer and more challenging benchmark for developing sophisticated classification algorithms. Following this trend, the Raabin dataset (2022)[7] offered a similarly large collection of 17,965 cells and is notable for its multi-task utility, supporting segmentation, detection, and 5-class classification. In 2023, the HRLS[8] and WBCAtt[9] datasets continued this trend of providing substantial data, with 16,027 cells across 9 classes and 10,298 cells across 5 classes, respectively, both focusing on the classification task. Our proposed dataset represents a significant advancement in the field by comprising 31,489 peripheral blood cell samples across 13 distinct cell classes. This makes it the largest publicly available dataset for PBC classification, representing a 71% increase over the previous largest classification dataset (Munich). The substantial increase in dataset size, combined with comprehensive class coverage, addresses the critical need for large-scale, diverse training data. This enhancement can support the development of more robust and generalizable machine learning models for automated blood cell analysis.

Dataset Motivation. The microscopic analysis of peripheral blood smears, particularly in counting white blood cells (WBCs), is a critical aspect of clinical hematology. This procedure is key to the accurate diagnosis and monitoring of various medical conditions, including infections, inflammatory diseases, and blood cancers such as leukemia. The use of computer-aided diagnostic systems presents a promising approach to enhance and streamline this traditionally time-consuming and subjective process. These systems can improve efficiency and reliability by automating and standardizing the analyses. However, to maximize their effectiveness, it is crucial to prioritize the quality, scale, and diversity of the datasets used for training and validation, as this will significantly enhance the performance and reliability of these automated systems in a clinical setting. Although PBC datasets can be found in the literature, they have several limitations. These include a generally limited number of cell classes, a restricted total number of images, and insufficient representation of rarer and clinically critical cell subtypes[10–12]. To address these critical gaps, a large-scale comprehensive dataset of PBCs has been created. The dataset contains more than 31,000 images across 13 different classes, including common types such as segmented neutrophils and lymphocytes, as well as diagnostically important but less common types, including myelocytes, platelet cluster, and reactive lymphocytes Fig. 1.

Dataset Overview. The dataset presented in this work is a large-scale collection of 31,489 digital microscope images of PBCs, organized into 13 distinct classes. These classes encompass a comprehensive range of white blood cells, their precursors, and other associated cell types, including Band Neutrophil, Basophil, Blast, Eosinophil, Erythroblast, Giant Platelet, Lymphocyte, Metamyelocyte, Monocyte, Myelocyte, Platelet Cluster, Reactive Lymphocyte, and Segmented Neutrophil. The distribution of the number of images belonging to each class in the dataset is given in Table 2, and sample images belonging to each class in the dataset are also shown in Fig. 2. We have also included a metadata.csv file to enhance dataset usability and enable rigorous experimental designs. This file maps each image to its class label, filename, and anonymized patient identifier. This structured metadata allows researchers to perform patient-level cross-validation, where all images from a given patient are assigned exclusively to either the training, validation, or test set. The metadata file also facilitates studies of inter-patient variability and its impact on model performance. It enables researchers to analyze classification accuracy across different patient populations and develop more robust automated hematology systems suitable for clinical deployment.

Potential Applications. The primary purpose of the dataset is to support the development and thorough validation of automated hematology systems. These tools can significantly decrease the time and subjectivity involved in manual blood smear analysis, enabling quicker and more reliable diagnoses of hematological malignancies and infections. For the computer vision community, it serves as a challenging new benchmark, with its

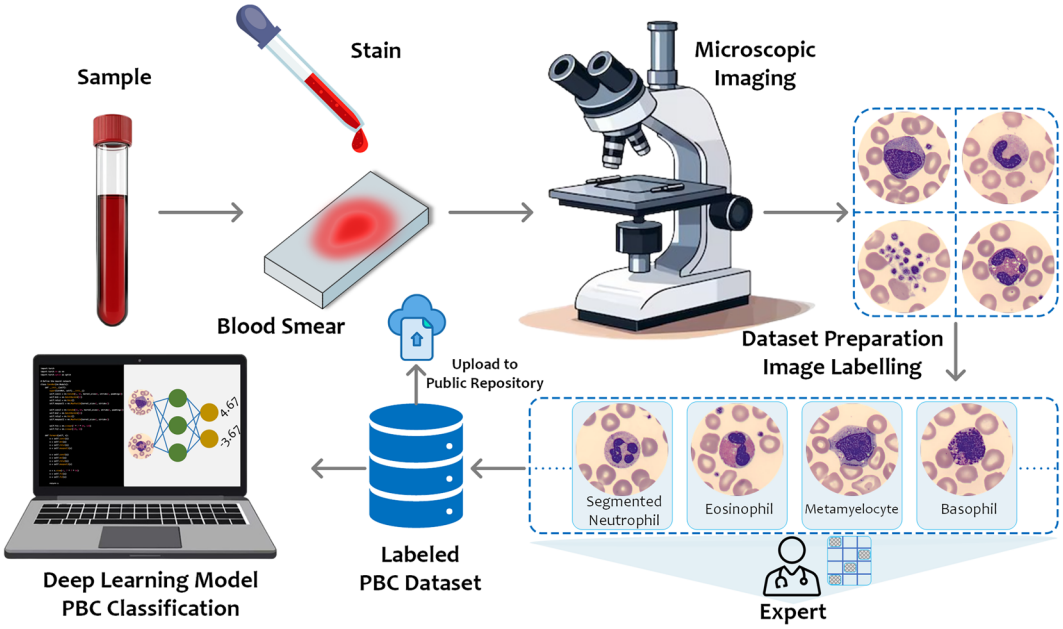

Fig. 1 Workflow for peripheral blood cell (PBC) classification using deep learning. This flowchart illustrates the step-by-step process for developing a deep learning model to classify PBCs from blood samples. The process begins with the collection of a blood sample, which is then used to prepare a blood smear on a glass slide. The smear is stained to highlight cellular components, making PBCs more visible under a microscope. Next, microscopic imaging is performed to capture high-resolution images of the stained blood smear. These images are then used for dataset preparation, where experts label each PBC according to its type (e.g., segmented neutrophil, eosinophil, metamyelocyte, basophil). The labeled images are compiled into a PBC dataset, which can be uploaded to a public repository for broader access. Finally, this labeled dataset is used to train and evaluate a deep learning model for the automated PBC classification.

