Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 9 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i09.61729

PAPER

A Calibrated Lightweight CNN Ensemble with CIELAB-CLAHE Preprocessing for White Blood Cell Classification and Hematological Disorder Screening

Aya Achir[1,2] (✉), **Ilham Battas**[2,3], **Ikram Debbarh**[2,3] **Hicham Medromi**[2,3], **Fouad Moutaouakkil**[1]

[1]System Architecture Team (EAS) – Networks, Computer Science, Telecommunications and Multimedia Laboratory (RITM) National High School of Electricity and Mechanic (ENSEM), Hassan II University of Casablanca, Morocco

[2]Research foundation for Development and Innovation in Science and Engineering (FRDISI), Casablanca, Morocco

[3]Graduate School of Biomedical Engineering and Health Techniques (SUPTECH-SANTE), Mohammedia, Morocco

aya.achir.doc23@ensem.ac.ma

ABSTRACT

White blood cell (WBC) classification is a clinically critical task in hematology, underpinning the diagnosis of blood malignancies, including leukemia, as well as infections and immune disorders. Accurate automated WBC differential counting is essential for early detection of conditions such as chronic lymphocytic leukemia (CLL) and acute myeloid leukemia (AML), where abnormal leukocyte counts and morphology are primary diagnostic indicators. Manual differential counting is time-consuming, subjective, and prone to inter-observer variability. This paper proposes a weighted ensemble framework combining three lightweight pretrained convolutional neural networks (CNNs): EfficientNet-B0, MobileNetV3-Small, and MobileNetV3-Large for automated four-class WBC classification on Paul Mooney's Blood Cell Images dataset (BCCD, Kaggle). A standardized preprocessing pipeline was employed to enhance nuclear contrast and normalize photometric variability. Each backbone was fine-tuned under a two-phase transfer-learning protocol with AdamW optimization and cosine annealing scheduling. Model evaluation was conducted via stratified 3-fold cross-validation, and ensemble predictions were aggregated using accuracy-weighted soft voting. The proposed ensemble achieved a classification accuracy of 98.57%, a weighted F1-score of 0.9857, a Matthews Correlation Coefficient (MCC) of 0.9810, and a mean AUC of 0.9994. These results demonstrate the effectiveness of ensemble-based transfer learning combined with targeted preprocessing for robust, well-calibrated WBC classification, with direct applicability to clinical laboratory automation.

KEYWORDS

white blood cell (WBC) classification, leukemia diagnosis, deep learning, ensemble learning, EfficientNet-B0, MobileNetV3, CLAHE, transfer learning, peripheral blood smear, convolutional neural network (CNN), hematology

Achir, A., Battas, I., Debbarh, I., Medromi, H., Moutaouakkil, F. (2026). A Calibrated Lightweight CNN Ensemble with CIELAB-CLAHE Preprocessing for White Blood Cell Classification and Hematological Disorder Screening. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(9), pp. 148–171. https://doi.org/10.3991/ijoe.v22i09.61729

Article submitted 2026-03-27. Revision uploaded 2026-05-17. Final acceptance 2026-05-19.

1 INTRODUCTION

White blood cells (WBCs), or leukocytes, are central to the human immune response, and their differential counts are vital for diagnosing a wide range of conditions from infections and autoimmune disorders to hematological malignancies such as leukemia [1]. Conventionally, expert pathologists examine stained blood smears under a microscope to categorize WBC subtypes. However, this manual process is slow, tedious, and subject to observer bias, often leading to inconsistent or error-prone results. These limitations, coupled with the growing clinical demand for rapid, high-throughput hematology screening, have motivated the development of automated, computer-aided systems. Automated WBC analysis promises fast and reproducible classification, thereby reducing technician workload and enabling more timely diagnosis and monitoring of diseases [2]. In recent years, deep learning [3] and, in particular, convolutional neural networks (CNNs) [4] have achieved strong performance across a broad range of medical image analysis tasks, including cancer detection, retinal image classification, and histopathology [1]. CNNs have been applied successfully to microscopic blood-cell images with considerable success. For example, Ali et al. combined a Vision Transformer (ViT) with a CNN backbone to classify leukocytes, reporting accuracy above 95% on a custom five-class dataset [5]. Transfer learning using ImageNet-pretrained models has become a widely adopted strategy for WBC classification, enabling strong performance even when annotated datasets are limited in size [1]. Recent work demonstrates that incorporating domain-specific normalization further improves discriminability across WBC subtypes [1] [2].

