

Comprehensive Benchmarking of YOLOv11 Architectures for Scalable and Granular Peripheral Blood Cell Detection

Mohamad Abou Ali^{1,3,4}, Mariam Abdulfattah^{3,4}, Baraah Al Hussein^{3,4}, Fadi Dornaika^{*1,2}, Ali Cherry^{3,4}, Mohamad Hajj-Hassan^{3,4}, and Lara Hamawy^{3,4}

¹*University of the Basque Country*, ²*IKERBASQUE*, ³*Lebanese International University (LIU)*,
⁴*The International University of Beirut*,

mohamad.abouali01@liu.edu.lb, 22030274@students.liu.edu.lb,
21930849@students.liu.edu.lb, fadi.dornaika@ehu.eus, ali.cherry@liu.edu.lb,
mohamad.hajjhassan@liu.edu.lb, lara.hamawy@liu.edu.lb

Abstract

Manual peripheral blood smear (PBS) analysis is labor-intensive and subjective. While deep learning offers a promising alternative, a systematic evaluation of state-of-the-art models such as YOLOv11 for fine-grained PBS detection is still lacking. In this work, we make two key contributions. First, we curate a large-scale annotated dataset for blood cell detection and classification, comprising 16,891 images across 12 peripheral blood cell (PBC) classes, along with the red blood cell class, all carefully re-annotated for object detection tasks. In total, the dataset contains 298,850 annotated cells. Second, we leverage this dataset to conduct a comprehensive evaluation of five YOLOv11 variants (ranging from Nano to XLarge). These models are rigorously benchmarked under two data-splitting strategies (70:20:10 and 80:10:10) and systematically assessed using multiple performance criteria, including mean Average Precision (mAP), precision, recall, F1-score, and computational efficiency. Our experiments show that the YOLOv11-Medium variant achieves the best trade-off, reaching a mAP@0.5 of 0.934 under the 8:1:1 split. Larger models (Large and XLarge) provide only marginal accuracy gains at substantially higher computational cost. Moreover, the 8:1:1 split consistently

outperforms the 7:2:1 split across all models. These findings highlight YOLOv11—particularly the Medium variant—as a highly effective framework for automated, fine-grained PBS detection. Beyond benchmarking, our publicly released dataset (<https://github.com/Mohamad-AbouAli/OI-PBC-Dataset>) offers a valuable resource to advance research on blood cell detection and classification in hematology.

Keywords—Peripheral Blood Cells (PBCs), object detection, YOLOv11, dataset refinement, re-annotation, overfitting, performance evaluation, mean Average Precision (mAP), precision–recall curves, model scalability, class expansion, MakeSense.ai.

1 Introduction

Peripheral blood cell (PBC) analysis plays a pivotal role in hematology, offering critical insights for the diagnosis and monitoring of infections, hematological malignancies, and immune system disorders [17]. Traditionally, this process relies on manual microscopic examination of stained blood smears a method that, while clinically effective, is inherently constrained by observer variability, fatigue, and the time-intensive expertise required [15]. These limitations underscore the urgent need for automated, reliable systems that can assist clinicians in delivering faster and more consistent diagnostic outcomes.

Recent advances in artificial intelligence (AI) and computer vision, particularly deep learning, have revolutionized medical image analysis. These technologies are increasingly being

*Corresponding author

applied to hematology for tasks including cell detection, segmentation, and classification [8]. However, the development of robust automated blood cell analysis systems faces two primary obstacles: (1) a scarcity of high-quality, publicly available datasets with detailed annotations for the full spectrum of clinically relevant cell types, and (2) the need for a systematic evaluation of modern, efficient object detection architectures tailored for this specific domain. Challenges such as high visual similarity among PBC subtypes, significant class imbalance—especially for rare cell types—and the necessity for computational efficiency in clinical settings further complicate progress [32, 1]. To address severe class imbalance, we introduced a novel data augmentation technique, Naturalize [5], which was further validated in other medical imaging applications [2].

This study directly addresses these gaps and makes two core contributions:

First, building on an existing dataset [3] originally dedicated to PBC classification, we curate a meticulously annotated resource for the detection of twelve distinct peripheral blood cell types. The resulting large-scale dataset comprises 16,891 images encompassing 12 peripheral blood cell (PBC) classes as well as the red blood cell class, all re-annotated with bounding boxes for object detection and classification tasks. This curated dataset offers a valuable benchmark to support and advance future research in automated blood cell analysis.

Second, we conduct a comprehensive and systematic evaluation of the recently released YOLOv11 architecture [20] for the task of blood cell detection. While convolutional neural network (CNN)-based classifiers have demonstrated strong performance, object detection models provide greater versatility by simultaneously localizing and classifying multiple cells within a single image. Frameworks such as Faster R-CNN, YOLO, and RetinaNet [23, 29] have achieved notable success; however, the adoption of the latest YOLOv11 model—incorporating architectural advancements that improve both speed and accuracy—remains largely unexplored and insufficiently evaluated in hematology [31]. To this end, we conduct a rigorous comparative analysis across five YOLOv11 variants (Nano, Small, Medium, Large, and XLarge) using our novel dataset. Performance is benchmarked using key metrics including mean Average Precision (mAP), precision, recall, F1-score, training duration, and inference speed. Our findings offer practical insights into the trade-offs between detection accuracy and computational efficiency, providing a foundational guide for developing robust and clinically applicable automated PBC analysis systems.

The remainder of this paper is structured as follows. Section 2 surveys related work in traditional and deep learning-based blood cell analysis. Section 3 describes our comprehensive

methodology, covering dataset curation, class refinement, annotation, and the YOLOv11 model family. Our experimental results and a detailed comparative analysis of all models are presented in Section 4. In addition, Section 4 discusses the clinical implications of our findings, acknowledges the study’s limitations, and suggests directions for future research. Finally, Section 5 provides concluding remarks.

2 Previous Works

Automated analysis of Peripheral Blood Smear (PBS) images has evolved significantly, transitioning from traditional machine learning approaches to advanced deep learning and transformer-based architectures. This progression reflects the growing demand for faster, more accurate, and standardized diagnostic tools in hematology, addressing the limitations of manual microscopic examination.

2.1 Traditional Machine Learning Approaches

Early efforts in blood cell analysis relied on handcrafted features—morphometric, textural, and colorimetric descriptors—combined with classifiers such as Support Vector Machines (SVMs) and decision trees. Ushizima et al. [28] developed an SVM-based system to classify six white blood cell (WBC) types, including malignant chronic lymphocytic leukemia (CLL) cells, achieving 90% accuracy using a tree-based multiclass strategy. Dsilva et al. [11] applied the Wavelet Scattering Transform (WST) with SVMs for dengue diagnosis, reaching 98.7% accuracy. However, these methods lacked scalability due to their dependence on manual feature engineering. In the same study, YOLOv8 was introduced for object detection, outperforming WST-SVM with a mean accuracy of 99.3%, underscoring the superiority of deep learning for heterogeneous datasets.

2.2 Deep Learning for Blood Cell Detection and Classification

Deep learning has revolutionized PBC analysis by enabling models to learn hierarchical features directly from raw image data. Jiang et al. [19] enhanced YOLOv3 with attention mechanisms, improving mAP from 0.872 to 0.943 on the BCCD dataset [21]. Chen et al. [24] integrated ResNet and DenseNet with a Spatial and Channel Attention Module (SCAM), achieving 97.84% accuracy across multiple datasets. Yao et al. [30]

compared Faster R-CNN and YOLOv4, noting YOLOv4’s real-time performance at 60 FPS. Rehman et al. [27] introduced AttriDet, extending YOLOv5 with attribute prediction capabilities. Guo et al. [33] further optimized YOLOv5 using attention modules and refining its bounding box regression loss function to improve detection accuracy, while Gan et al. [14] proposed the TXL-PBC dataset with semi-automatic annotation via YOLOv8n, achieving mAP@0.5 of 97.0%.