White Blood Cell Class	Number of Images
Band Neutrophil	199
Basophil	546
Blast	816
Eosinophil	1,862
Erythroblast	243
Giant Platelet	2,835
Lymphocyte	5,764
Metamyelocyte	262
Monocyte	1,381
Myelocyte	98
Platelet Cluster	188
Reactive Lymphocyte	1,039
Segmented Neutrophil	16,256
Total	**31,489**

Table 2. Distribution of images across the 13 peripheral blood cell classes in the dataset.

severe, clinically realistic class imbalance and fine-grained classes providing an ideal testbed for advancing algorithms in complex recognition tasks, such as imbalanced learning and fine-grained visual classification. Finally, the dataset is an invaluable educational asset, forming the basis for digital atlases and training modules that can help medical students and hematology trainees master the identification of a wide spectrum of diagnostically critical white blood cells, including rare subtypes they may not frequently encounter in practice.

Methods

Data Preparation. *Dataset Selection and Standardization.* Peripheral blood smear images are collected retrospectively using the Sysmex DI-60 system (Sysmex Corporation, Japan), a widely adopted automated digital morphology analyzer. All images are acquired under standardized laboratory conditions at the Koç University Hospital laboratory using May-Grünwald-Giemsa-stained smears. The study has been approved by the Koç

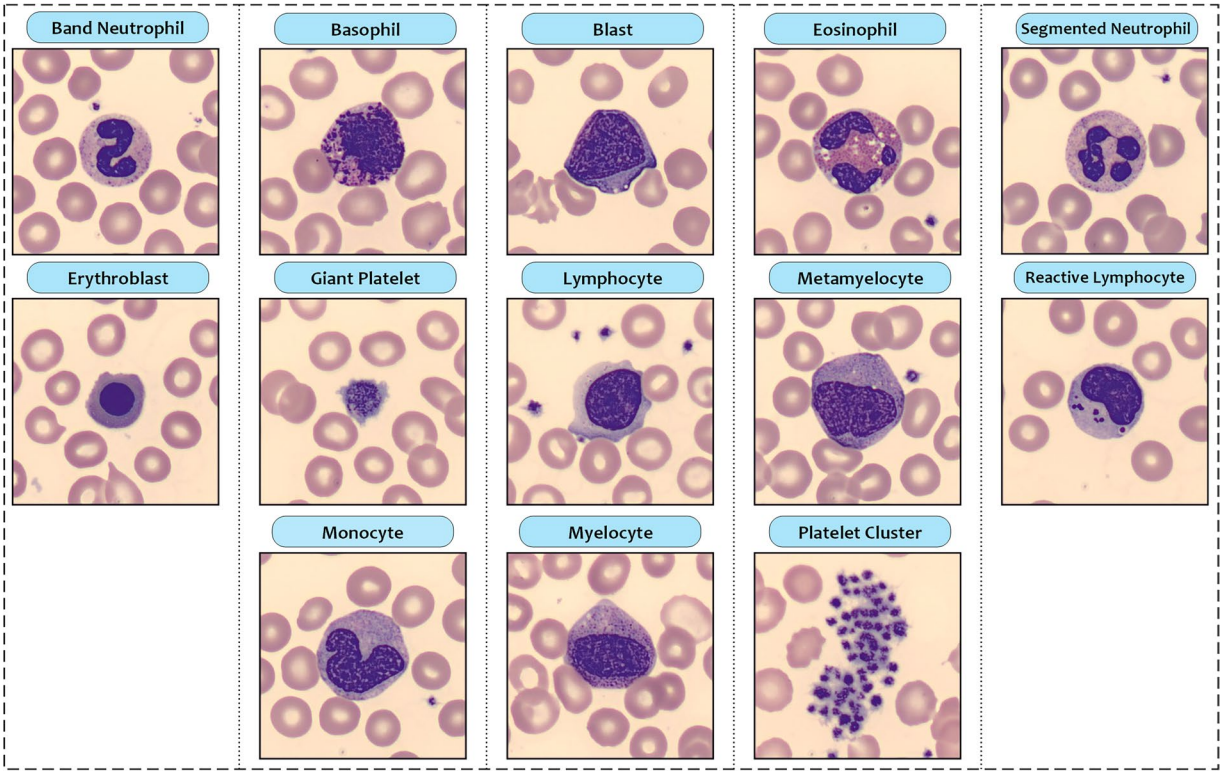

Fig. 2 An overview of the visual diversity of the data set, with sample images of each of the 13 classes.

University Clinical Research Ethics Committee (approval no. 2025.153.IRB1.023). The data are collected retrospectively from routine laboratory practice and fully anonymized before annotation and public release. The ethics approval covers the use and public release of these nonidentifying data, and consent requirements are addressed by the ethics committee under this approval and the published license. Images are captured at 100 × magnification and stored in JPEG format with fixed pixel dimensions[1,2]. Only cells that could be clearly visualized and morphologically classified are included. Images with artifacts, blurriness, overlapping cells, or staining issues are excluded to ensure consistency and reliability in subsequent model training and evaluation.

Expert Annotation and Quality Control. Each image is independently labeled by two expert laboratory technicians, each with over 10 years of professional experience in hematology. In cases where the initial annotations are discordant, a third expert technician with similar seniority served as an adjudicator. Final ground truth labels were assigned based on majority agreement (2 out of 3 experts). This consensus method helps minimize annotation bias and reflects current clinical practice. To quantify the agreement between reviewers, Cohen's kappa (K) coefficient is calculated. This statistic provides a measure of inter-rater reliability beyond chance and is widely used in medical image annotation studies[1,14]. High K values (>0.80) indicate strong agreement, ensuring robust reference standards for training and evaluating machine learning models[15].