Despite these successes, WBC image classification poses significant challenges. Blood smear images can vary widely in staining quality and color balance due to differing laboratory protocols, and white cells exhibit complex morphology that may overlap between subtypes. Annotated medical datasets are often limited in size, which hinders training of large networks. Much of the existing literature focuses primarily on optimizing individual architectures for maximum accuracy, often relying on complex pipelines or highly specialized network designs, making direct and fair comparison between architectures difficult. Moreover, cross-dataset comparisons, which are common in the WBC literature, can obscure true performance differences due to variations in class counts, acquisition conditions, and annotation protocols.

In this context, the present study proposes a principled and reproducible ensemble framework that combines three efficient pretrained architectures, EfficientNet-B0, MobileNetV3-Small, and MobileNetV3-Large, under a unified two-phase transfer-learning protocol, evaluated via rigorous 3-fold stratified cross-validation on the BCCD benchmark. The main contributions of this work are as follows:

- **i)** A systematically designed integration pipeline combining CLAHE-based luminance preprocessing, on-the-fly augmentation, two-phase fine-tuning, and accuracy-weighted soft-voting ensemble aggregation, evaluated under a fully reproducible cross-validation framework.
- **ii)** An empirical comparison of three lightweight CNN architectures under strictly identical experimental conditions, identical splits, hyperparameters, augmentation, and evaluation metrics demonstrating the benefit of ensemble aggregation over individual models for WBC classification.

iii) A comprehensive probabilistic evaluation including Brier Scores, calibration curves, and MCC demonstrating that a carefully integrated ensemble of lightweight models achieves strong performance with computational efficiency suitable for resource-constrained hardware deployment.
iv) An explicit analysis of cross-dataset comparison limitations in the WBC classification literature, with recommendations for more rigorous and reproducible benchmarking practices.

2 RELATED WORK

Automated WBC classification has been extensively investigated as a core component of computer-aided leukemia diagnosis. Over the past decade, research in this domain has progressively evolved from traditional machine learning pipelines relying on handcrafted features toward deep learning and, more recently, transformer-based architectures. This section reviews the main methodological trends, highlights their limitations, and positions the contribution of the present study.

2.1 Classical machine learning approaches for WBC classification

Early studies on WBC classification primarily relied on manually engineered features describing cell morphology, texture, and color distribution, combined with conventional classifiers such as support vector machines (SVMs), k-nearest neighbors (k-NN), and random forests. These pipelines typically followed a sequential workflow: image segmentation to isolate leukocytes from erythrocytes and background, followed by feature extraction targeting nucleus shape (area, perimeter, circularity, eccentricity), cytoplasmic texture (co-occurrence matrices, Haralick features), and color histograms in HSV or LAB color spaces. Representative works demonstrated moderate to high classification performance on relatively small, institution-specific datasets. For instance, Habibzadeh et al. employed pretrained deep architectures, including ResNet and Inception variants, for early leukocyte classification comparisons, documenting the utility of feature-based and transfer-learning strategies on blood smear images [6]. Subsequent studies explored multi-domain feature extraction, including spectral and hyperspectral information, to enhance discrimination between leukocyte subtypes with overlapping morphological characteristics. Similar machine learning frameworks have been applied productively to other medical imaging domains. Research proposed an ML-based diagnostic system for breast cancer detection that combined structured feature extraction with classical classifiers, demonstrating that well-designed feature engineering pipelines can achieve clinically competitive performance even without deep neural networks. This broader body of work underscores the value of principled feature design and task-specific preprocessing as a foundation for medical image classification systems. Although classical ML approaches demonstrated the feasibility of automated WBC analysis, their strong dependence on handcrafted features introduced significant limitations: inter-observer variability in feature definition, sensitivity to staining and imaging conditions, poor scalability to large and heterogeneous datasets, and fundamentally constrained representational capacity compared to learned representations. These limitations motivated the shift toward end-to-end deep learning methods.

2.2 CNN-based deep learning methods

With the availability of larger annotated datasets, CNNs rapidly became the dominant paradigm for WBC classification. CNN-based approaches enable end-to-end feature learning directly from raw images, eliminating the need for manual feature design and providing hierarchical representations that capture morphological structure at multiple spatial scales. Early CNN-based systems already demonstrated classification accuracies above 96% on WBC and leukemia-related datasets [7], and subsequent studies employing deeper architectures and domain-adapted training protocols further improved performance [8]. Transfer learning using pre-trained backbones such as VGG16 [8] and ResNet variants [9] became particularly popular due to their strong representational capacity and reduced training requirements on limited medical datasets. Comparisons between VGG-16 and ResNet-50 architectures on blood smear images highlighted that architecture selection and fine-tuning strategy substantially influence final classification accuracy [10] [11].