2.3 Transformer-Based and Hybrid Architectures

Recent work has explored transformer-based models for blood cell detection, leveraging self-attention to capture global context. Nugraha et al. [12] combined YOLOv8 with Detection Transformer (DETR), significantly improving basophil and lymphocyte detection. Li et al. [22] utilized Fourier Ptychographic Microscopy (FPM) and Deep Convolutional GANs (DCGANs) to enhance dataset quality, achieving an mAP of 0.936 with DETR. Chen et al. [6] proposed MFDS-DETR, a hybrid CNN–Transformer model with multi-scale deformable attention, reaching AP50 scores up to 99.9% on dense cell datasets. Table 1 provides a comprehensive assessment of recent advancements in automated blood cell detection. In this table, we present a comparison summarizing the strengths and limitations of prior studies. This table consolidates key performance indicators—including mAP, accuracy, precision, recall, and F1-score—across various machine learning, deep learning, and transformer-based approaches applied to peripheral blood smear analysis.

All previous works in Table 1 are on the BCCD dataset, which is limited by the number of images “364 images”, the number of classes “3” (WBC, RBC and Platelet), and the image quality.

While the aforementioned studies demonstrate significant progress in automating PBC analysis, several key limitations persist. Many approaches are evaluated on limited cell taxonomies, often overlooking clinically relevant but morphologically similar sub-classes (e.g., banded vs. segmented neutrophils, or the spectrum of immature granulocytes). Furthermore, the performance of newer, state-of-the-art architectures like the YOLOv11 family remains largely unexplored for this specific domain, particularly in a comprehensive framework that evaluates the trade-off between accuracy and computational efficiency across multiple variants. Most critically, there is a lack of systematic benchmarking that can guide clinical deployment decisions based on real-world constraints. To address these gaps, this paper presents a thorough evaluation of the entire YOLOv11 architecture family (Nano, Small, Medium,

Large, and XLarge) for the detection of an expanded 12-class PBC taxonomy. Our novel contribution lies in the meticulous re-annotation of a large-scale public dataset for object detection and a rigorous comparative analysis across two data splitting strategies. This work provides much-needed practical insights into model selection, establishing the YOLOv11-Medium variant as the optimal architecture for achieving a critical balance between high diagnostic accuracy and operational efficiency in clinical settings.

3 Materials and Methods

This study introduces a comprehensive methodological framework for automated peripheral blood smear (PBC) analysis, encompassing critical stages from dataset curation to model evaluation. Our approach is designed for robustness and clinical relevance, beginning with the acquisition and significant refinement of a public microscopic dataset. To enhance diagnostic granularity, we expanded the original 8-class taxonomy into a more clinically meaningful 12-class system. Each cell instance in this refined dataset was meticulously re-annotated at the object level to facilitate precise localization and classification. To rigorously assess model generalization, we employed two distinct dataset splitting strategies (7:2:1 and 8:1:1). Subsequently, a full suite of YOLOv11 model variants—from the lightweight Nano to the high-performance XLarge—were trained and optimized using advanced augmentation techniques. Finally, model performance was benchmarked exhaustively using standard object detection metrics. The complete end-to-end pipeline is illustrated in Figure 1.

3.1 Peripheral Blood Cell (PBC) Dataset

The PBC dataset used in this study is derived from a dataset of microscopic peripheral blood cell images for the development of automatic recognition systems [3], acquired using a CellaVision DM96 automated hematology imaging analyzer. Images were captured in RGB color space, stored in JPEG format with a resolution of 360×363 pixels, and annotated by experienced clinical pathologists. The original dataset contains 17,092 images categorized into eight classes: neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes (myelocytes, metamyelocytes, and promyelocytes), erythroblasts, and platelets. Representative examples of these classes are shown in Figure 2 [3], and their distribution is detailed in Table 2. Each image in the PBC dataset contains one instance belonging to one of the eight classes. This image contains several red blood cells.

Table 1: Performance Comparison of Blood Cell Detection Methods

Paper	mAP	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Ushizima et al. [28]	N/A	94.9	N/A	N/A	N/A
Dsilva et al. [11]	mAP50-95 of 94.1	99.3 $\pm$ 1.4	98.6	97.55	98.76
Jiang et al. [19]	94.3	RBCs=97.44/WBCs=99.46/ Platelets=96.99	N/A	N/A	N/A
Chen et al. [8]	93.51	97.42	N/A	92.35	92.66
Yao et al. [30]	N/A	96.25	High	High	High
Rehman et al. [27]	44.2	N/A	N/A	N/A	28.2
Guo et al [33]	97.4	99.4	97.9	93.5	High
Gan et al. [14]	97.4	N/A	96.4	94.6	95.4
Nughaha [12]	N/A	N/A	83.24	77.8	78.65
Li et al. [22]	0.936 for DETR	N/A	N/A	N/A	N/A
Chen et al. [6]	79.7 (WBCDD)/97.2 (AP50)	N/A	N/A	N/A	N/A

Table 2: Distribution of Cell Types in the Peripheral Blood Cell (PBC) Dataset

#	Cell Type	Images	Percent (%)
1	Neutrophils	3329	19.48
2	Eosinophils	3117	18.24
3	Basophils	1218	7.13
4	Lymphocytes	1214	7.10
5	Monocytes	1420	8.31
6	Immature Granulocytes (IG)	2895	16.94
7	Erythroblasts	1551	9.07
8	Platelets (Thrombocytes)	2348	13.74
Total		17,092	100

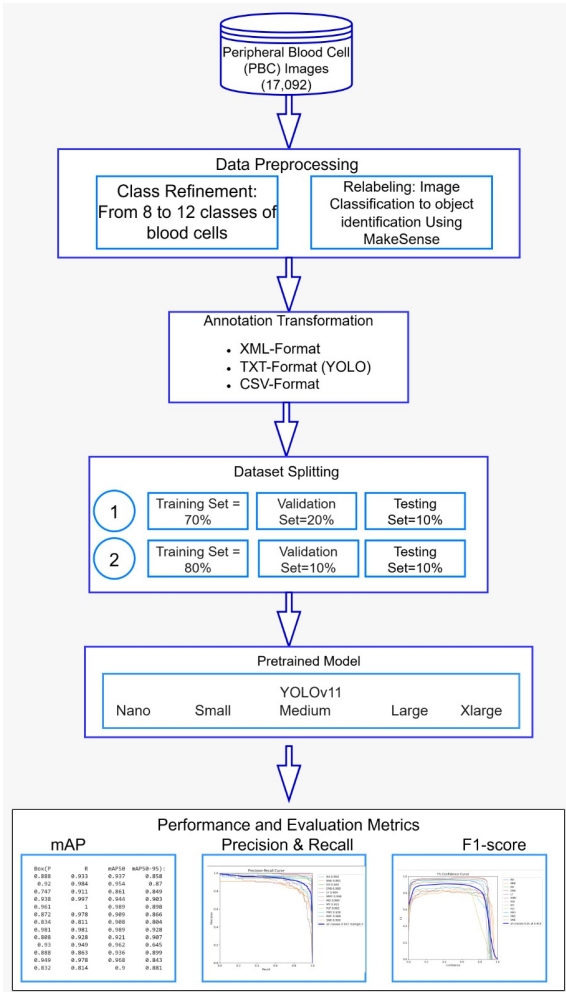


Figure 1: Comprehensive pipeline for automated PBC analysis, showing sequential stages from dataset preparation to model evaluation.