Inclusion and Exclusion Criteria. *Inclusion criteria:* (1) Single, well-centered leukocytes in focus; (2) No cell overlaps or artifacts; (3) Proper staining and intact morphology. *Exclusion criteria:* (1) Out-of-focus or low-resolution images; (2) Multiple or overlapping cells; (3) Cytoplasmic or nuclear artifacts; (4) This curation process was essential for ensuring high-quality annotations and reducing noise in model training.

Imaging Details. All image acquisition procedures followed standardized protocols as implemented on the Sysmex DI-60. Captured images are anonymized prior to annotation, removing all patient identifiers and linking information.

Inter-Rater Agreement Analysis. To assess the inter-observer agreement for individual WBC types, Cohen's kappa (K) coefficients are calculated separately for each cell class. For this purpose, a binary 2 × 2 confusion matrix is constructed per class, where each cell type is treated as the positive class and all other types as negative. The values used in the computation—true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN)—are derived from the manual comparison between the reference labels and the independent review by a second expert. The results are presented (Cell Type - Cohen's Kappa): Segmented Neutrophil - 0.99; Band Neutrophil - 0.85; Lymphocyte - 0.99; Monocyte - 0.99; Basophil - 0.99; Eosinophil - 1.00; Myelocyte - 0.98; Metamyelocyte - 0.94; Blast - 1.00; Reactive Lymphocyte - 0.99. All evaluated cell types demonstrated substantial to almost perfect agreement according to Landis and Koch's interpretation scale ($K > 0.81$)[15]. Notably, basophils, eosinophils, and blasts achieved perfect agreement ($K = 1.00$), indicating absolute consistency between

observers. Band neutrophils, typically more prone to interpretational variability due to their morphological overlap with segmented neutrophils, still reached a high kappa value of 0.85, reflecting strong agreement.

Image Resizing and Label Mapping. All images are resized to a uniform resolution of 368 × 368 pixels to ensure consistency during the training process. However, Vision Transformer (ViT) models require images to be resized to 224 × 224 pixels. Resizing is performed using a bilinear interpolation method to preserve spatial information. All images are provided in RGB color space and in a tensor format, without any normalization. No additional pre-processing steps or data augmentation techniques are applied. Class labels are automatically extracted from the folder names, resulting in 13 distinct categories (see Table 2). The images available on Zenodo are in their original, unprocessed form, corresponding directly to what was captured by the Sysmex DI-60 system. Resize operations are only performed during our baseline experiments. These resizing steps are not applied to the images stored in the repository, which keep their original resolution as produced by the imaging device.

Train/Validation/Test Splits. An image-level stratified splitting strategy is employed to ensure that the classes are evenly distributed across the training, validation, and test sets. Images from each class are split according to a 7:1:2 ratio for the training, validation, and test sets, respectively. Consequently, the entire dataset is partitioned into training, validation, and test sets with proportions of 70%, 10%, and 20%, respectively. Additionally, we provide a patient-based splits file that ensures all images from the same patient remain in the same split. This approach supports more rigorous evaluation at the patient level. Compared with image-level splits, patient-level splits tend to exhibit greater class imbalance because different patients contribute different numbers of images across cell types. This happens because different patients contribute different numbers of cells, and not all cell types are present for every patient. As a result, some rare classes may become even less represented in a given split, thereby affecting generalization performance and making comparisons across studies more difficult. Furthermore, maintaining patient-level separation limits the ability to balance classes at a finer level. A key feature of our dataset is that morphological diversity within white blood cell classes mainly occurs within each class, not on a patient-by-patient basis. This is because cell appearance is primarily determined by cell type and maturation stage, rather than individual patient factors. Therefore, both splitting strategies are available via our metadata files, enabling researchers to choose the most appropriate approach based on their specific research goals.

The folder structure is presented in the Data Records section.

Data Records

The dataset has been made publicly available to support research in hematological image analysis and is stored in the Zenodo repository[16]. It contains images of peripheral blood cells belonging to 13 different classes, including 11 types of white blood cells and two types of platelets. All images are standardized to a uniform resolution of 368 × 368 pixels (with the exception of those used in ViT models), and are stored in the common JPG format to ensure a balance between image quality and manageable file sizes. The technical specifications are consistent across the entire dataset: each image is an RGB color file with a 24-bit depth (uint8 data type). Furthermore, the images have a resolution of 96 DPI. For ease of use and to facilitate supervised learning tasks, the data is intuitively organized into folders named according to their respective peripheral blood cell class. First, all images are sorted into folders named after their specific peripheral blood cell class. Second, we have partitioned the entire collection into training, validation, and testing sets to ensure a fair and reproducible model evaluation. This structure is preserved within each split, meaning the train, val, and test directories each contain the same 13 class-named subfolders. This consistent hierarchy makes it straightforward to load and manage the data for model training and analysis. The directory is structured as follows:

```
train
├── Blast
├── Monocyte
├── Myelocyte
└── ...(other classes)
val
├── Blast
├── Monocyte
├── Myelocyte
└── ...(other classes)
test
├── Blast
├── Monocyte
├── Myelocyte
└── ...(other classes)
```

The dataset is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), and the source code is released under the Apache License, Version 2.0.