More advanced CNN designs incorporated attention mechanisms, multi-scale feature extraction, or hybrid CNN–SVM classification heads to capture subtle morphological variations among leukocyte subtypes. Multi-scale feature pyramids, in particular, have enabled models to simultaneously capture fine nuclear texture and coarser cytoplasmic context, yielding improved discrimination between morphologically similar classes such as eosinophils and neutrophils. While these works demonstrated the effectiveness of deep CNNs for hematological image classification, many rely on complex architectures, heavy preprocessing pipelines, or aggressive data augmentation strategies, which may inflate performance on specific benchmarks while reducing reproducibility and practical deployment potential. The computational cost of large CNN models further limits their applicability in resource-constrained laboratory environments, motivating the investigation of lightweight and efficient alternatives.

2.3 Transfer learning, hybrid models, and state-of-the-art performance

Recent studies have reported accuracies approaching or exceeding 99% by combining deep CNN backbones with hybrid classification strategies and advanced preprocessing. Transfer learning frameworks employing multi-fold preprocessing with optimized CNN architectures have demonstrated strong performance gains on public WBC datasets [12]. Previous studies have proposed a framework integrating deep and transfer learning for WBC classification in medical educational settings, reporting competitive accuracy with efficient training protocols [11]. Similarly, lightweight and optimized CNN architectures have been shown to match the performance of heavier models when preprocessing and training pipelines are carefully designed [12]. The use of Inception-based [12] and residual network architectures [9] further extended the representational diversity available to ensemble-based approaches. Importantly, the success of AI-powered image analysis has extended well beyond WBC classification to a broad range of clinical imaging tasks. The integration of artificial intelligence and advanced image processing techniques has demonstrated strong potential for improving diagnostic accuracy across diverse medical imaging domains, including retinal analysis and neurological imaging.

Recent studies employing transformer-based architectures with multi-scale feature fusion strategies have further shown promising performance in complex diagnostic tasks involving MRI data [13], [14]. These cross-domain investigations collectively illustrate that the core principles underlying effective AI-based image analysis, robust preprocessing, multi-scale feature extraction, and domain-adapted transfer learning are broadly applicable across diverse biomedical imaging modalities and diagnostic targets and directly inform the design choices of the present WBC classification framework.

In parallel, transformer-based models have recently been introduced for WBC classification. ViT and hybrid ViT–CNN architectures have achieved competitive performance while offering alternative representational mechanisms compared to purely convolutional networks [5]. These attention-based models demonstrate strong discriminability on multi-class WBC datasets, though they typically require larger computational budgets and more training data than lightweight CNN alternatives. Swin Transformer architectures, which leverage hierarchical windowed self-attention, have emerged as a computationally efficient transformer variant well-suited to dense prediction tasks in medical imaging, and their application to hematological image analysis is an active area of investigation.

Another important trend concerns explainable artificial intelligence (XAI). Several studies have incorporated interpretability tools such as Grad-CAM, SHAP, or LIME to visualize discriminative regions in WBC images and improve clinical trust in model predictions. Although these approaches enhance transparency and facilitate regulatory acceptance, they often rely on highly optimized, dataset-specific pipelines that may not generalize well across different acquisition conditions or staining protocols.

2.4 Lightweight CNN architectures in biomedical imaging

The deployment of deep learning models in clinical and laboratory settings imposes practical constraints on model size, inference speed, and power consumption that large-scale architectures such as VGG19 or ResNet-152 are ill-suited to satisfy. This has motivated a growing body of research into lightweight CNN architectures specifically designed to balance accuracy with computational efficiency. MobileNet variants, SqueezeNet, ShuffleNet, and EfficientNet-B0 have all been explored in the context of biomedical image classification, demonstrating that depthwise separable convolutions, inverted residuals, and compound scaling strategies can substantially reduce parameter counts without proportional loss of discriminative power. EfficientNet-B0 applies a principled compound scaling of network depth, width, and input resolution, achieving favorable accuracy-per-parameter ratios on ImageNet [15] that transfer effectively to medical imaging tasks. MobileNetV3, which incorporates inverted residual blocks augmented with squeeze-and-excitation attention modules and hard-swish activations, further reduces both latency and memory footprint while maintaining competitive representational capacity. Studies in WBC classification have confirmed that lightweight architectures pretrained on ImageNet achieve strong performance on blood smear images with substantially lower computational cost than heavier alternatives [12], making them particularly attractive for resource-constrained point-of-care diagnostics and automated laboratory systems. The complementary

architectural properties of EfficientNet-B0 and MobileNetV3 variants provide a strong motivation for their combination in an ensemble framework, as each model's distinct feature extraction pathway can capture distinct aspects of leukocyte morphology.