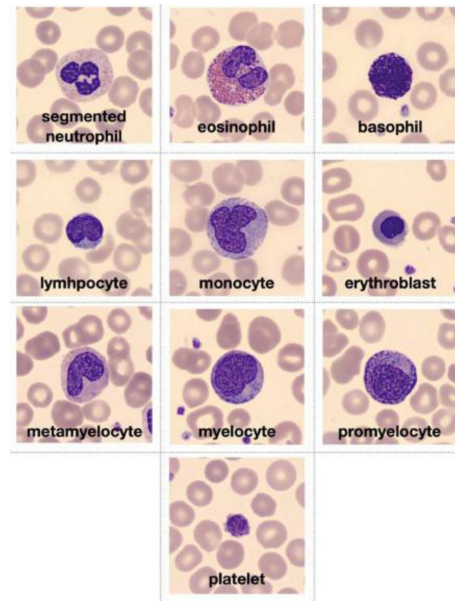


Figure 2: Representative examples of the eight original cell categories in the PBC dataset [3]. **Note:** the image meta-myelocyte, myelocyte and pro-myelocyte are found under the class Immature Granulocyte in the original PBC dataset.

3.2 Data Preprocessing

Data preprocessing is a critical phase designed to enhance the quality and utility of the raw PBC images for subsequent machine learning tasks. This stage addresses inherent challenges such as variations in image quality, inconsistencies in cell morphology, and the need for standardized labeling. The preprocessing pipeline involves two primary components: Class Refinement and Relabeling.

3.2.1 Class Refinement: From 8 to 12 Classes of Blood Cells

To enhance the clinical relevance and granularity of automated blood cell detection, the dataset was refined from an initial eight classes to a total of twelve distinct categories. This refinement involved two key stages. First, the original eight classes were expanded to eleven distinct blood cell types by subdividing neutrophils and immature granulocytes into more specific subcategories: Neutrophils were subdivided into banded neutrophils (BNE) and segmented neutrophils (SNE). Immature granulocytes were further separated into pro-myelocytes (PMY), myelocytes (MY), and meta-myelocytes (MMY). During this refinement process, 151 images from the original dataset were excluded due to insufficient morphological clarity. The refined dataset comprised 16,891 images, with the updated class distribution shown in Table 3. As mentioned earlier, each image in the PBC dataset contains one instance belonging to one of the eight classes. This image contains several red blood cells.

We introduced a twelfth class, Red Blood Cells (RBCs), as an independent category. This addition is essential for comprehensive object detection, as RBCs are the most prevalent cell type, constituting over 90% of all cells. The complete distribution of the 298,850 cell masks across all twelve classes is shown in Table 4.

The original set of 16,891 peripheral blood cell (PBC) images (detailed in Table 3) contains these 298,850 individual cell instances.

3.2.2 Relabeling for Object Detection Using MakeSense.ai

To enable precise per-cell localization and classification—a necessity for clinical diagnostic applications—the dataset was transformed from an image-level classification format to an object detection framework. This crucial transition was executed using MakeSense [25], an open-source tool for image annotation. The process involved meticulously drawing tight bounding boxes around every individual cell instance across all 12 re-

defined classes and assigning the corresponding class label. This labor-intensive manual annotation is paramount, as it generates the high-fidelity ground truth data required to train models to accurately classify and localize cells within complex microscopic fields.

The annotation workflow was standardized using MakeSense.ai to ensure consistency. To guarantee label quality and minimize human error, a rigorous cross-verification protocol was implemented, with multiple annotators reviewing a subset of images to achieve high inter-annotator agreement. The resulting annotations provide a rich dataset where each cell is precisely localized (e.g., 'lymphocyte at [x1, y1, x2, y2]'), a significant enhancement over the original global image labels (e.g., 'image contains lymphocytes'). All annotations were exported in both Pascal VOC (XML) and YOLO (TXT) formats to ensure compatibility with standard object detection architectures. A representative example of the annotated dataset, showcasing color-coded bounding boxes for various cell types, is presented in Figure 3.

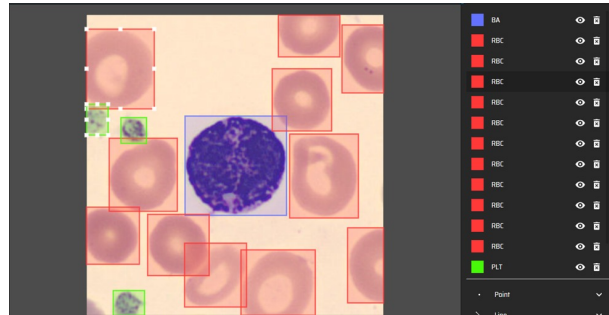


Figure 3: Example of a peripheral blood smear image annotated with bounding boxes for red blood cells (RBCs, red), white blood cells (WBCs, blue), basophils (BA, dark blue), and platelets (PLT, green) for object detection analysis.

3.3 Annotation Transformation for Interoperability

To maximize the utility, interoperability, and long-term accessibility of the curated dataset, the annotated data was programmatically converted into three standard machine learning formats. This transformation ensures compatibility with a wide range of current and future object detection frameworks and facilitates different stages of the research workflow.

The PASCAL VOC (Visual Object Classes) XML format [13] was generated to provide a structured, human-readable

Table 3: Distribution of Blood Cell Types in the Images After Class Refinement to 11 Categories

#	Cell Type	Symbol	Total Images	Percent (%)
1	Basophil	BA	1218	7
2	Banded Neutrophil	BNE	1633	10
3	Eosinophil	EO	3117	18
4	Erythroblast	ERB	1551	9
5	Lymphocyte	LY	1214	7
6	Meta-myelocyte	MMY	1015	6
7	Monocyte	MO	1420	8
8	Myelocyte	MY	1137	7
9	Platelet	PLT	2348	14
10	Pro-myelocyte	PMY	592	4
11	Segmented Neutrophil	SNE	1646	10
Total			16891	100

Table 4: Collected Masks from the PBC dataset

#	Class	Masks	Percent (%)
1	Basophil	1,192	0.40
2	Banded Neutrophil	1,694	0.57
3	Eosinophil	3,143	1.05
4	Erythroblast	1,763	0.59
5	Lymphocyte	1,246	0.42
6	Metamyelocyte	1,057	0.35
7	Monocyte	1,450	0.49
8	Myelocyte	1,151	0.39
9	Platelet	14,652	4.90
10	Promyelocyte	638	0.21
11	Segmented Neutrophil	1,754	0.59
12	Red Blood Cell	269,110	90.05
Total		298,850	100

record of annotations, ideal for archival and use with numerous deep learning libraries. For model training, the annotations were converted to the native TXT format [26] required by the YOLO (You Only Look Once) family of models, which provides a compact representation using normalized bounding box coordinates. Finally, a CSV (Comma-Separated Values) format [18] was created to enable efficient data inspection, filtering, and analysis using tabular data tools like Pandas, supporting streamlined data management and collaborative research. All conversions were executed using automated scripts to guarantee accuracy and eliminate manual error, thereby preserving the integrity of the meticulously annotated ground truth data.

3.4 Dataset Splitting Strategy

Effective dataset partitioning is critical for developing generalizable models and mitigating overfitting. To rigorously evaluate the robustness and performance of the YOLOv11 variants under different data availability scenarios, two distinct stratified splitting strategies were employed. Stratified sampling was applied in both cases to preserve the original class distribution across all subsets, ensuring a representative and unbiased evaluation.

The first strategy utilized a conventional 7:2:1 split, allocating 70% of the images for training, 20% for validation, and 10% for a hold-out test set. This configuration provides a substantial training corpus for feature learning, a dedicated validation set for hyperparameter tuning and early stopping, and a completely independent test set for the final, unbiased assessment of model performance on unseen data.

A second strategy employed an 8:1:1 split to investigate the impact of increased training data volume on model performance. This approach prioritizes maximum feature exposure during training by allocating a larger portion of the data (80%), while maintaining a sufficient validation set (10%) for monitoring generalization and a consistent, independent test set (10%) for fair comparative analysis between the two strategies. The unchanged test set size ensures that performance comparisons are equitable and not influenced by variations in test set difficulty or composition.