Technical Validation

Baseline methods. We have selected a diverse range of well-established deep learning architectures as baselines to validate our datasets thoroughly. The selection is primarily based on the models evaluated in the MedSegBench study[17], which is important in terms of using the most common architectures in the medical image analysis literature. The selected models include various architectural categories, including traditional Convolutional Neural Networks (CNNs), lightweight efficient models, and modern transformer-based approaches. We have included several well-known CNNs as baseline models. The ResNet variants are chosen due to their widespread use in computer vision research. Specifically, we have selected ResNet-18, ResNet-34, and ResNet-50[18]. This range allows us to evaluate performance across different network depths, with ResNet-18 offering a good balance between depth and efficiency, and ResNet-50 being capable of learning more complex and abstract features. Additionally, DenseNet-121 is included for its unique architecture that promotes feature reuse and improves gradient flow through dense connections[19]. This design is especially parameter-efficient and is effective at capturing the fine-grained details in medical images. We have also used models designed for efficiency to evaluate performance within computational constraints. EfficientNetV2[20], MobileNetV2[21], and MobileNetV3[22] are chosen because they are state-of-the-art lightweight models that serve as effective alternatives to larger networks like ResNet and DenseNet, with minimal loss of accuracy. We also included MNASNet, which differs from handcrafted networks as it is derived from Platform-Aware Neural Architecture Search (AutoML)[23]. This method optimizes the architecture to specific hardware, enhancing efficiency. We have evaluated four MNASNet variants with depth multipliers of 0.5, 0.75, 1.0, and 1.3 to examine the balance between model size and performance. Finally, we have added three new models, the Vision Transformer (ViT)[24], Mobile Transformer[25] and Max Vision Transformer[26] to compare against the latest advancements in computer vision. Unlike traditional models that focus on small parts of an image, these models use a method called self-attention to understand the entire image at once. Including these models helps us compare different approaches to designing computer vision systems, indicating a shift towards using attention-based methods.

All baseline models are implemented using the torchvision library within the PyTorch ecosystem[27] except for Vision Transformer[28] and Mobile Vision Transformer[29]. They are trained from scratch on our dataset without using pre-trained ImageNet weights. This ensures that the evaluation accurately measures the models' capacity to learn features directly from the white blood cell images, rather than relying on pre-trained general-purpose knowledge. The experiments are performed on a single NVIDIA RTX 4090 GPU. Each model is trained for 200 epochs using the Adam optimizer with a learning rate of $1e$-3. The categorical cross-entropy loss function guides the training process. A batch size of 32 is used for all experiments, except for the Transformer-based models, where the batch size is reduced to 16 due to higher memory requirements. We have not applied any weight decay or data augmentation techniques to evaluate the raw architectural performance of each model. Throughout the training process, the model's performance is monitored on the validation set. The model weights corresponding to the epoch with the highest validation F1 score are saved for final evaluation on the test set. For complete transparency and reproducibility, the full implementation details, training scripts, and evaluation protocols are made publicly available in our code repository GitLab.

Performance Measures. In this study, classification performance is evaluated based on the following metrics: Accuracy, Recall, Precision, and F1-score. In the following equations, *TP* presents the true positives, *TN* presents the true negatives, *FP* presents the false positives, and *FN* presents the false negatives.

- **Accuracy (ACC)** measures the overall accuracy of the model by computing the ratio of correctly classified instances to the total number of instances:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Recall (REC)** measures the proportion of actual positive instances that are correctly classified by the model as positive.

$$REC = \frac{TP}{TP + FN} \tag{2}$$

- **Precision (PREC)** measures the correctness of positive predictions and is calculated as the ratio of true positives to all predicted positives.

$$PRE = \frac{TP}{TP + FP} \tag{3}$$

- **F1-Score** is defined as the harmonic mean of precision and recall.

In addition to these performance metrics, the confusion matrix is used to visually present the classification results of the best-performing models and to provide an in-depth, class-wise analysis of their performance.

Results. In this section, we present the results of our experiments using various deep learning models on our newly developed peripheral blood cell dataset. The evaluation is structured to provide a clear understanding of both the overall model performance and the specific challenges posed by the dataset's class distribution. We first compare the architectural trade-offs between traditional CNNs, mobile-optimized networks, and modern

Models	#FLOPs	#Params	Accuracy	Precision	Recall	F1 Score
DenseNet-121	7.62	6.96	**0.9523**	0.8283	0.7913	**0.8059**
ResNet-18	4.99	11.18	0.9493	0.8194	0.7604	0.7692
ResNet-34	10.04	21.29	0.9437	0.7800	0.7565	0.7613
ResNet-50	11.20	23.53	0.9387	0.7750	**0.8015**	0.7846
MobileNetV2	0.82	2.24	0.9444	0.8121	0.7832	0.7912
MobileNetV3	0.58	4.21	0.9441	0.7953	0.7439	0.7578
EfficientNetV2	7.82	20.19	0.9460	0.8315	0.7965	0.7974
Vision Transformer	0.20	204.71	0.7582	0.3775	0.2842	0.2757
Max Vision Transformer	2.01	30.41	0.9326	0.6994	0.6869	0.6917
Mobile Vision Transformer	0.84	4.94	0.9374	0.7615	0.7111	0.7284
MNASNet_0.5	**0.28**	**0.95**	0.9388	**0.8569**	0.7167	0.7385
MNASNet_0.75	0.59	1.90	0.9303	0.7312	0.6686	0.6810
MNASNet_1.0	0.87	3.11	0.9170	0.6909	0.6718	0.6719
MNASNet_1.3	1.47	5.01	0.9309	0.7428	0.7498	0.7357

Table 3. Performance and complexity comparison of different models. Best results for each metric are highlighted in bold. #FLOPs and #Params are reported in gigabytes and megabytes, respectively.

transformer-based models, highlighting key differences in accuracy, efficiency, and other performance metrics. Subsequently, we conduct a detailed class-wise analysis of the top-performing models to evaluate their effectiveness across both well-represented and rare cell types. These analyses not only establish strong performance benchmarks but also validate the technical quality and utility of the dataset for developing robust classification systems.