2.5 Ensemble learning strategies in medical diagnosis

Ensemble learning, which combines the predictions of multiple base classifiers to produce a final decision, is a well-established strategy for improving classification accuracy and robustness in machine learning. In medical image analysis, ensemble methods are particularly valuable because individual models may exhibit systematically different error patterns corresponding to distinct morphological ambiguities, and combining their predictions can correct errors that no single model resolves reliably. Common ensemble strategies employed in the medical imaging literature include majority voting, soft (probability-averaging) voting, stacking, and weighted averaging based on cross-validation performance. Soft-voting ensemble approaches, wherein posterior class probability vectors are averaged or weighted across constituent models, have consistently demonstrated advantages over hard-voting schemes in probabilistic classification settings, producing better-calibrated confidence estimates in addition to improved point accuracy. In the WBC classification domain specifically, ensemble combinations of CNN architectures have been shown to surpass individual model performance on standard benchmarks, confirming the complementary nature of diverse CNN feature extractors for leukocyte recognition [2]. More broadly, AI-based image analysis and ensemble strategies have been applied across a spectrum of medical imaging tasks, including retinal disease screening and pulmonary disease identification [16], [17], where the aggregation of diverse model predictions can improve decision boundary stability. The accuracy-weighted soft-voting scheme employed in the present study represents a principled, data-driven instantiation of this broad strategy, where each model's contribution to the ensemble is scaled according to its empirically measured cross-validation performance, naturally down-weighting less accurate models without manual tuning.

2.6 Image enhancement and preprocessing methods in medical imaging

Preprocessing plays a foundational role in medical image classification pipelines, directly influencing the quality of features available to downstream learning algorithms. In hematological imaging, photometric variability arising from differences in Giemsa-Wright staining concentration, microscope illumination, and camera sensor response can substantially degrade classifier performance when images are not normalized prior to training. Histogram equalization and its adaptive variants have been widely applied to mitigate this variability. Contrast-Limited Adaptive Histogram Equalization (CLAHE) [18], which performs local contrast normalization with a clip limit to prevent noise over-amplification, is particularly well-suited to blood smear images, where nuclear regions of interest occupy only a fraction of the image field. Applying CLAHE in the luminance channel of the CIELAB color space,

rather than directly to the RGB channels, decouples contrast enhancement from color information and prevents hue distortion that would otherwise corrupt staining-color features important for cell type discrimination. Color space transformations have been similarly employed in retinal image preprocessing for diabetic retinopathy detection [16] , where normalized luminance channels have been shown to improve vessel and lesion contrast without sacrificing color-based diagnostic cues. Gaussian filtering, applied as a preliminary denoising step, reduces high-frequency shot noise from image sensors prior to contrast normalization, preventing noise amplification during CLAHE processing. The mathematical properties of structured image transformations have also been studied in specialized computational imaging contexts, underscoring the broader relevance of algorithmic image manipulation as a foundation for robust feature extraction in biomedical analysis. Collectively, the preprocessing literature supports the conclusion that targeted, physics-informed preprocessing pipelines tailored to the specific properties of the imaging modality and tissue type are consistently more effective than generic augmentation or normalization strategies. The preprocessing pipeline employed in the present study was designed with this principle in mind, combining Gaussian denoising and luminance-channel CLAHE within a reproducible, standardized workflow applied uniformly across all images and folds.

The present study addresses these limitations by systematically integrating an established preprocessing techniques, CLAHE in the CIELAB color space, with a lightweight ensemble of pretrained CNN backbones under a principled cross-validation scheme. Rather than introducing entirely new algorithmic components, the contribution of this work lies in the careful engineering integration, empirical optimization, and rigorous evaluation of established methods on the four-class BCCD benchmark, demonstrating that a well-designed pipeline of standard components can achieve strong classification performance with computational efficiency.

3 MATERIALS AND METHODS

3.1 Research methodology

We adopted a structured, reproducible pipeline for WBC image classification, built around three interconnected stages: standardized image preprocessing, transfer-learning-based model training, and ensemble aggregation. The pipeline was designed to ensure fair comparison across architectures by maintaining identical data splits, augmentation strategies, optimization hyperparameters, and evaluation protocols throughout all experiments. A fixed random seed (seed = 42) was applied at all stochastic stages to ensure full reproducibility.

Overall, the methodological workflow presented in Figure 1 reflects a deliberate design choice: rather than pursuing a single highly tuned architecture, we systematically combine three complementary lightweight CNNs under a principled ensemble scheme, evaluated through stratified cross-validation, to produce robust and well-calibrated WBC classification predictions.

