White Blood Cell Classification

Dina Zahran[1], Ali Marzban[2], and Karim ElTohamy[2]

[1]CIS Research Department, Nile University
[2]ITCS Department, Nile University

Abstract

Automated white blood cell (WBC) classification is an essential yet challenging component of digital hematopathology, especially when images are strongly affected by noise, blur, and staining variability. In this work, a robust pipeline is developed for a 13-class WBC classification challenge in which images contain multiple WBCs, red blood cells, and platelets, lack segmentation masks, and often exhibit severe synthetic degradations. Several preprocessing strategies are systematically explored, including classical denoising and deconvolution, learning-based restoration, and morphology-preserving stain normalization, together with modern classification backbones such as EfficientNetV2-B0, ConvNeXt-Tiny, and ResNet-50, as well as pretrained WBC detection with YOLOv8m to obtain WBC-focused crops. An ensemble of CNN classifiers—achieves a macro F1-score of approximately 0.63 on the competition leaderboard, highlighting both the promise and the difficulty of fine-grained WBC recognition under extreme image degradation. These results motivate future work on more expressive augmentation and fusion strategies to better exploit rare classes and improve robustness.

1 Introduction

Automated analysis of peripheral blood smears has the potential to reduce workload, standardize reporting, and extend expert-level morphology to settings where trained hematologists are scarce [1]. Deep learning methods have achieved strong performance on several public WBC datasets, but these collections typically contain well-focused, clean, and consistently stained images that do not reflect the variability and degradation encountered in practice [2]. In contrast, recent robustness-oriented benchmarks introduce heavy noise, blur, and color perturbations, making fine-grained subclassification of leukocytes substantially more challenging and exposing the fragility of standard pipelines [3].

Under such conditions, naïve preprocessing can easily destroy diagnostically relevant features, such as nuclear contours, chromatin texture, cytoplasmic granules, and subtle differences in staining, while overly powerful restoration models may hallucinate structures that do not correspond to the underlying biology [3]. At the same time, modern detectors and classifiers, including YOLO-based WBC localization and CNN/transformer backbones, offer a rich design space for building robust systems if the input morphology is preserved [4][5]. This work focuses on designing and evaluating a morphology-aware pipeline for robust WBC classification under extreme image degradation, comparing classical filters, learning-based denoisers, and stain-normalization strategies, and integrating them with state-of-the-art detection and classification models to achieve strong performance on a challenging leukocyte classification benchmark.

2 Related Works

Recent work on WBC classification has been dominated by convolutional and transfer-learning architectures trained on curated microscopic datasets. Many studies focus on five major leukocyte types and report high accuracies by combining standard preprocessing with deep CNN backbones such as DenseNet, ResNet, and EfficientNet. For example, a DenseNet121-based model with normalization and data augmentation achieved about 99% accuracy on a four-class Kaggle dataset, while other studies combine hand-crafted features such as SIFT or texture descriptors with CNN embeddings to improve performance on LISC and similar datasets. These systems demonstrate the effectiveness of deep learning when image quality is high and staining conditions are controlled.

In parallel, several papers have investigated optimized pipelines that integrate more elaborate preprocessing and model selection. A recent Scientific Reports study proposed "multi-fold preprocessing" with extensive data augmentation followed by an optimized CNN, reporting 0.99 accuracy and arguing that careful preprocessing plus a well-tuned custom network can outperform heavier transfer-learning models on standard WBC benchmarks. Other works employ hybrid strategies, such as super-resolution GANs to enhance low-resolution cells before classification with VGG-based networks, or architectures like ReRNet designed to improve convergence and generalization on limited WBC data. A comprehensive 2024 review confirms that CNN-based approaches now dominate the field and generally exceed classical machine-learning pipelines in both accuracy and robustness but also notes that most evaluations are conducted on relatively clean, low-noise images.

More recently, attention has turned to domain shift and alternative imaging modalities. Deep feature-based classifiers have been proposed for label-free WBC imaging, where a CNN encoder replaces manual feature engineering and achieves competitive performance compared to traditional cytometry. Nonetheless, even these newer studies typically assume moderate image quality and do not systematically address extreme degradations such as heavy synthetic noise, blur, and severe staining variation. Robustness-oriented benchmarks and analyses remain scarce, and there is limited evidence on how aggressive denoising, restoration, or stain normalization interact with downstream WBC classification performance. The present work specifically targets this gap by evaluating deep learning–based WBC classification under highly degraded imaging conditions and by comparing morphology-preserving versus restoration-focused preprocessing strategies.