Based on the results (see Table 3), different network architectures show trade-offs between performance and efficiency. Among traditional convolutional networks, DenseNet-121 performs best, with an accuracy of 95.23%, precision of 82.83%, and F1-score of 80.59%. Its dense connectivity pattern enables efficient feature reuse, resulting in competitive computational efficiency with 7.62 GFLOPs and a relatively modest parameter count of 6.96M. The ResNet variants exhibit interesting scaling characteristics that challenge conventional assumptions about depth-performance relationships. ResNet-18, despite being the shallowest variant, achieves competitive accuracy (94.93%) with reasonable precision (81.94%), although it suffers from lower recall (76.04%) and F1-score (76.92%). Notably, ResNet-18 requires significantly more parameters (11.18M) than DenseNet-121 while achieving lower performance, highlighting the efficiency advantages of dense connections. ResNet-34 and ResNet-50 demonstrate decreasing returns with increased depth, where ResNet-34 achieves 94.37% accuracy with doubled computational cost (10.04 GFLOPs) and tripled parameters (21.29M) compared to ResNet-18. ResNet-50 also illustrates this trend, achieving only 93.87% accuracy despite requiring the highest computational resources in this group (11.2 GFLOPs, 23.53M parameters) This suggests that deeper networks may be overfitting or not trained optimally for this task.

Mobile-optimized architectures focus on efficiency. MobileNetV2 performs well, achieving 94.44% accuracy and requiring low computational resources (0.82 GFLOPs, 2.24M parameters). Its balanced performance across all metrics—precision (81.21%), recall (78.32%), and F1-score (79.12%)—demonstrates the effectiveness of depthwise separable convolutions and inverted residual blocks. This architecture achieves nearly comparable accuracy to much larger models while requiring orders of magnitude fewer computational resources, making it ideal for resource-constrained environments. MobileNetV3 reduces computational needs further (0.58 GFLOPs) and still achieves highly competitive accuracy (94.41%), but with a significant decrease in precision, recall, and F1-score. EfficientNetV2 strikes a balance between mobile efficiency and traditional architecture performance. It requires higher computational resources (7.82 GFLOPs, 20.19M parameters) but achieves superior precision (83.15%) and overall accuracy (94.60%) compared to other mobile architectures.

The transformer-based architectures reveal significant challenges in adapting vision transformers to this specific task, with performance substantially lagging behind convolutional counterparts. The standard Vision Transformer performs poorly across all metrics, achieving only 75.82% accuracy with extremely low precision (37.75%) and recall (28.42%), despite requiring massive parameter count (204.71M) with relatively low FLOPs (0.20G). This suggests severe underfitting or inadequate training methodology for the transformer architecture. Max Vision Transformer shows considerable improvement, achieving 93.26% accuracy with more balanced precision (69.94%) and recall (68.69%), though still requiring substantial parameters (30.41M) and computational resources (2.01 GFLOPs). Mobile Vision Transformer presents the best transformer performance with 93.74% accuracy and reasonable efficiency (0.84 GFLOPs, 4.94M parameters), demonstrating that mobile-optimized transformer designs can achieve competitive results, though still falling short of the best convolutional architectures. The extremely poor performance of the standard Vision Transformer in our experiments can be attributed to the fundamental data requirements of transformer-based architectures. Vision Transformers lack the built-in inductive biases of convolutional neural networks, such as translation equivariance and locality. Therefore, they require substantially larger datasets to learn spatial relationships effectively from scratch. While our dataset of 31,489 images is substantial for medical imaging applications, it remains orders of magnitude smaller than the datasets typically used for successful ViT training, such as ImageNet-21k or JFT-300M. The superior

performance of mobile-adapted transformer variants in our study is particularly evident when comparing the Mobile Vision Transformer to standard ViTs. This demonstrates that architectural modifications incorporating beneficial inductive biases can partially compensate for the limited availability of data. However, for datasets of this scale, convolutional architectures remain more appropriate, or alternatively, transfer learning approaches using pre-trained transformer models should be employed to leverage learned representations from large-scale corpora.

The MNASNet models provide valuable insights into how well neural architecture search works at different sizes. The smallest model, MNASNet_0.5, achieves the highest precision (93.88%) with very few resources (0.28 GFLOPs, 0.95M parameters). Moreover, it achieves the best accuracy and F1-scores compared to other MNASNet variants. The scaling behavior within the MNASNet models reveals concerning trends. Larger variants (MNASNet_0.75, MNASNet_1.0, and MNASNet_1.3) generally exhibit degraded performance despite increased computational requirements. MNASNet_1.0 represents the worst performer in terms of precision (69.09%) and F1-score (67.19%), achieving only 91.70% accuracy with 0.87 GFLOPs and 3.11M parameters. This suggests that making the models larger may not always result in improved performance.

The analysis shows that a model's efficiency doesn't always mean it performs worse. MobileNetV2 is very good at balancing performance and efficiency, while DenseNet-121 outperforms larger ResNet models. This suggests that for this task, how well a model reuses features and its architecture are more important than just size. Overall, DenseNet-121 yields the best performance, but MobileNetV2 is more suitable if you require a balance between efficiency and performance, especially in resource-constrained settings.

We also evaluated the class-wise performance of the top models from each group (traditional networks, Mobile-optimized, transformer-based, and MNASNet). The results are shown in Table 4 and displayed in Fig. 3. The class distribution within the dataset heavily influences the performance of the DenseNet-121, EfficientNetV2, MNASNet_0.5, and Mobile Vision Transformer architectures. A clear correlation exists between the number of training images available for a class and the models' ability to classify it accurately. For instance, classes with abundant samples, such as Segmented Neutrophil (16,256 images), Lymphocyte (5,764 images), and Eosinophil (1,862 images), are classified with high F1-scores, often near-perfect, across all three models. This suggests that, with sufficient data, all architectures are capable of effectively learning the distinguishing features of these cell types.