Algorithm 1: Proposed Methodology for WBC Ensemble Classification

Description: A three-architecture, stratified 3-fold cross-validated, two-phase transfer learning and soft-voting ensemble approach for accurate WBC classification.

Input: Dataset with TRAIN and TEST folders — 4 WBC classes: Eosinophil, Lymphocyte, Monocyte, Neutrophil
Define CON; img_size = 128, batch_size = 64, epochs = 30
- lr_phase1 = 3e-4, lr_phase2 = 5e-5, wd = 1e-4, K_folds = 3, seed = 42, patience = 8, max_norm = 1.0

Load images → apply Gaussian blur (3×3) + CLAHE (L channel, clip = 2.0, grid 8×8)* + Resize 128×128 px
all_paths, all_labels + merge TRAIN + TEST
for each model *M* in {EfficientNet-B0, MobileNetV3-Small, MobileNetV3-Large} **do**
    SetRandomSeed(42)
    skf ← StratifiedKFold (folds = 3, shuffle = True, random_state = 42)
    for each (*train_idx*, *val_idx*) **in** *skf* (all_paths, all_labels) **do**
        Build TrainLoader with augmentation · ValLoader normalisation only
        Load pretrained backbone M (ImageNet-1k) · attach custom head:
            Dropout(0.3) → Linear(feat_dim = 256) → BN → ReLU →
            Dropout(0.15) → Linear(256 → 4)
        Phase 1 (ep. 1–10): freeze backbone · train head only · AdamW ·
            LR = lr_phase1
        Phase 2 (ep. 11–30): unfreeze all layers · full fine-tune · AdamW ·
            LR = lr_phase2 · CosineAnnealingLR · AMP
        Apply GradientClipping(max_norm = 1.0) · EarlyStopping(patience = 8)
        logits ← M(ValLoader) · probs ← Softmax(logits) · preds ← argmax(logits)
        Compute: Accuracy, F1-W, MCC, AUC (OvR), Brier Score, Cohen's κ
        Store: fold_probs[M], fold_preds[M], fold_accs[M]
    end for (inner fold loop)
end for (outer model loop)
Compute ensemble weights:
- $w_k = \frac{\text{acc}_k}{jinfolds \sum_j \text{acc}_j}$
- $P_{\text{ens}}(c|i) = \sum w_k \cdot P_k(c|i)$
- $\hat{y}(i) = \text{argmax}_c \, P_{\text{ens}}(c|i)$

Aggregate: Mean Accuracy ± Std, F1-W, MCC, AUC
- Mean Accuracy ± F1, MCC
- AUC across 3 stross folds × 3 models

Output: confusion matrices, ROC/PR curves, calibration plots, SOTA comparison table

Fig. 1. Methodological workflow followed in the WBC classification process, from image acquisition and preprocessing through model training, cross-validation, ensemble aggregation, and final evaluation

3.2 Dataset

We utilized Paul Mooney's Blood Cell Images dataset (Kaggle) as the primary data source for this study [19]. This open-access dataset contains approximately 12,500 microscopic images of blood smear cells, distributed into four major WBC classes: Neutrophils, Eosinophils, Monocytes, and Lymphocytes, with each class represented by roughly 3,000 labeled images, providing a balanced corpus for training and evaluation. The images are Giemsa-Wright-stained peripheral blood smear fields, which highlight WBC nuclei (typically purple-blue) while erythrocytes remain unstained or pink. The Kaggle release includes 410 original microscope images with bounding-box annotations that were augmented (through rotations, flips, etc.) to produce the full 12,500-image corpus [20]. Table 1 summarizes the four WBC classes included in the dataset, together with their morphological characteristics, physiological prevalence, and associated classification challenges.

We selected this dataset for several reasons. First, it has become a widely used benchmark for WBC image classification in recent literature [1]. Second, the dataset is class-balanced, ensuring that each WBC type has sufficient examples for robust training. Third, the data are publicly available and well-annotated, promoting reproducibility. Finally, the four cell categories, neutrophils, eosinophils, monocytes, and lymphocytes, cover the most clinically significant leukocyte subtypes, spanning both granulocytes and agranulocytes [21].

Table 1. Summary of the four WBC classes in Paul Mooney's Blood Cell Images dataset, with morphological characteristics, physiological prevalence, and classification difficulty

WBC Class	Images	Morphology	Prevalence (%)	Challenge
Eosinophil	~3,000	Bilobed nucleus, pink granules	1–4	Confusion with Neutrophil
Lymphocyte	~3,000	Large round nucleus, scant cytoplasm	20–40	Blast cell overlap
Monocyte	~3,000	Kidney-shaped nucleus, abundant cytoplasm	2–8	Variable morphology
Neutrophil	~3,000	Multilobed nucleus, neutral granules	50–70	Confusion with Eosinophil

3.3 Preprocessing and augmentation

A standardized three-stage preprocessing pipeline was applied uniformly to every image prior to model input. This pipeline was designed to normalize photometric variability across the dataset while preserving the morphological discriminative features of WBC nuclei and cytoplasm.