3.5 Model Selection, Architecture, and Training Configuration

The YOLOv11 (You Only Look Once version 11) object detection framework [16] was selected as the core architecture for this study due to its state-of-the-art performance in real-time object detection and its optimal balance between inference speed and accuracy, which is particularly suitable for medical image analysis applications requiring both precision and efficiency. To comprehensively evaluate the performance-efficiency trade-offs across different model complexities, all five primary variants of the architecture—Nano (n), Small (s), Medium (m), Large (l), and XLarge (x) [4]—were implemented and compared.

Each variant was initialized with pre-trained weights from the MS COCO (Common Objects in Context) dataset, leveraging transfer learning to accelerate convergence and enhance feature extraction capabilities on the specialized peripheral blood cell (PBC) dataset. The architectural foundation common to all variants consists of three main components, as illustrated in Figure 4: a CSPDarknet backbone for hierarchical feature extraction, a feature pyramid network (FPN) and path aggregation network (PAN) neck for multi-scale feature fusion, and an anchor-free detection head that generates bounding box coordinates, objectness scores, and class probability predictions.

The model variants represent a progressive scaling of this fundamental architecture, with increasing depth, width, and feature capacity from Nano to XLarge. The YOLOv11-Nano variant, designed for edge deployment, prioritizes computational efficiency with minimal parameters, potentially at the cost of discriminating morphologically similar cell types. The Small variant offers a balanced performance-efficiency profile suitable for most clinical applications. The Medium variant provides enhanced representational capacity for capturing nuanced cellular features, while the Large and XLarge variants target maximum accuracy in resource-abundant environments, aiming to address the most challenging cases of cellular differ-

entiation.

All models were trained using a consistent, optimized set of hyperparameters to ensure fair comparative analysis. The training regimen employed stochastic gradient descent with momentum (0.937) and weight decay (0.0005) regularization, with an initial learning rate of 0.01 following a cosine annealing schedule. Extensive data augmentation strategies were implemented, including Mosaic augmentation (probability=1.0), random horizontal flipping (probability=0.5), color space adjustments in HSV domains (hue: ± 0.015 , saturation: ± 0.7 , value: ± 0.4), and RandAugment automatic augmentation. Models were trained for 100 epochs with a batch size of 16 at 640×640 resolution, using automatic mixed precision to accelerate training while maintaining numerical stability. Early stopping with a patience of 15 epochs was implemented to prevent overfitting, with comprehensive validation performed after each epoch.

A detailed summary of the primary hyperparameters is provided in Table 5.

3.6 Detection-Oriented Evaluation Metrics

The performance of all YOLOv11 variants was rigorously evaluated using standard object detection metrics to facilitate a comprehensive comparison across model sizes and data splitting strategies. This multi-faceted assessment ensures a balanced interpretation of each model’s capabilities beyond a single performance indicator. The primary metric was Mean Average Precision (mAP) [7], which provides a holistic measure of both precision and recall across various Intersection over Union (IoU) thresholds. We report two key mAP variants:

- mAP@0.5: The average precision at an IoU threshold of 0.5, serving as the primary indicator of overall detection accuracy under a common localization criterion.
- mAP@0.5:0.95: The mean average precision computed over IoU thresholds from 0.5 to 0.95 (in increments of 0.05), providing a stringent assessment of bounding box localization accuracy.

To deconstruct the constituents of mAP and provide interpretability for clinical deployment, we also report precision, recall, and F1-score [9]. Precision, the ratio of true positives to all positive detections, quantifies the model’s reliability and its propensity for false alarms. Recall, the ratio of true positives to all actual positives in the dataset, measures the model’s sensitivity and ability to identify all relevant cell instances. The F1-score, as the harmonic mean of precision and recall, offers a single balanced metric that is particularly informative for evaluating performance on rare cell classes amidst the inherent dataset imbalance.

This suite of metrics collectively enables a nuanced analysis of the trade-offs between detection accuracy, localization pre-

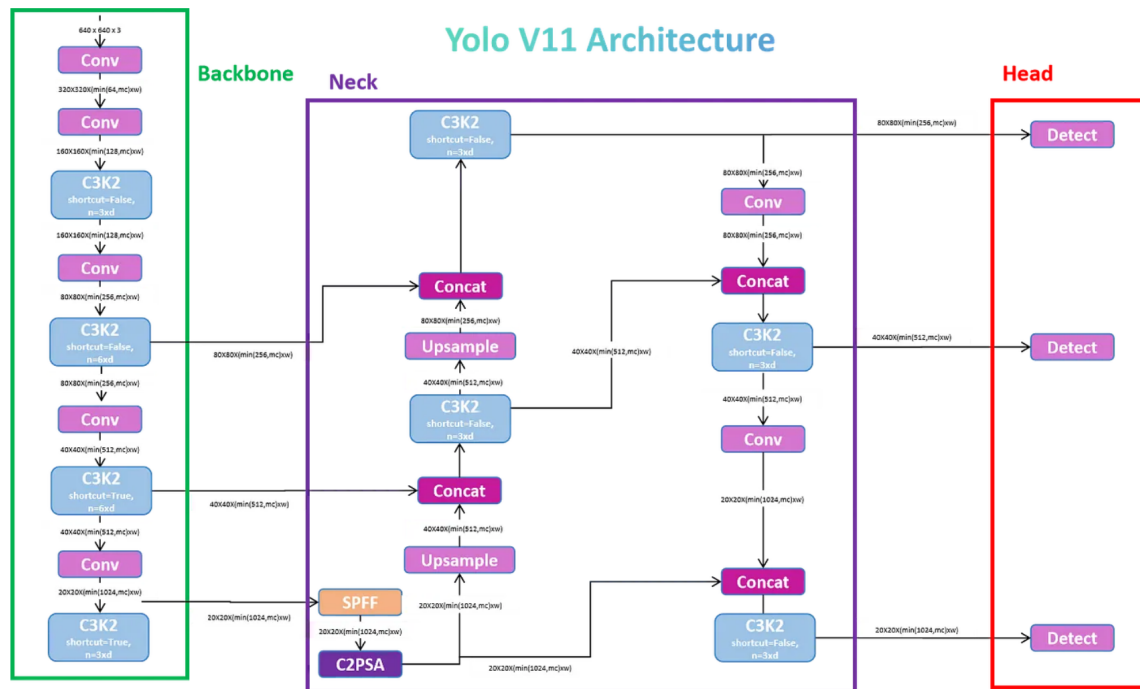


Figure 4: YOLOv11 architecture, showing convolutional backbone, spatial pyramid pooling, and anchor-free detection heads

Table 5: Hyperparameter Configuration for YOLOv11 Training

Category	Parameter	Value	Function
Training Strategy	epochs	100	Maximum training iterations
	batch	16	Batch size for gradient computation
	imgsz	640	Input image resolution
	patience	15	Early stopping tolerance
Optimization	optimizer	SGD	Stochastic Gradient Descent with momentum
	lr0	0.01	Initial learning rate
	lrf	0.0001	Final learning rate (cosine annealing)
	momentum	0.937	SGD momentum parameter
	weight_decay	0.0005	L2 regularization coefficient
Data Augmentation	mosaic	1.0	Mosaic augmentation probability
	fliplr	0.5	Horizontal flip probability
	hsv_h	0.015	Hue augmentation range
	hsv_s	0.7	Saturation augmentation range
	hsv_v	0.4	Value augmentation range
	auto_augment	randaugment	Automatic augmentation policy
Hardware	amp	true	Automatic Mixed Precision enabled
	workers	8	Data loading threads

cision, and classification reliability. The insights derived are critical for guiding model selection based on practical clinical requirements, where the balance between high recall (minimizing missed cells) and high precision (minimizing misdiagnoses) is paramount. A classification report of confusion matrix [10] evaluates prediction quality using precision, recall, and F1-score per class, along with macro and weighted average accuracies. Accuracy [10], calculated as a percentage of correct predictions, is determined by Equation 1:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision [10] measures the quality of a positive prediction made by the model and the Equation 2 demonstrates its computational process:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall [10] measures how many of the true positives (TPs) were recalled (found) and calculated using the Equation 3:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

F1-Score [10] is the harmonic mean of precision and recall and can be calculated using the Equation 4:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

4 Results & Discussion

This section presents a comprehensive evaluation of YOLOv11 model variants for automated peripheral blood cell detection and classification, analyzing performance metrics across architectures and data splitting strategies to identify optimal configurations for clinical hematology applications.