Conversely, the models struggle most directly with the most underrepresented classes. The Myelocyte class, with only 98 images, has the lowest F1-score, especially 0.08 with MobileVit and 0.09 with MNASNet_0.5. This poor performance is attributed to extremely low recall, indicating that the models failed to learn the features of Myelocytes due to a lack of training examples. A similar trend is observed for other minority classes like Band Neutrophil (199 images) and Metamyelocyte (262 images), where F1-scores are consistently among the lowest. Notably, MobileViT's F1-score (0.13) is substantially lower than that of the other models. This data imbalance creates a bias, where models are good at identifying majority classes but fail to generalize to minority classes.

Some classes are easier to identify because they look very different from others, even if they have fewer images. For example, Platelet Cluster (188 images) and Erythroblast (243 images) achieve high F1-scores despite having few samples, indicating their features are unique and easily learned. On the other hand, the Reactive Lymphocyte class, despite having a moderate sample size (1,039 images), presents a challenge for all models, resulting in average F1-scores. This suggests an inherent difficulty in distinguishing it from other lymphocyte types. The poor performance on certain rare classes cannot be attributed solely to the limited sample size but reflects the complex interplay between data availability and morphological variability. Blast cells, despite having only 816 samples, achieved high F1-scores (0.80-0.86) due to their distinctive and homogeneous morphological features, including large nuclei with fine chromatin and prominent nucleoli. In contrast, myelocytes (98 samples) showed substantially lower F1-scores (0.08-0.28) due to high intra-class morphological variability and overlapping features with adjacent maturation stages. Similarly, band neutrophils (199 samples) showed poor performance (F1: 0.13-0.39) because their distinction from segmented neutrophils relies on subtle nuclear segmentation patterns that exist on a morphological sequence. Reactive lymphocytes, despite having 1,039 samples, remained challenging (F1: 0.48-0.61) due to their heterogeneous appearance resulting from various activation states. These findings demonstrate that successful classification of rare cell types requires both adequate sample representation and morphologically distinct features. Future dataset expansion should prioritize morphologically variable classes. Additionally, explore advanced methods like generating synthetic data or adding more morphological features.

Additional experiments are conducted using weighted cross-entropy loss on DenseNet-121 to investigate the impact of class imbalance on model performance. In these experiments, weights are assigned in inverse proportion to class frequencies. This approach is intended to increase the penalty for misclassifying rare classes. The results are presented in Fig. 4. The comparative results reveal nuanced trade-offs between majority and minority class performance. The weighted loss approach showed modest improvements for certain rare classes, with band neutrophils increasing from 13 to 15 correct classifications out of 41 test samples, while platelet clusters showed minimal improvement from 34 to 35 out of 38 samples. However, this came at the cost of reduced overall accuracy (94.50% vs. 95.23%) and F1-score (79.89% vs. 80.59%), with some classes such as monocytes experiencing severe overprediction with 256 correct classifications but numerous false positives. These results demonstrate that weighted loss redistributes model attention from the majority to the minority classes but does not uniformly improve performance across all rare classes. The effectiveness varies substantially depending on class-specific characteristics such as morphological distinctiveness and intra-class variability. While the dataset enables realistic benchmarking, out-of-the-box model performance on rare classes can be limited. Future work may benefit from targeted augmentation, synthetic sample generation, or specialized imbalance learning methods to better capture the variability of these rare cell types.

Class Name	Precision				Recall				F1-Score			
	DN-121	EN_V2	MNAS	MobileViT	DN-121	EN_V2	MNAS	MobileViT	DN-121	EN_V2	MNAS	MobileViT
Band Neutrophil	0.52	**0.71**	0.67	0.21	**0.32**	0.24	0.2	0.10	**0.39**	0.36	0.3	0.13
Basophil	**0.99**	**0.99**	0.92	**0.99**	0.99	**1.00**	0.99	0.99	0.99	**1.00**	0.95	0.99
Blast	**0.88**	0.77	0.82	0.85	0.83	**0.91**	0.79	0.77	**0.86**	0.83	0.8	0.81
Eosinophil	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**1.00**	0.99	**1.00**	**1.00**	**1.00**	0.99	**1.00**
Erythroblast	0.94	0.9	**0.97**	0.79	**0.98**	0.92	0.76	0.76	**0.96**	0.91	0.85	0.77
Giant Platelet	**0.99**	**0.99**	0.97	0.97	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**0.99**	0.98	0.98
Lymphocyte	0.92	**0.93**	0.92	0.88	**0.94**	0.89	0.9	0.93	**0.93**	0.91	0.91	0.91
Metamyelocyte	**0.68**	0.62	0.57	0.64	0.68	**0.85**	0.32	0.53	0.68	**0.72**	0.41	0.58
Monocyte	0.9	**0.96**	0.8	0.89	**0.91**	0.85	0.95	0.86	**0.91**	0.9	0.87	0.87
Myelocyte	0.33	0.5	**1.00**	0.25	**0.19**	**0.19**	0.05	0.05	0.24	**0.28**	0.09	0.08
Platelet Cluster	**1.00**	0.92	**1.00**	0.94	0.87	**0.9**	0.82	0.82	**0.93**	0.91	0.9	0.88
Reactive Lymphocyte	**0.63**	0.52	0.52	0.50	0.6	**0.62**	0.57	0.46	**0.61**	0.57	0.55	0.48
Segmented Neutrophil	**0.99**	**0.99**	**0.99**	**0.99**	**1.00**	**1.00**	0.99	0.99	**0.99**	**0.99**	**0.99**	**0.99**

Table 4. Class-wise performance comparison of best models in terms of Precision, Recall, F1-Score. DN-121: DenseNet-121; EN_V2: EfficientNetV2; MNAS: MNASNet_0.5; MobileViT: Mobile Vision Transformer. Best values are highlighted in bold.