3 Materials and Methods

3.1 Data Collection and Preprocessing

The experiments in this work are conducted on a recent white blood cell classification challenge that targets robustness under realistic acquisition variability and synthetic degradations. The dataset consists of color microscopic patches centered on WBCs and annotated into 13 morphologically fine-grained classes covering mature leukocytes, immature precursors, and atypical/reactive forms. Each RGB image typically contains at least one WBC but also includes surrounding red blood cells (RBCs), platelets, and background artifacts, reflecting realistic smear composition rather than pre-cropped single-cell thumbnails.

Several properties of the data make the task particularly challenging:

1. The class distribution is highly imbalanced. Frequent types such as segmented neutrophils and lymphocytes dominate the training set, while rare precursors and atypical cells (e.g. PLY, PC, VLY) appear in small numbers. This increases bias and creates unstable decision boundaries for under-represented classes.

2. Some images contain more than one white blood cell, yet each image has a single label. This assumes a primary or central cell defines the class. The ambiguity complicates supervised learning, especially when secondary cells belong to different categories.

3. Many images show synthetic noise, motion blur, defocus blur, and color perturbations. These distort nuclear and cytoplasmic morphology and make expert visual assessment difficult.

4. The competition does not provide segmentation masks, which makes supervised WBC/RBC separation infeasible and complicates the use of mask-based segmentation networks, motivating the exploration of pretrained detection models instead.

In our trials, classical preprocessing techniques (median and bilateral filtering, CLAHE, deconvolution) and learning-based restoration with models such as CARE and related denoisers improved global smoothness but often failed to recover diagnostically useful structures or even over-smoothed and washed out the nucleus and granules. These characteristics motivated the design of a morphology-preserving pipeline that relies on gentle preprocessing, robust WBC localization, and strong classifiers rather than attempting aggressive "perfect" restoration of heavily corrupted inputs.

3.2 Baseline Classification Experiments on Raw Images

Given these constraints, the first group of experiments evaluated how far one can go by applying modern classifiers directly to the raw competition images, without any explicit segmentation or heavy restoration. We experimented with vision transformers (ViTs) such as DinoBloom and SCKansformer as well as a YOLOv11-based classifier, alongside strong convolutional baselines. For CNN-based models, all images were converted to 3-channel RGB, resized to 224×224 using padding to preserve aspect ratio, and normalized with ImageNet statistics. Data augmentation included random rotations, zooms, shifts, crops, and moderate brightness/contrast jitter, blur, and noise;

for rare classes (PLY, PC, VLY) we applied stronger geometric and photometric perturbations and employed balanced sampling so that mini-batches approximated a uniform class distribution. CutMix augmentation was enabled with moderate probability to further enrich training examples for under-represented classes.

Two ImageNet-pretrained backbones, EfficientNetV2-B0 and ConvNeXt-Tiny, were fine-tuned with a custom 13-class classification head. Training proceeded in stages:

1. Head-only training with the backbone frozen and a learning rate of 1e-4.

2. (ii) Partial unfreezing of the top blocks with a reduced learning rate of 1e-5.

3. Optional full-model fine-tuning with a small learning rate if validation macro-F1 continued to improve.

Early stopping was based on validation macro-F1 with a patience of five epochs, reflecting the competition metric. After training each model independently, we performed soft fusion at inference time by averaging class-probability outputs from EfficientNetV2-B0 and ConvNeXt-Tiny and taking the argmax of the fused probabilities as the final prediction. This baseline pipeline provided a strong reference point but still suffered from confusion on images where multiple WBCs, RBCs, and platelets crowded the field or where extreme noise and blur obscured morphology, motivating a move toward explicit WBC localization.