4.1 YOLOv11-Nano Model Performance

As the most lightweight architecture, the YOLOv11-Nano model established a performance baseline. On the 7:2:1split, the model achieved a final mAP@0.5 of 0.878. A clear dichotomy in performance was observed. Morphologically distinct cell types, such as Eosinophils (EO), Monocytes (MO), and Basophils (BA), were detected with high reliability, as evidenced by F1-scores consistently above 0.7, illustrated in the F1-confidence curve Figure 5.

Conversely, the model exhibited notable difficulty in fine-grained differentiation, particularly among morphologically similar cell types within the granulocytic lineage. Immature granulocytes—promyelocytes (PMY), myeloblasts (MY), and metamyelocytes (MMY)—as well as neutrophil subclasses—banded (BNE) and segmented (SNE)—were frequently misclassified due to their overlapping visual features. This challenge is clearly reflected in the normalized confusion

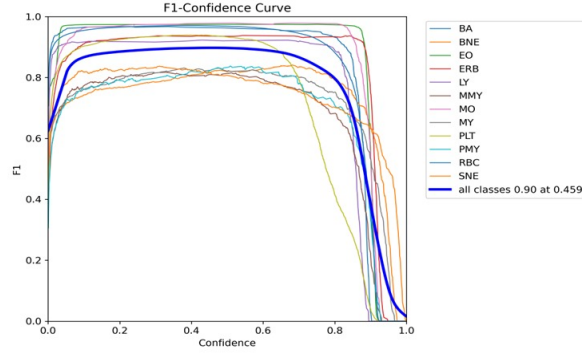


Figure 5: F1-Confidence Curve for YOLO Nano model across all 12 PBC classes (7:2:1 Split).

matrix Figure 6, which highlights substantial misclassification rates, including PMY misidentified as MY (15.7%) and MMY as MY (18.3%). Additionally, platelets (PLT) demonstrated high recall at lower confidence thresholds, yet precision declined sharply at higher thresholds, suggesting persistent confusion with small cellular artifacts. Although the dataset is explicitly defined with twelve biologically relevant blood cell categories, the normalized confusion matrix includes an additional category—background—as part of the evaluation process. This arises from the nature of object detection frameworks such as YOLO, which are trained not only to distinguish among annotated object classes but also to differentiate objects from non-object regions within an image. Any region that does not correspond to a labeled blood cell is implicitly treated as background, allowing the model to suppress false positives and enhance localization accuracy. As a result, the background class is incorporated during evaluation to reflect the model’s ability to correctly identify true cellular instances while avoiding misclassification of irrelevant image regions. This inclusion provides a more comprehensive and diagnostically meaningful assessment of detection performance.

Utilizing the larger training proportion of the 8:1:1 split provided a noticeable performance boost for the Nano model, with its mAP@0.5 rising to 0.892. Improvements were most pronounced in the F1-confidence curve, where the overall macro F1-score peaked at 0.90 (figure 8). Classes such as Erythroblasts (ERB) and Red Blood Cells (RBC) showed more stable performance across confidence thresholds. However, the model’s fundamental limitations in distinguishing morphologically similar cell types persisted, as confirmed by the persistent patterns of misclassification visible in the corresponding confusion matrix.

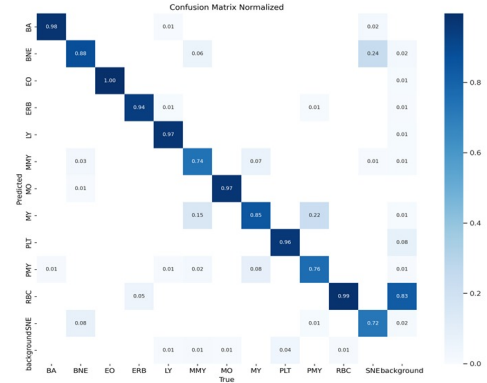


Figure 6: Normalized Confusion Matrix for YOLO Nano model (7:2:1 Split) .

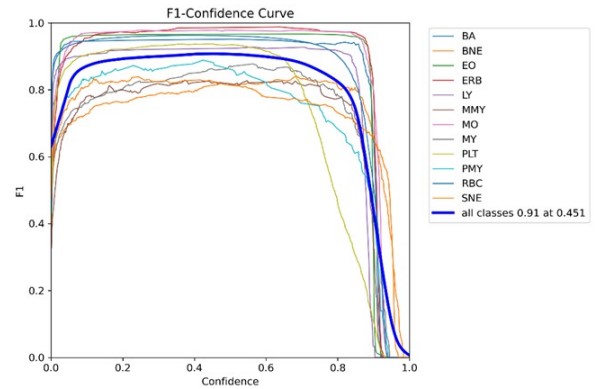


Figure 7: F1-score curve for YOLO Nano model showing performance (8:1:1 Split).

4.2 YOLOv11-Small Model Performance

The YOLOv11-Small model, offering increased capacity, consistently outperformed the Nano variant, achieving a superior balance between computational efficiency and detection accuracy. On the 7:2:1 split, the Small model attained a higher mAP@0.5 of 0.904. The F1-confidence curves Figure 8 showed modest but consistent improvements across all classes, with the most notable gains observed for the challenging immature granulocytes and neutrophils.

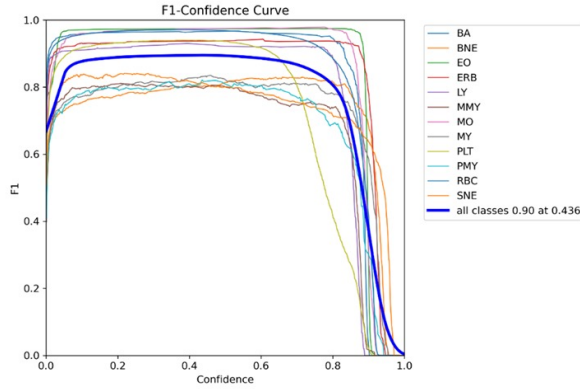


Figure 8: F1-score curve for YOLO Nano model showing performance (7:2:1 Split).

This enhanced discriminatory power is further demonstrated by the normalized confusion matrix, which showed reduced misclassification rates between BNE and SNE, as well as less confusion of PLT with artifacts. When trained on the 8:1:1 split, the Small model's performance peaked at a mAP@0.5 of 0.915. It achieved a superior precision-recall balance compared to the Nano model, reflected in an overall mAP@0.5 of 0.915. The Precision-Recall curve Figure 9 indicates strong performance across most classes, with exceptionally high Average Precision (AP) values for MO (0.989) and ERB (0.989).

The training process was also more efficient, with convergence occurring at 80 epochs compared to the 92 required for the Nano model, highlighting its optimized learning dynamics, as shown in the training results plot Figure 10.

4.3 YOLOv11-Medium Model Performance

The YOLOv11-Medium model represented a significant turning point in the performance hierarchy, delivering substantial gains that often surpassed the incremental improvements

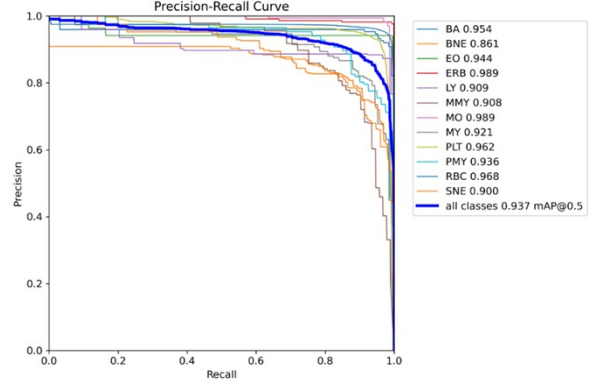


Figure 9: Precision-Recall curve for YOLO Small model showing the trade-off between precision and recall across different confidence thresholds (8:1:1 Split).