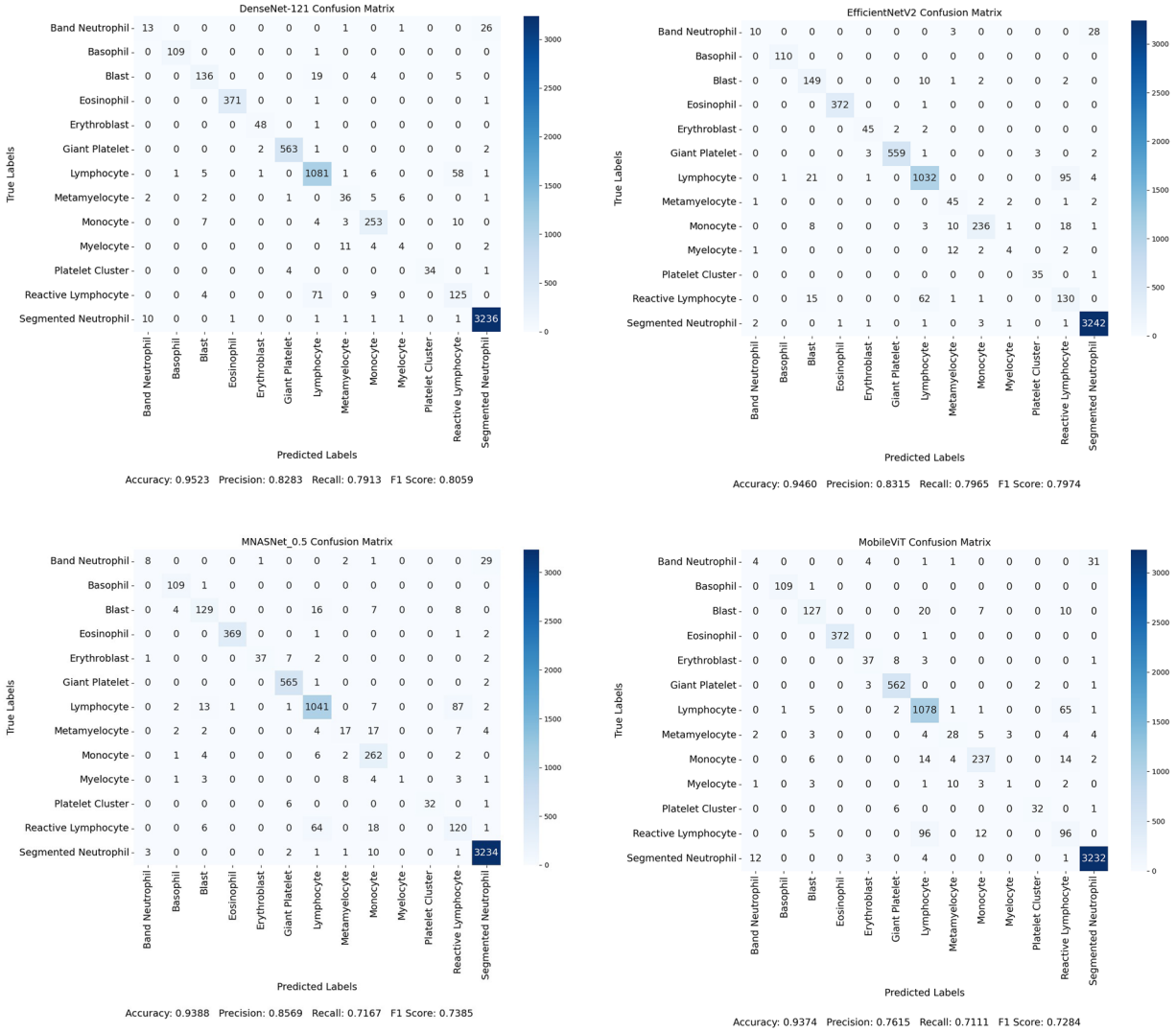

Fig. 3 Confusion matrices for four deep learning models (DenseNet-121, EfficientNetV2, MNASNet_0.5 and MobileViT) classifying 13 blood cell types.

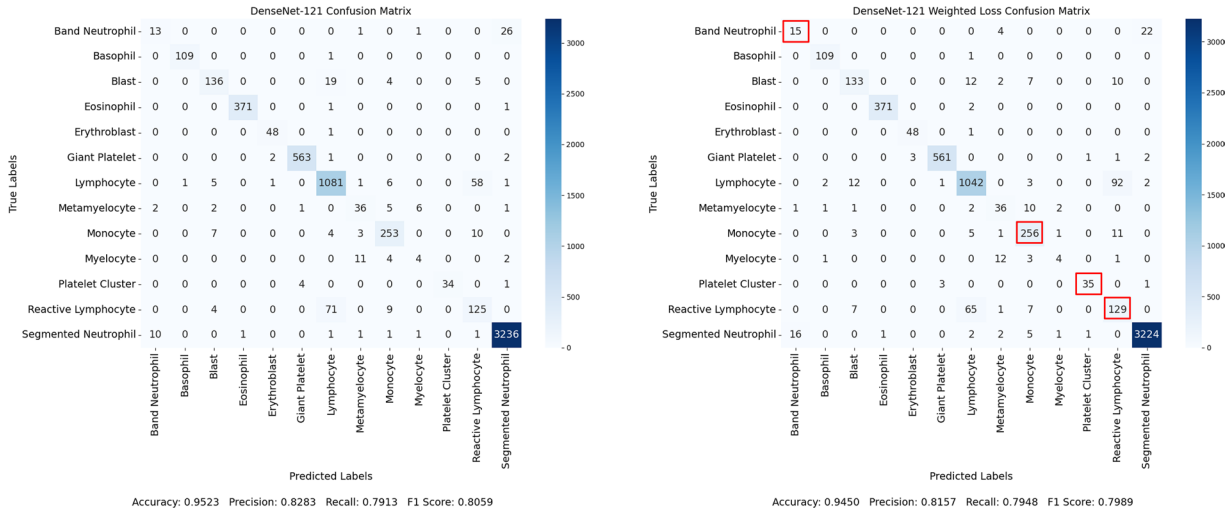

Fig. 4 Comparison of DenseNet-121 performance with (left) and without a weighted loss (right) function for addressing class imbalance.