3.3 WBC Localization via Pretrained YOLO and Re-classification

Because each image contained not only the target WBC but also numerous RBCs and platelets, we hypothesized that classification performance would improve if models operated on WBC-focused crops rather than full, cluttered patches. The lack of segmentation masks ruled out fully supervised mask-based segmentation, so instead, we searched for publicly available blood-cell detectors and adopted the pretrained YOLOv8m model released by keremberke for blood cell detection [6], which distinguishes WBCs, RBCs, and platelets and has been trained on images that include both clean and noisy examples. Using this detector, we automatically generated a new dataset by running YOLOv8m on all competition images, extracting bounding boxes corresponding to WBC detections, and cropping these regions without any additional preprocessing.

The resulting WBC-only crops were then used to train a ResNet-50 classifier with a 13-class output layer, using a training schedule similar to the baseline CNNs but now with substantially reduced background clutter. This two-stage pipeline—pretrained WBC detection followed by fine-grained classification—was designed to focus learning on leukocyte morphology while leveraging the robustness of YOLOv8m to moderate blur and noise.

4 Results

4.1 Baseline Architecture Comparison

Our initial experiments established a baseline performance using raw input data without extensive preprocessing. We evaluated three state-of-the-art architectures: EfficientNetV2, ResNet50, and ConvNeXt. As shown in Table 1, **EfficientNetV2** demonstrated superior feature extraction capabilities in early trials, achieving a Test F1-score of **0.6236**, significantly outperforming ResNet50 (0.5433).

4.2 Preprocessing and Progressive Learning Strategies

Given the severe class imbalance and image degradation, we evaluated several strategies to improve generalization.

4.2.1 Imbalance Correction (Loss vs. Data)

We attempted to explicitly penalize the model for misclassifying minority classes using Manual Class Weights and Focal Loss. Contrary to standard practice, these interventions degraded performance (EfficientNet F1 dropped to 0.5801 and 0.5787, respectively). We hypothesize that because the rare classes (e.g., Prolymphocytes) are heavily degraded, aggressive penalization forced the model to overfit on noise artifacts rather than learning generalized features.

4.2.2 Gradual Unfreezing

In contrast to loss manipulation, data-centric strategies yielded positive results. A pipeline combining Median Filtering, Gaussian Deblurring, and targeted augmentation raised the F1-score to 0.6082. Furthermore, implementing a **Gradual Training** strategy—initially freezing the backbone and progressively unfreezing layers—provided the most significant boost, raising the EfficientNet performance to **0.6308** and ConvNeXt to 0.6200.

4.2.3 ROI Extraction Analysis (YOLOv8)

We tested a detection-based approach using a pretrained YOLOv8 model to crop WBCs prior to classification. This yielded a Test F1-score of **0.5950**, lower than our baseline. The performance drop is attributable to domain mismatch; the YOLO model, not fine-tuned on our specific 13 classes, occasionally truncated nuclei or missed cells entirely, confirming that detection requires domain-specific training.

4.3 Detailed Per-Class Performance

To understand the specific failure modes of our pipeline, we analyzed the class-wise Accuracy for our top two models (EfficientNet-B0 and ConvNeXt-Tiny) on the validation set.

The results (Table 2) reveal a stark dichotomy in performance between mature and immature cells:

Table 1: Summary of experimental results on the Phase 1 WBC Benchmark.

Experiment / Method	**Test F1-Score**	**Key Observation**
ResNet50 (Baseline)	0.5433	Failed to capture fine textures.
ConvNeXt (Baseline)	0.5800	Competitive but heavier.
EfficientNetV2 (Baseline)	0.6236	Best single-model baseline.
EffNet + Class Weights	0.5801	Overfitting on noisy samples.
EffNet + Preproc + Aug	0.6082	Denoising helped generalization.
EffNet + Gradual Finetune	**0.6308**	Significant gain from unfreezing.
YOLOv8 Crop + EffNet	0.5950	Detection errors hurt classification.
Ensemble	**0.6345**	**Best Result (Rank #34).**

- **Mature Cells (Solved):** Both models achieved near-human performance on mature cells. Segmented Neutrophils (SNE) reached **98.01%** accuracy (ConvNeXt), and Eosinophils (EO) reached **98.82%**.
- **The Precursor Trap:** Performance collapsed for the immature granulocytic lineage. Band Neutrophils (BNE), which differ from SNE only by subtle nuclear indentation, achieved only **28.17%** accuracy. Similarly, Metamyelocytes (MMY) hovered around **47%**. This indicates the models struggle to distinguish subtle maturation stages in the presence of noise.
- **The Rare Class Limit:** The rarest classes suffered the most. Prolymphocytes (PLY) had **0.00%** accuracy (0/1 correct), and Plasma Cells (PC) achieved only **50–62%**.