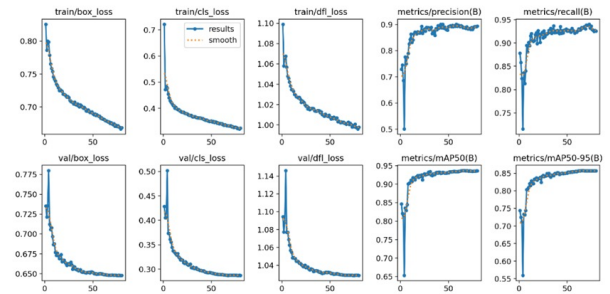


Figure 10: Training Results for YOLO Small model (8:1:1 Split).

seen in subsequent larger variants. On the 7:2:1 split, the Medium model delivered a dramatic improvement, achieving an mAP@0.5 of 0.943. This enhancement was most evident in the classification of the most challenging cells. The normalized confusion matrix Figure 11 reveals a stark reduction in misclassification among immature granulocytes and neutrophils; for instance, the PMY→MY error rate was nearly halved compared to the Nano model. The F1-scores for these classes saw significant increases, indicating the model’s augmented capacity is critical for capturing subtle morphological differences.

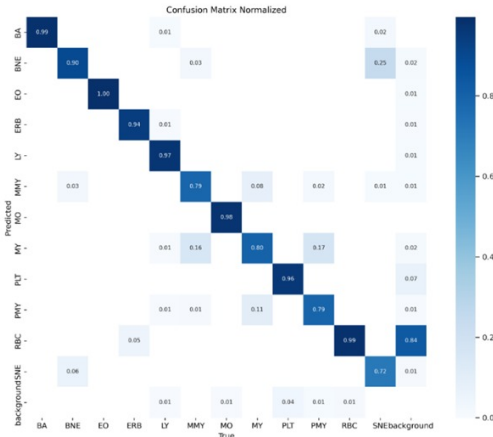


Figure 11: Normalized Confusion Matrix for YOLO Medium model (7:2:1 Split).

The model maintained its strong performance on the 8:1:1 split, achieving an mAP@0.5 of 0.934. The F1-confidence curve (Figure 12) showed further gains for immature granulocytes (e.g., PMY peak 0.83 vs 0.80 on Small). The model excelled particularly in recall (0.935), making it highly suitable for screening applications where minimizing false negatives is critical.

Analysis of the confusion matrix confirmed these findings, showing improved differentiation for all cell types, with PLT misclassification with artifacts dropping to 15.8%.

4.4 YOLOv11-Large and XLarge Models Performance

The YOLOv11-Large and XLarge models demonstrated the law of diminishing returns, whereby performance gains became marginal and were not commensurate with the substantial increase in computational cost and complexity. On the 7:2:1 split, the Large model’s mAP@0.5 of 0.945 and the XLarge’s

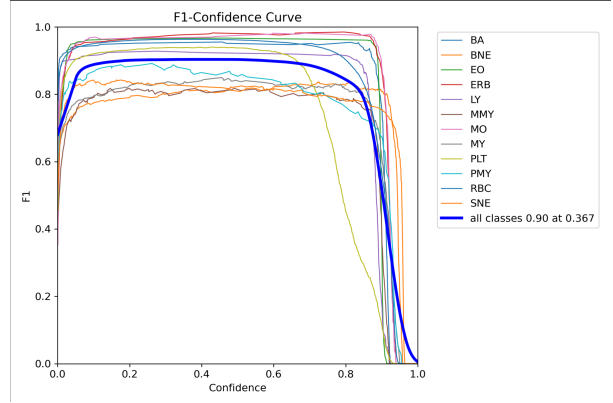


Figure 12: F1 confidence medium model (8:1:1 Split).

value of 0.943 were only marginally higher than the Medium model’s 0.943. Analysis of the precision-recall curves for the Large model Figure 13 and the F1-confidence curves for the XLarge model Figure 14 revealed nearly identical performance between the Large, XLarge, and Medium models. The confusion matrices for these large variants were also virtually indistinguishable from the Medium model’s, confirming that the additional millions of parameters did not translate to meaningfully better feature extraction or classification for this specific task.

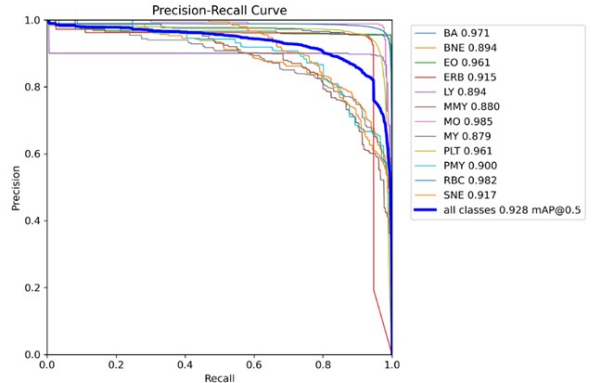


Figure 13: Precision-Recall Curve for YOLO Large model (7:2:1 Split).

This trend continued on the 8:1:1 split. The Large model achieved an mAP@0.5 of 0.947, while the XLarge model reached 0.948—a negligible 0.1% improvement incurred at

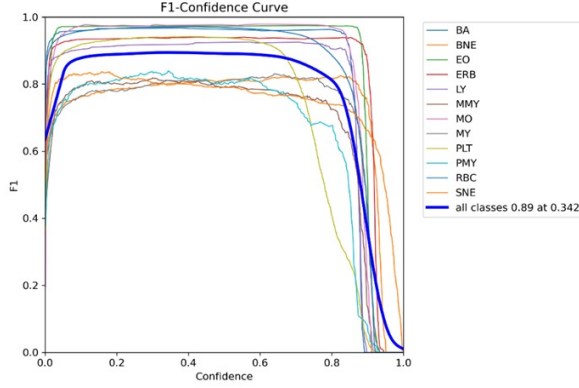


Figure 14: F1-Confidence Curve for YOLO XLarge model across all 12 PBC classes (7:2:1 Split).

a great computational expense. The training results for the XLarge model Figure 15 showed convergence dynamics and final metric values that were highly similar to those of the Large and Medium models, reinforcing the conclusion that model performance had saturated at the Medium architecture level.

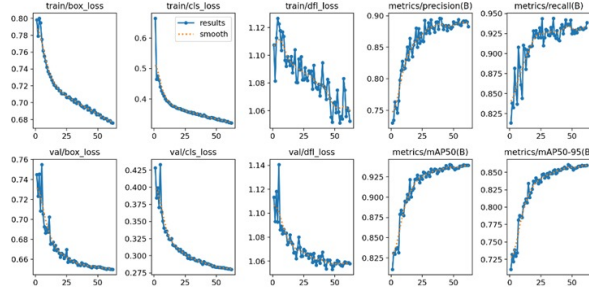


Figure 15: Training Results for YOLO XLarge model (8:1:1 Split).

4.5 Comparative Analysis Across Architectures and Splits

A holistic comparison of all models and data splits, summarized in Table 6, yields several critical insights. The 8:1:1 data split (80% training data) consistently outperformed the 7:2:1 split across all models and evaluation metrics. Smaller models (Nano, Small) benefited more substantially from the larger training dataset, while larger models (Medium and above)

showed smaller gains, as their inherent capacity was already substantial. Furthermore, performance gains plateaued at the Medium model variant. The minimal increase in mAP from Medium to XLarge was accompanied by a drastic increase in computational cost, both in training time and inference speed. Consequently, the YOLOv11-Medium model trained on the 8:1:1 data split emerged as the optimal choice, achieving 98.6% of the Large model’s performance while requiring significantly fewer computational resources, thereby offering the best practical balance for clinical deployment.