Stage 1 Gaussian Denoising. Each image was first converted from BGR to RGB color space. A 3×3 Gaussian blur with $\sigma = 0$ was then applied in the spatial domain to suppress high-frequency shot noise without degrading nuclear boundary sharpness. This corresponds to the convolution:

$$I_2(x,y) = \left(\frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2} \right) * I_1 \right)(x,y) \quad (1)$$

where I_1 is the input image and $G\sigma$ is the Gaussian kernel.

Stage 2 CLAHE enhancement. CLAHE [18] was applied exclusively to the luminance (L*) channel of the CIELAB color space, with clip limit 2.0 and tile grid size 8×8. The CIELAB conversion decouples luminance from chrominance so that contrast enhancement operates independently of color information. The CLAHE redistribution for a tile with histogram h(k) and pixel count N is:

$$h'(k) = \min(h(k), \beta), \text{ then redistributed uniformly, with } \beta = \text{clip} \cdot (N/L) \quad (2)$$

where L is the number of grey levels and β is the per-tile clip limit. This approach enhances local contrast, particularly within the heterochromatin regions of WBC nuclei, while preventing over-amplification of background noise that would occur with standard global histogram equalization. Figure 2 illustrates the preprocessing pipeline on representative samples from each class.

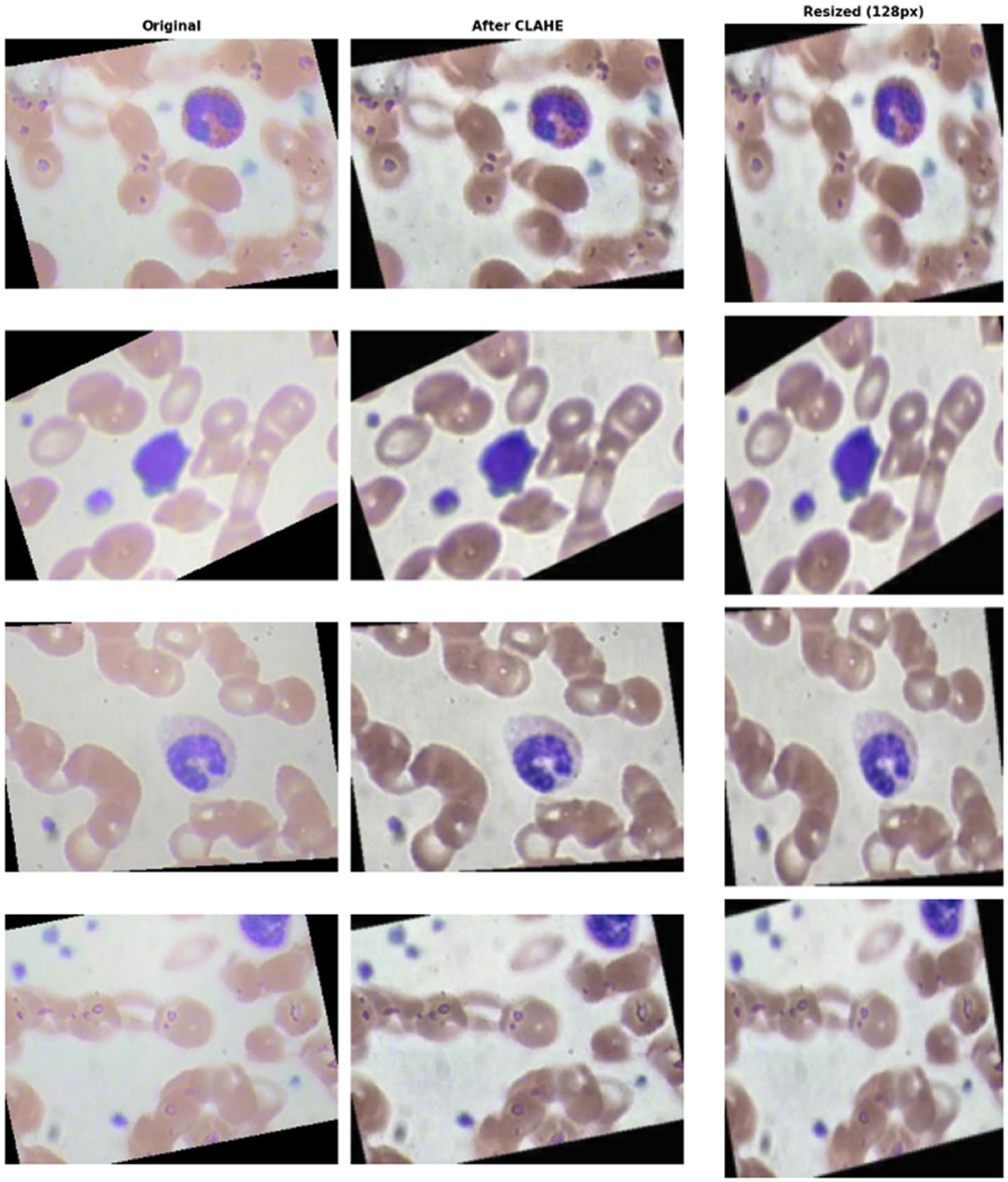

Fig. 2. Preprocessing pipeline visualization (Original → CLAHE → Final). Each row corresponds to one representative sample. Columns (left to right): original RGB image; image after CLAHE enhancement on the L* channel; final 128×128 resized output fed to the network

Stage 3 Resizing. All images were bilinearly interpolated to a uniform 128×128 pixel resolution. This input dimension balances model capacity, GPU memory constraints, and inference latency.

To improve generalization and reduce overfitting on the training split of each fold, an augmentation chain was applied on-the-fly via PyTorch transforms: random horizontal flip (p = 0.50), random vertical flip (p = 0.30), random rotation within [–15°, +15°], color jitter (brightness ±0.2, contrast ±0.2, saturation ±0.1), and random erasing (p = 0.15). All augmented tensors were normalized using ImageNet channel statistics (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]). Augmentation was exclusively applied during training; validation inference used normalization only.

3.4 Model architectures