Table 2: Validation Accuracy by Class. Note the massive performance gap between mature cells ($> 95\%$) and precursors ($< 50\%$).

Class	**EfficientNet-B0 Acc.**	**ConvNeXt-Tiny Acc.**	**Winner**
Mature Cells			
Neutrophil (SNE)	97.44%	**98.01%**	ConvNeXt
Lymphocyte (LY)	95.67%	**96.44%**	ConvNeXt
Eosinophil (EO)	98.22%	**98.82%**	ConvNeXt
Immature/Rare			
Band Neutrophil (BNE)	**32.39%**	28.17%	EffNet
Metamyelocyte (MMY)	**48.94%**	46.81%	EffNet
Plasma Cell (PC)	50.00%	**62.50%**	ConvNeXt
Prolymphocyte (PLY)	0.00%	0.00%	–

4.4 Evaluation of Vision Transformers (ViT)

Given the popularity of Foundation Models, we evaluated two Transformer-based architectures: **DinoBloom-B** (pre-trained on histology) and **SCKansformer** (trained from scratch). Both failed to match CNN performance (Table 3).

- **DinoBloom (Domain Shift):** Despite massive pre-training, DinoBloom suffered from severe overfitting (Train F1: 0.79 vs. Val F1: 0.52). The model's features, learned from clean histology slides, did not transfer robustly to the noisy, degraded smear images in this dataset.

- **SCKansformer (Data Starvation):** Training a Transformer from scratch proved infeasible. Over 14 epochs, the model failed to converge (Best Val F1: 0.25), confirming that Transformers lack the inductive bias required to learn effective features from small datasets (∼15k images).

Table 3: Comparison of CNN vs. Transformer architectures. Transformers failed to match CNN performance due to the lack of domain-aligned pre-training and insufficient data scale.

Model	Best Val F1	Best Val Acc	Failure Mode
EfficientNetV2 (CNN)	**0.6398**	**92.15%**	– (Baseline)
DinoBloom-B (ViT)	0.5204	59.10%	Domain Shift / Overfitting
SCKansformer (KAN)	0.2509	14.71%	Data Starvation

4.5 Dataset Gap Analysis

To address the data scarcity identified in the previous section, we conducted a systematic audit of six public datasets (Raabin, PBC, AML-LMU, MICCAI '24, SegPC, ALL-IDB).

As shown in Table 4, external data can solve the scarcity of Blasts and Plasma Cells. However, **Prolymphocytes (PLY)** remain absent from all surveyed public repositories. This "Critical Gap" mathematically justifies the need for Generative AI, as no external source exists to augment this class.

Table 4: Gap Analysis. While specialist datasets exist for Leukaemic Blasts and Plasma Cells, Prolymphocytes are missing from the public domain.