In summary, although the design of the models affects their performance slightly, EfficientNetV2 performs slightly better on some challenging classes, such as Metamyelocyte. MobileViT, on the other hand, struggles notably with certain low-sample classes, such as Band Neutrophil and Erythroblast. The main issue is that some classes have very few examples. The models achieve better results on well-represented classes but struggle significantly with underrepresented ones. This shows that we need methods like data augmentation or special sampling to help improve accuracy for these rare cell types.

Data availability

The KU-Optofil Peripheral Blood Cell (PBC) dataset generated and analyzed in this study is publicly available on Zenodo[16]. It can be accessed using the following link https://doi.org/10.5281/zenodo.17333317. The repository contains all 31,489 anonymized peripheral blood cell images (JPG format, 368 × 368 pixels, 96 DPI) organized into training, validation and test splits, as well as accompanying metadata files (including class labels and anonymized patient identifiers) that enable both image-level and patient-level splitting strategies.

Code availability

The source code files and evaluation scripts for traditional, mobile, transformer-based, and MnasNet models are shared on GitLab https://gitlab.com/optofil/ku-optofil-peripheral-blood-cell-dataset.

Received: 7 August 2025; Accepted: 29 January 2026;
Published online: 06 February 2026

References

1. Acevedo, A. *et al.* Recognition of peripheral blood cell images using convolutional neural networks. *Computer Methods and Programs in Biomedicine* **180**, 105020, https://doi.org/10.1016/j.cmpb.2019.105020 (2019).
2. Fırat, H. Classification of microscopic peripheral blood cell images using multibranch lightweight cnn-based model. *Neural Computing and Applications* **36**, 1599–1620, https://doi.org/10.1007/s00521-023-09158-9 (2023).
3. Scotti, F. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *CIMSA. 2005 IEEE international conference on computational intelligence for measurement systems and applications, 2005.*, 96–101 (IEEE, 2005).
4. Labati, R. D. *et al.* All-idb: The acute lymphoblastic leukemia image database for image processing. In *2011 18th IEEE International Conference on Image Processing*, 2045–2048, https://doi.org/10.1109/ICIP.2011.6115881 (2011).
5. Rezatofighi, S. H. & Soltanian-Zadeh, H. Automatic recognition of five types of white blood cells in peripheral blood. *Computerized Medical Imaging and Graphics* **35**, 333–343, https://doi.org/10.1016/j.compmedimag.2011.01.003 (2011).
6. A Single-cell Morphological Dataset of Leukocytes — kaggle.com. kaggle.com/datasets/rashasalim/blood-smear-images-for-aml-diagnosis. [Accessed 31-07-2025].
7. Kouzehkanan, Z. M. *et al.* A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. *Scientific Reports***12**, https://doi.org/10.1038/s41598-021-04426-x (2022).
8. Bodzas, A. *et al.* A high-resolution large-scale dataset of pathological and normal white blood cells. *Scientific Data***10**, https://doi.org/10.1038/s41597-023-02378-7 (2023).
9. Tsutsui, S. *et al.* Wbcatt: A white blood cell dataset annotated with detailed morphological attributes. In Oh, A. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 36, 50796–50824 (Curran Associates, Inc., 2023).
10. Bodzas, A. *et al.* A high-resolution large-scale dataset of pathological and normal white blood cells. *Scientific Data* **10**, 466 (2023).
11. Rehman, A. *et al.* 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*, 553–563 (Springer, 2024).
12. Acevedo, A. *et al.* A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief* **30**, 105474, https://doi.org/10.1016/j.dib.2020.105474 (2020).
13. Rehman, A. *et al. A Large-Scale Multi Domain Leukemia Dataset for the White Blood Cells Detection with Morphological Attributes for Explainability*, 553-563 (Springer Nature Switzerland, 2024).

14. Alférez, S. *et al.* Automatic recognition of atypical lymphoid cells from peripheral blood by digital image analysis. *American Journal of Clinical Pathology* **143**, 168–176, https://doi.org/10.1309/ajcp78ifstogzzjn (2015).
15. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159, https://doi.org/10.2307/2529310 (1977).
16. Yarıkan, A. E. *et al.* Ku-optofil pbc: Ku-optofil peripheral blood cell dataset, https://doi.org/10.5281/zenodo.17333317 (2025).
17. Kuş, Z. & Aydin, M. Medsegbench: A comprehensive benchmark for medical image segmentation in diverse data modalities. *Scientific Data* **11**, https://doi.org/10.1038/s41597-024-04159-2 (2024).
18. He, K. *et al.* Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
19. Huang, G. *et al.* Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
20. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 10096–10106 (PMLR, 2021).
21. Sandler, M. *et al.* Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
22. Howard, A. *et al.* Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).
23. Tan, M. *et al.* MnasNet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).
24. Dosovitskiy, A. *et al.* An image is worth 16 × 16 words: Transformers for image recognition at scale, https://doi.org/10.48550/ARXIV.2010.11929 (2020).
25. Mehta, S. & Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, https://doi.org/10.48550/ARXIV.2110.02178 (2021).
26. Tu, Z. *et al.* Maxvit: Multi-axis vision transformer, https://doi.org/10.48550/ARXIV.2204.01697 (2022).
27. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision (2016).
28. Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, https://doi.org/10.5281/zenodo.4414861 (2019).
29. Wang, P. lucidrains/vit-pytorch. https://github.com/lucidrains/vit-pytorch (2025).

Acknowledgements

This project is supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) TEYDEB 1501 grant with a project number of 3231130. A. Kiraz acknowledges partial support from the Turkish Academy of Sciences (TÜBA)

Author contributions

A.E.Y. conducted the experiment(s), C.Ö., V.A., K.E.P., S.İ. conducted data collection, cleaning, pre-processing, and annotation steps, Z.K., M.A., B.K., A.K. and C.Ö. designed experimental studies and analyzed the results, K.B. provided supervision. All authors wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026