Three lightweight CNN architectures, all pretrained on ImageNet-1k [22], were selected as individual classifiers within the ensemble: EfficientNet-B0 [15], MobileNetV3-Small, and MobileNetV3-Large [23]. The selection criteria were: (i) high accuracy-to-parameter ratio enabling edge-device deployment; (ii) demonstrated transferability to medical imaging domains; and (iii) architectural diversity to ensure complementary error distributions in the ensemble. Each backbone was loaded via the timm library with global average pooling (num_classes = 0, global_pool = 'avg'). A shared custom classification head was appended after an automated feature-dimension dry-run. The head architecture was: Dropout(p = 0.3) → Linear(feat_dim → 256) → BatchNorm1d(256) → ReLU → Dropout(p = 0.15) → Linear(256 → 4). The final layer outputs raw logits for the four WBC classes; class probabilities are computed via softmax at inference time. Table 2 provides a comparative overview of the three backbone architectures [24].

Table 2. Comparative overview of the three CNN backbone architectures used in the ensemble

Architecture	Params(M)	ImageNet Top-1 Acc.	Key Design Feature	Head Feat. Dim
EfficientNet-B0	5.3	77.1%	Compound scaling (depth × width × resolution)	1280
MobileNetV3-Small	2.5	67.7%	Inverted residuals + Hard-Swish + SE modules	576
MobileNetV3-Large	5.4	75.2%	Wider inverted residuals + SE + Hard-Swish	960

3.5 Training protocol and hyperparameters

Training followed a two-phase transfer-learning strategy designed to stabilize head convergence before full fine-tuning. In Phase 1 (epochs 1-10), all backbone parameters were frozen, and only the classification head was optimized, allowing rapid convergence of the randomly initialized head layers without disrupting pre-trained feature representations. In Phase 2 (epochs 11-30), all parameters were unfrozen, and the entire network was fine-tuned end-to-end at a reduced learning rate to allow domain-specific adaptation of low-level features.

The AdamW optimizer [25] was used throughout both phases, with weight decay $\lambda = 1 \times 10^{-4}$ and gradient clipping at max_norm = 1.0. Learning rates were set to

3×10^{-4} in Phase 1 and 5×10^{-5} in Phase 2. A CosineAnnealingLR scheduler was applied in Phase 2 to provide smooth learning-rate decay. Mixed-precision training (torch.cuda.amp) was applied throughout to accelerate computation. An early-stopping mechanism monitored mean validation accuracy with a patience of 8 epochs, restoring the best model checkpoint upon triggering. Batch size was fixed at 64. Table 3 summarizes all key hyperparameters.