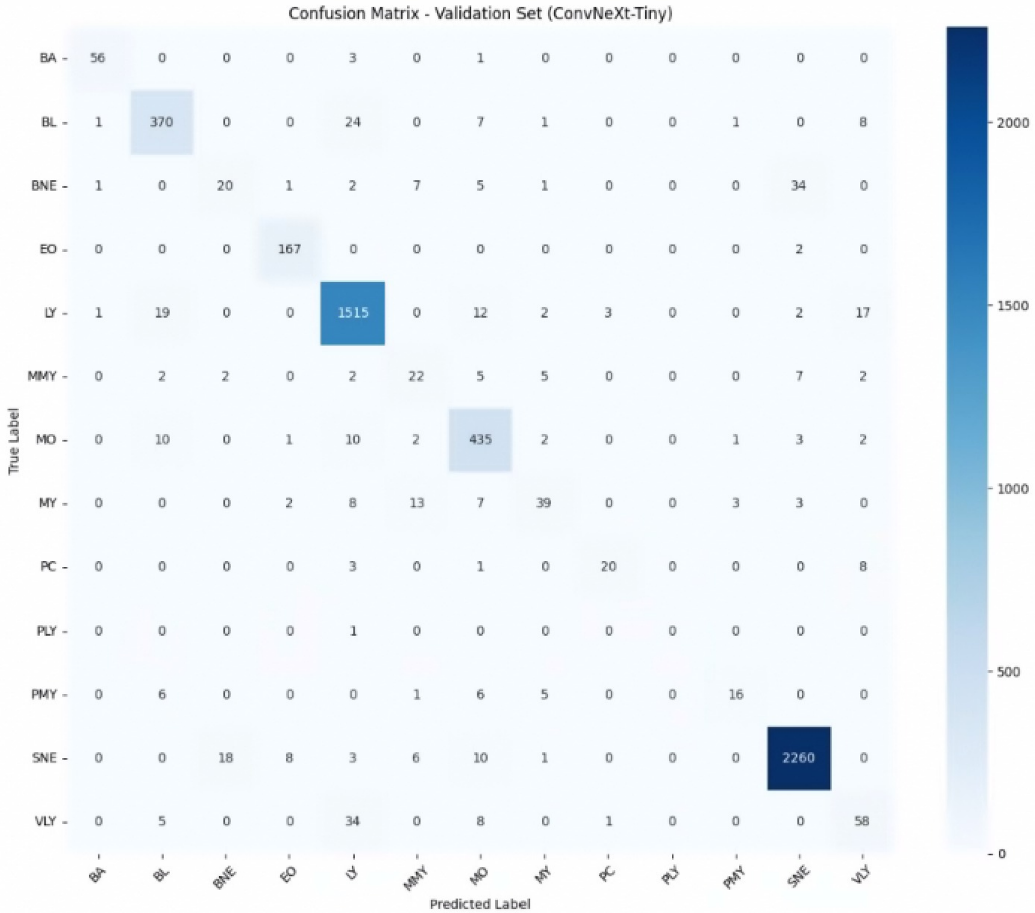
Class	Our Data	Best External Source	Gain	Status
Neutrophil	15,048	Raabin-WBC / PBC	>2.0x	Solved
Blast (BL)	2,271	MICCAI 2024	7.5x	Solved
Plasma (PC)	58	SegPC-2021	>10x	Solved
Prolymphocyte	**13**	**None (0 images)**	**0.0x**	**Critical Gap**

4.6 Final Ensemble Performance

Our best result was achieved by creating a Soft-Voting Ensemble of our two best fine-tuned models: EfficientNetV2 (0.6308) and ConvNeXt (0.6200). By averaging the softmax probabilities, the ensemble smoothed out individual model variances—leveraging ConvNeXt's precision on mature cells and EfficientNet's sensitivity to precursors—achieving a final Test F1-score of **0.6345**. This result secured rank **#34** on the competition leaderboard.

5 Conclusion and Future Work

This project investigated a range of preprocessing and modeling strategies for 13-class WBC classification under challenging conditions that include class imbalance, multiple cells per image, absence of



	BA	BL	BNE	EO	LY	MMY	MO	MY	PC	PLY	PMY	SNE	VLY
BA	56	0	0	0	3	0	1	0	0	0	0	0	0
BL	1	370	0	0	24	0	7	1	0	0	1	0	8
BNE	1	0	20	1	2	7	5	1	0	0	0	34	0
EO	0	0	0	167	0	0	0	0	0	0	0	2	0
LY	1	19	0	0	1515	0	12	2	3	0	0	2	17
MMY	0	2	2	0	2	22	5	5	0	0	0	7	2
MO	0	10	0	1	10	2	435	2	0	0	1	3	2
MY	0	0	0	2	8	13	7	39	0	0	3	3	0
PC	0	0	0	0	3	0	1	0	20	0	0	0	8
PLY	0	0	0	0	1	0	0	0	0	0	0	0	0
PMY	0	6	0	0	0	1	6	5	0	0	16	0	0
SNE	0	0	18	8	3	6	10	1	0	0	0	2260	0
VLY	0	5	0	0	34	0	8	0	1	0	0	0	58