It should be noted that the YOLOv11-XLarge model was trained and evaluated on an NVIDIA A100 GPU, while the Nano, Small, Medium, and Large variants were executed on an NVIDIA T4 GPU. This hardware discrepancy accounts for the unexpected observation where the XLarge model exhibits comparatively shorter training and inference times despite its larger parameter count and higher computational complexity. Consequently, direct runtime comparisons across all variants should be interpreted with caution, as the results are influenced not only by model architecture but also by differences in computational hardware.

4.6 Qualitative Results: Ground-Truth versus Prediction

This section provides a qualitative analysis of the model performances by directly comparing their predictions against the ground-truth annotations for the test set. This visual assessment complements quantitative metrics by highlighting specific strengths, such as accurate classification and localization, and weaknesses, such as misclassifications between morphologically similar cell types.

Figure 16 demonstrates the robust performance of the YOLOv11 Small model. The visual comparison reveals precise alignment between the predicted bounding boxes and the ground-truth annotations for Segmented Neutrophils (SNE), Red Blood Cells (RBC), and Platelets (PLT). The high confidence scores associated with these correct predictions indicate the model’s certainty in detecting these common cell types.

This high standard of performance is also evident in the results from the smaller YOLOv11 Nano model. As shown in Figure 17, the Nano variant successfully identifies and localizes the less frequent Basophils (BA) in addition to RBCs and Platelets. The consistency between its predictions and the ground-truth labels confirms that this compact model also generalizes effectively on the PBC dataset for these specific classes.

However, not all predictions are flawless, and qualitative analysis is key to identifying specific failure modes. Fig-

Table 6: Performance comparison of YOLOv11 variants on PBC dataset (splits: 811 and 721).

Metric	Split	Nano	Small	Medium	Large	XLarge
PARAMS (M)	-	2.6	9.4	20.1	25.3	86.9
mAP50-95	8:1:1	0.892	0.915	0.934	0.947	0.948
	7:2:1	0.845	0.853	0.851	0.850	0.851
Precision	8:1:1	0.887	0.908	0.926	0.941	0.943
	7:2:1	0.892	0.887	0.878	0.882	0.877
Recall	8:1:1	0.895	0.921	0.939	0.952	0.953
	721	0.910	0.909	0.919	0.916	0.921
F1-Score (%)	8:1:1	0.891	0.914	0.932	0.946	0.948
	7:2:1	89.75	89.24	89.82	89.98	89.84
Training Time (hrs)	8:1:1	3.2	4.5	6.8	6.0	4.0
	7:2:1	8.039	8.794	14.066	17.393	4.707
Inference Speed (ms/img)	8:1:1	12.3	18.7	27.4	42.8	50.5
	7:2:1	2.2	4.8	10.7	14.6	3.7



Figure 16: Exemplary detection results from the YOLOv11 Small model (8:1:1 Split). The top row shows the model's predictions with bounding boxes and confidence scores for Segmented Neutrophils (SNE), Red Blood Cells (RBC), and Platelets (PLT). The bottom row provides the corresponding ground-truth annotations.

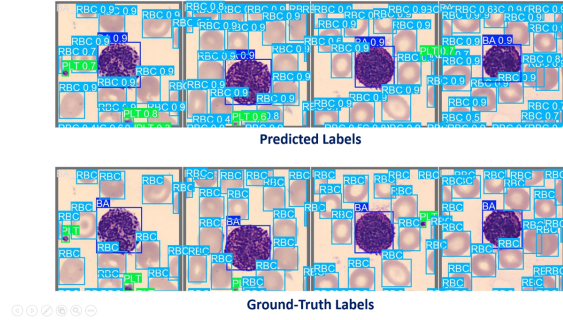


Figure 17: Detection results for the YOLOv11 Nano model (8:1:1 Split) on images containing Basophils (BA). The top row shows the model's output, which correctly identifies BA, RBC, and PLT. The bottom row confirms the accuracy of these predictions.

ure 18 provides a critical example where the YOLOv11 Nano model, while correctly identifying most cells, misclassifies a Segmented Neutrophil (SNE) as a Metamyelocyte (MMY) with a high confidence score of 0.9. This error is highly informative, as it pinpoints a confusion between two developmentally related white blood cells that can appear morphologically similar—a common and challenging nuance in cytological image analysis. This insight is valuable for guiding future work, suggesting a need for targeted data augmentation or a more refined network architecture to improve discriminative ability between such similar classes.

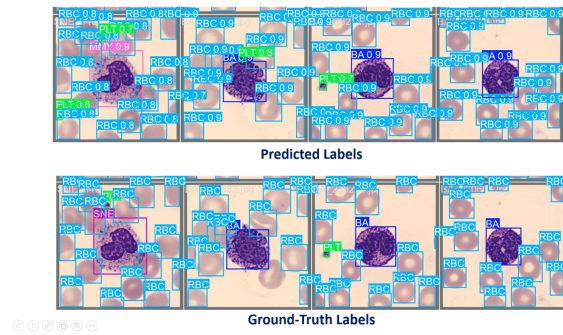


Figure 18: A case study of a misclassification by the YOLOv11 Nano model (8:1:1 Split). The model correctly identifies RBCs and PLTs but misclassifies a Segmented Neutrophil (SNE) as a Metamyelocyte (MMY) with high confidence (0.9).

In summary, the qualitative results demonstrate that both YOLOv11 variants perform robustly in detecting and classifying the majority of blood cells, showing high precision in localizing objects and assigning correct labels to distinct classes. The primary area for improvement, as identified by the specific misclassification, lies in enhancing the model’s ability to discriminate between specific, similar types of white blood cells to reduce confident false positives.

4.7 Discussion

This study presents a comprehensive benchmark of the YOLOv11 architecture family for automated peripheral blood cell detection, demonstrating its strong potential to augment hematological diagnostics. Our central finding identifies YOLOv11-Medium as the Pareto-optimal model, achieving an exceptional balance between high accuracy (mAP@0.5 of 0.934) and computational efficiency, making it the most suitable variant for real-world clinical deployment. The high per-

formance across all variants, particularly the elevated mAP metrics, underscores the capability of deep learning-based object detection to mitigate the labor-intensive and subjective limitations inherent in manual peripheral blood smear examination.

A critical insight from this work is the significant influence of dataset partitioning strategy on model performance. The consistent superiority of the 8:1:1 split over the conventional 7:2:1 split across all models indicates that for complex, imbalanced cytological datasets, maximizing the volume of training data is a more decisive factor for robustness than maintaining a larger validation set for hyperparameter tuning. This finding provides a valuable practical guideline for developing medical imaging AI systems where annotated data is scarce but precious.

Despite these promising results, several limitations must be acknowledged to contextualize our findings. Firstly, the models were trained and validated on images acquired from a single system (CellaVision DM96) under a specific staining protocol, which may impair generalizability to images from other hematology analyzers. Secondly, the absence of external validation on a multi-center dataset limits our ability to assess robustness across diverse patient populations and laboratory conditions. Finally, persistent challenges, such as severe class imbalance and the inherent morphological similarity between certain cell lineages (e.g., immature granulocytes), remain a source of residual misclassification, suggesting that architectural advances alone may be insufficient to fully resolve these ambiguities.

Future research should prioritize multi-institutional collaboration to build large, diverse, and multi-source datasets for external validation, which is the gold standard for demonstrating clinical utility. To address the lingering challenge of inter-class heterogeneity, investigations into advanced domain adaptation techniques, vision transformer-based architectures, and explainable AI (XAI) methods are warranted. These approaches could enhance model generalization and provide clinicians with interpretable diagnostic insights. Ultimately, the translation of these systems into diagnostic workflows requires rigorous clinical usability studies, incorporating a “pathologist-in-the-loop” framework to evaluate the practical synergy between AI-assisted screening and expert morphological interpretation.