Figure 1: ConvNeXt Confusion Matrix.

segmentation masks, and severe noise, blur, and staining variability. Starting from direct classification on raw images using CNNs and vision transformers and progressing to a two-stage pipeline with pretrained YOLOv8m WBC localization followed by fine-grained ResNet-50 classification, the experiments showed that heavy restoration methods such as aggressive classical filtering or learning-based denoisers did not consistently recover reliable morphology and sometimes degraded performance, whereas light, morphology-preserving processing combined with robust backbones provided a more stable foundation. The final ensemble of models reached a macro F1-score of about 0.63 on the competition metric, which is encouraging given the dataset difficulty but still leaves a substantial gap to clinically desirable performance.

Future work will focus on two complementary directions aimed at closing this gap. First, conditional fusion and targeted augmentation will be explored for rare classes: by learning class-aware or difficulty-aware fusion schemes and generating additional synthetic examples for under-represented categories, the model may better capture subtle morphological patterns in precursors and atypical cells. Second, heavier and more diverse augmentations will be applied upstream of both segmentation and classification, including intensity, geometric, and degradation-style perturbations that mimic the

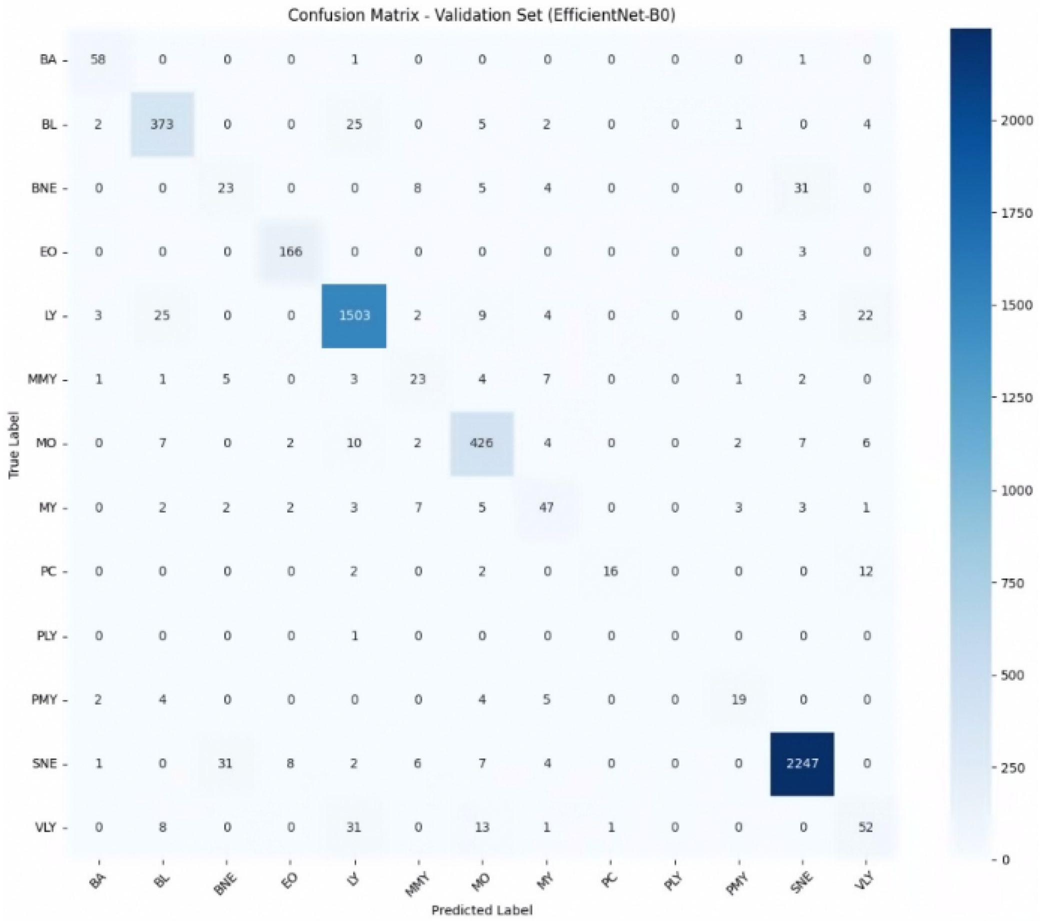


Figure 2: EffecientNET Confusion Matrix.

full spectrum of artifacts present in the competition data, with the goal of training detectors and classifiers that are explicitly robust to extreme noise and blur.

6 Team Members Contribution

1. Dina Zahran: Explored different data preprocessing methods, trained the YOLOv11 classifier, and constructed pipeline of segmentation and classification using YOLOv8m.

2. Ali Marzban: Trained and fine-tuned different CNN-based models, including ResNET50, ConvNeXt, EffecientNET, etc., and also ensembling some of them together using soft voting.

3. Karim ElTohamy: Trained and fine-tuned Vision Transformers, applied segmentation pipeline, and trained on Resnet50.

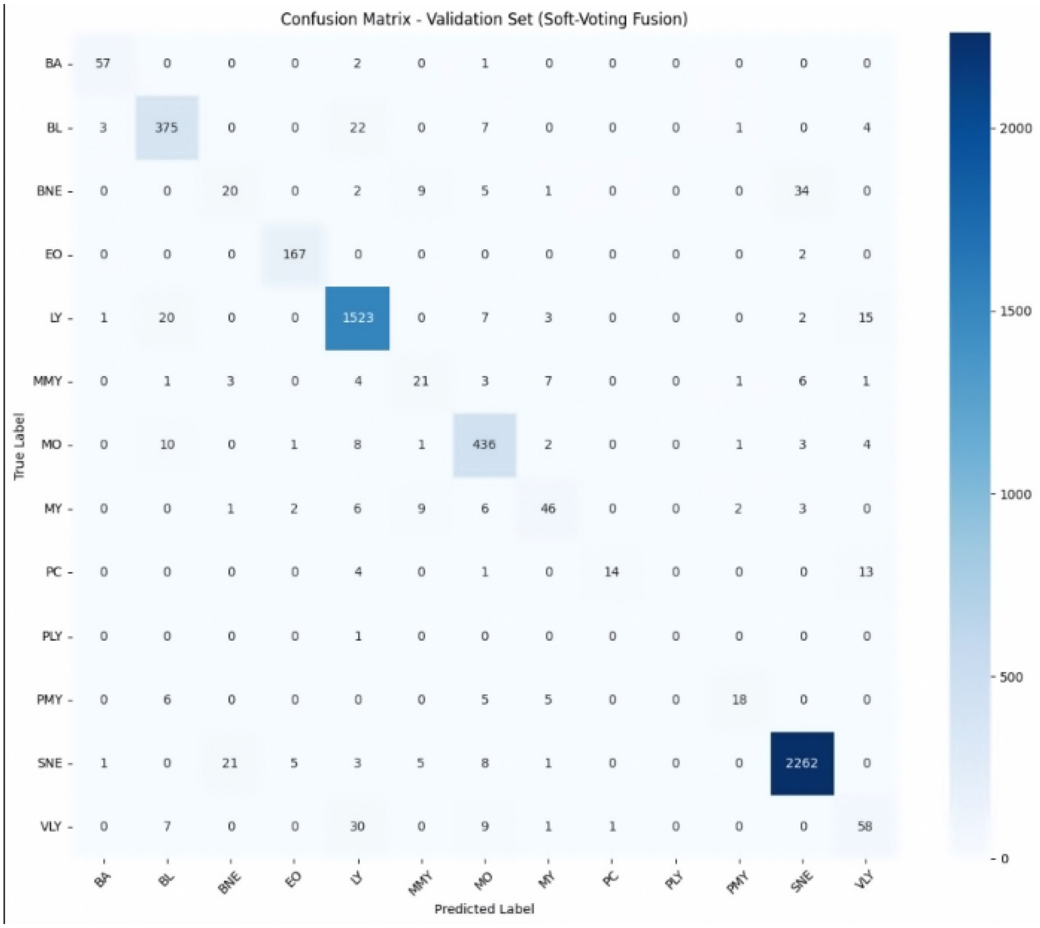


Figure 3: Soft-Voting Confusion Matrix.

References

1. Asghar R, Kumar S, Shaukat A, and Hynds P. Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review. PLoS One 2024;19:e0292026.