In conclusion, this work establishes a robust, evidence-based framework for selecting object detection models in medical imaging, challenging the notion that larger models are invariably superior. By providing a detailed analysis of the accuracy-efficiency trade-offs across the YOLOv11 spectrum, we offer a clear roadmap for integrating scalable AI solutions into clinical hematology. The adoption of such systems holds the poten-

tial to standardize peripheral blood smear analysis, enhance diagnostic reproducibility, and streamline laboratory workflows globally, ultimately paving the way for more accessible and consistent hematological diagnostics.

5 Conclusion

This study conducted a rigorous evaluation of the YOLOv11 architecture for automated peripheral blood cell detection, systematically comparing five model variants under two data partitioning strategies. Our findings demonstrate that while model capacity is crucial for capturing complex cellular morphology, diminishing returns occur beyond the Medium variant, where larger models yield negligible performance gains despite significantly increased computational demands. Also, An annotated dataset combining detection and classification labels for cell type has been created and released on GitHub to aid future studies.

The YOLOv11-Nano variant provided competent baseline performance but showed limitations in distinguishing morphologically similar cell types. The Small and Medium variants delivered substantial improvements, with the Medium variant emerging as the optimal architecture, achieving near-peak detection accuracy (mAP@0.5 of 0.934) while maintaining practical computational efficiency. Our analysis further revealed that increased training data (8:1:1 split) consistently enhanced performance, with smaller models benefiting disproportionately from additional data.

Based on our comprehensive evaluation, we recommend the YOLOv11-Medium model trained with an 8:1:1 data split as the optimal configuration for clinical deployment. This selection represents the Pareto-optimal balance between accuracy and efficiency, making it suitable for integration into real-time diagnostic workflows.

This work challenges the conventional "larger is better" paradigm in medical AI and provides an evidence-based framework for model selection in hematological image analysis. Our findings represent a significant step toward routine clinical adoption of automated PBC analysis, with potential to enhance diagnostic precision, standardize morphological interpretation, and streamline laboratory workflows globally.

Future research will focus on external validation through multi-center collaborations and development of advanced techniques to address persistent challenges in fine-grained cell differentiation, particularly for rare cell types and immature morphological variants. In addition, different object-identification architectures will be evaluated using our generated dataset.

References

- [1] M. Abou Ali, F. Dornaika, and I. Arganda-Carreras. Blood cell revolution: Unveiling 11 distinct types with 'naturalize' augmentation. *Algorithms*, 16(12):562, 2023.
- [2] M. Abou Ali, F. Dornaika, I. Arganda-Carreras, H. Ali, and M. Karaoui. Naturalize revolution: Unprecedented AI-driven precision in skin cancer classification using deep learning. *BioMedInformatics*, 4(1):638–660, 2024.
- [3] A. Acevedo, A. Merino, S. Alf  rez, A. Molina, L. Bold  , and J. Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020.
- [4] M. S. Al-Quraishi and M. Al-Mekhlafi. Comprehensive analysis on ultralytics-supported YOLO models for detection and recognition of large office objects for indoor navigation. *Procedia Computer Science*, 232:1400–1409, 2024.
- [5] M. A. Ali, F. Dornaika, A. Cherry, M. A. Ali, and A. Hage-Diab. Advancing medical image processing: A comparative analysis of 'naturalize' and traditional augmentation techniques for enhanced diagnostic accuracy. In *2024 International Conference on Smart Systems and Power Management (IC2SPM)*, pages 138–143, 2024.
- [6] C. Chen, B. Zhang, Y. Chen, Y. Huang, C. Sun, X. Wang, Y. Fu, F. Dai, Y. Qin, Y. Peng, and Y. Gao. Accurate leukocyte detection based on Deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases, 2024. HDU-ITMO Joint Institute, Hangzhou Dianzi University, Hangzhou, China.
- [7] X. Chen, X. Wang, and K. Zhang. End-to-end training of object class detectors for mean average precision. In *Computer Vision - ECCV 2016*, pages 509–525, Cham, 2021. Springer.
- [8] X. Chen, X. Wang, and K. Zhang. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 71, 2021.
- [9] D. Chicco and G. Jurman. Improving the quality evaluation process of machine learning models: A comparative analysis of balanced accuracy, MCC, and F1 score. *Applied Soft Computing*, 149, 2024.
- [10] Hercules Dalianis. *Clinical Text Mining*. Springer, Cham, 2018.
- [11] L. Dsilva et al. Wavelet scattering- and object detection-based computer vision for identifying dengue from peripheral blood microscopy. *Int. J. Imaging Syst. Technol*, 34, 2023.

- [12] B. Erfianto and A. Nugraha. White blood cell detection using YOLOv8 integration with DETR to improve accuracy. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 7, 2023.
- [13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.
- [14] L. Gan and X. Li. TXL-PBC: A freely accessible labeled peripheral blood cell dataset, 2021. Data set publication.
- [15] R. B. Hegde, K. Prasad, H. Hebbar, B. M. K. Singh, and I. Sandhya. Automated decision support system for detection of leukemia from peripheral blood smear images. *Journal of Digital Imaging*, 33(2):361–374, 2019.
- [16] P. Hidayatullah, N. Syakrani, M. R. Sholahuddin, T. Gelar, and R. Tubagus. YOLOv8 to YOLOv11: A comprehensive architecture in-depth comparative review. *arXiv [cs.CV]*, 2025.
- [17] A. V. Hoffbrand, D. Higgs, D. M. Keeling, and A. Mehta. *Hoffbrand's Essential Haematology*. Wiley-Blackwell, Hoboken, New Jersey, USA, 2023.
- [18] R. Ingh and A. Singh. Tensorflow with user friendly graphical framework for object detection api, 2020.
- [19] Z. Jiang. Improved detection performance in blood cell count by an attention-guided deep learning method. *OSA Continuum*, 4(2), 2021.
- [20] R. Khanam and M. Hussain. YOLOv11: An overview of the key architectural enhancements, 2024.
- [21] Shin-Jye Lee, Pei-Yun Chen, and Jeng-Wei Lin. Complete blood cell detection and counting based on deep neural networks. *Applied Sciences*, 12(16), 2022.
- [22] M. Li et al. Peripheral blood leukocyte detection based on an improved detection transformer algorithm. *Sensors*, 23(16), 2023.
- [23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.
- [24] J. Liu. Accurate classification of white blood cells by coupling pre-trained ResNet and DenseNet with SCAM mechanism. *BMC Bioinformatics*, 2022.
- [25] A. Musleh, S. A. AlRyalat, and A. Qasem. Image annotation software for artificial intelligence applications. *High Yield Medical Reviews*, 1(2), 2023.
- [26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [27] A. Rehman, T. Meraj, A. M. Minhas, A. Imran, M. Ali, and W. Sultani. A large-scale multi domain leukemia dataset for the white blood cells detection with morphological attributes for explainability, 2023. Unpublished document or pre-print, not fully formatted.
- [28] D. Ushizima and A. A. D. Lorena. Support vector machines applied to white blood cell recognition. In *Proc. IEEE Int. Conf. Hybrid Intelligent Systems (ICHIS)*, pages 195–200, 2005.
- [29] Qiang Wei, Xiaoyang Hu, Xuna Wang, and Haodong Wang. Improved retinanet target detection model. In *2022 2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence (AH-PCAI)*, pages 470–476, 2022.
- [30] J. Yao and X. Huang. High-efficiency classification of white blood cells based on object detection. *Journal of Healthcare Engineering*, 2021, 2021.
- [31] M. Yurdakul and S. Tasdemir. BC-YOLO: MBCConv-ECA based YOLO framework for blood cell detection. *Signal, Image and Video Processing*, 19, 2025.
- [32] L. Zhang, X. Wang, and Y. Li. Challenges and solutions in automated blood cell analysis: A comprehensive review. *Journal of Medical Imaging and Health Informatics*, 11(4):1025–1038, 2021.
- [33] Y. Zhang and M. Guo. Blood cell detection method based on improved YOLOv5. *IEEE Access*, 3:67987–67995, 2023.