2. Shahzad M, Ali F, Shirazi SH, et al. Blood cell image segmentation and classification: a systematic review. Journal Name 2010.

3. Saidani O, Umer M, Alturki N, et al. White blood cells classification using multi-fold preprocessing and optimized CNN model. Scientific Reports 2024;14:3570.

4. Tarimo SA, Jang MA, Ngasa EE, Shin HB, Shin H, and Woo J. WBC YOLO-ViT: 2 Way - 2 stage white blood cell detection and classification with a combination of YOLOv5 and vision transformer. Computers in Biology and Medicine 2024;169:107875.

5. Rao A, Kho KW, Mykytiv V, and Visentin A. Deep Learning Pipeline for Blood Cell Segmentation, Classification and Counting. In: *Proceedings of the 32nd Irish Conference on Artificial In-*

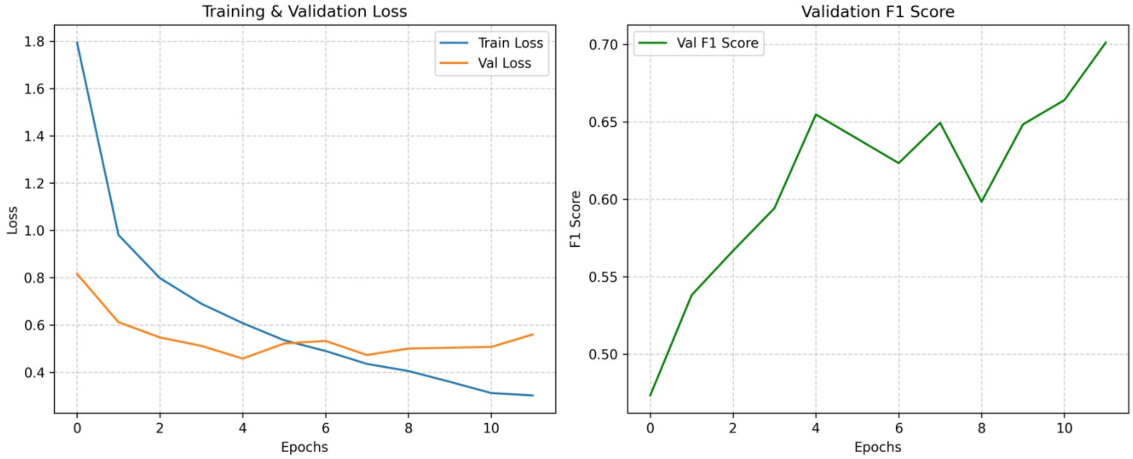


Figure 4: ResNET50 before Segmentation

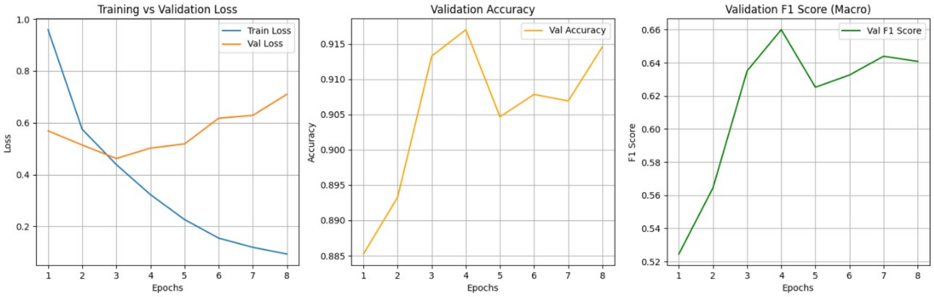


Figure 5: ResNET50 after Segmentation

telligence and Cognitive Science (AICS 2024). Vol. 3910. CEUR Workshop Proceedings. Dublin, Ireland: CEUR-WS.org, 2024:460–72. URL: https://ceur-ws.org/Vol-3910/aics2024_p72.pdf.

6. keremberke. keremberke/yolov8m-blood-cell-detection. https://huggingface.co/keremberke/yolov8m-blood-cell-detection. Accessed on December 20, 2025. 2025. URL: https://huggingface.co/keremberke/yolov8m-blood-cell-detection.