Table 3. Summary of training hyperparameters used for all three architectures

Hyperparameter	Phase 1 (Head Only)	Phase 2 (Full Fine-Tuning)
Optimizer	AdamW	AdamW
Learning rate	3×10^{-4}	5×10^{-5}
Weight decay	1×10^{-4}	1×10^{-4}
LR scheduler	Constant	CosineAnnealingLR
Epochs	1–10	11–30
Batch size	64	64
Early stopping patience	—	8 epochs
Gradient clipping	1.0	1.0

3.6 Cross-validation and ensemble strategy

Model generalization was estimated using stratified 3-fold cross-validation (K = 3, random_state = 42). Stratification ensures that each fold retains the original class distribution (25% per class), which is critical for balanced evaluation. Each fold yielded approximately 1,867 training and 933 validation samples. For every fold, all three architectures were trained and evaluated independently; no data leakage occurred between splits.

Following cross-validation, a soft-voting ensemble was constructed by aggregating the softmax probability vectors from the three individual models. Per-model ensemble weights were computed proportionally to each model's mean cross-validation accuracy:

$$w_k = \bar{a}cc_k / \Sigma_j \bar{a}cc_j \tag{3}$$

$$P\_ens(c \mid i) = \Sigma_k \, w_k \cdot P_k(c \mid i) \tag{4}$$

$$\hat{y}(i) = argmax\_c \; P\_ens(c \mid i) \tag{5}$$

where w_k denotes the normalized weight of model k, $P_k(c \mid i)$ is the softmax probability of class c for sample i from model k, and $\hat{y}(i)$ is the final predicted class.

3.7 Evaluation metrics

Model performance was assessed using a comprehensive suite of metrics computed over the full aggregated out-of-fold predictions: (i) Accuracy and weighted F1-score (F1-W) capture overall classification correctness accounting for class balance; (ii) Matthews Correlation Coefficient (MCC) provides a balanced measure robust to

class imbalance [26]; (iii) mean one-vs-rest AUC quantifies discriminability; (iv) Brier Score measures probabilistic calibration (lower is better; naive baseline = 0.25); (v) Cohen's Kappa assesses inter-rater agreement beyond chance; (vi) calibration curves (one-vs-rest) and Precision-Recall (PR) curves with mean Average Precision (mAP) were computed for the ensemble.

4 RESULTS

This section reports the quantitative performance of all individual models and the final ensemble, evaluated over 3-fold stratified cross-validation on the BCCD dataset. Results are presented across all classification metrics, confusion matrices, ROC and PR curves, and probability calibration diagnostics.

4.1 Overall classification performance

The complete evaluation metrics for each model and the weighted ensemble, computed over aggregated out-of-fold predictions, are summarized in Table 4. The proposed ensemble achieved the highest scores across all primary metrics, attaining an accuracy of 0.9857, a weighted F1-score of 0.9857, an MCC of 0.9810, and a mean AUC of 0.9994. Among individual architectures, MobileNetV3-Small demonstrated the strongest standalone performance (accuracy = 0.9832), followed by EfficientNet-B0 (0.9732) and MobileNetV3-Large (0.9700). The consistent superiority of the ensemble over all individual models confirms the complementary discriminative properties of the three architectures.

Table 4. Classification performance of individual models and ensemble on the BCCD dataset (3-fold CV)

Model	Acc.	F1-W	MCC	AUC	Brier	Kappa
EfficientNet-B0	0.9732	0.9732	0.9643	0.9973	0.0159	0.9643
MobileNetV3-Small	0.9832	0.9832	0.9776	0.9990	0.0097	0.9776
MobileNetV3-Large	0.9700	0.9700	0.9600	0.9973	0.0184	0.9600
Ensemble (Ours)*	0.9857	0.9857	0.9810	0.9994	0.0116	0.9810

Notes: Acc. = accuracy; F1-W = weighted F1-score; MCC = Matthews Correlation Coefficient; AUC = mean one-vs-rest ROC-AUC; Brier = mean Brier score; Kappa = Cohen's κ. * = proposed method.

4.2 Training curves

The 3-fold mean training and validation curves for each architecture over 30 epochs are illustrated in Figure 3, with individual fold trajectories overlaid as faint lines and cross-fold means represented by bold lines. All three models converged reliably within the allotted training budget. EfficientNet-B0 exhibited a smooth, stable loss descent with a minimal training–validation gap, reaching a best validation accuracy of 0.9729. MobileNetV3-Small converged to a best validation accuracy of 0.9832, while MobileNetV3-Large achieved 0.9693 with a slightly wider training–validation discrepancy. The observed convergence behavior is consistent with the two-phase training strategy, characterized by rapid loss reduction during Phase 1 (epochs 1–10) and stable fine-tuning throughout Phase 2 (epochs 11–30).

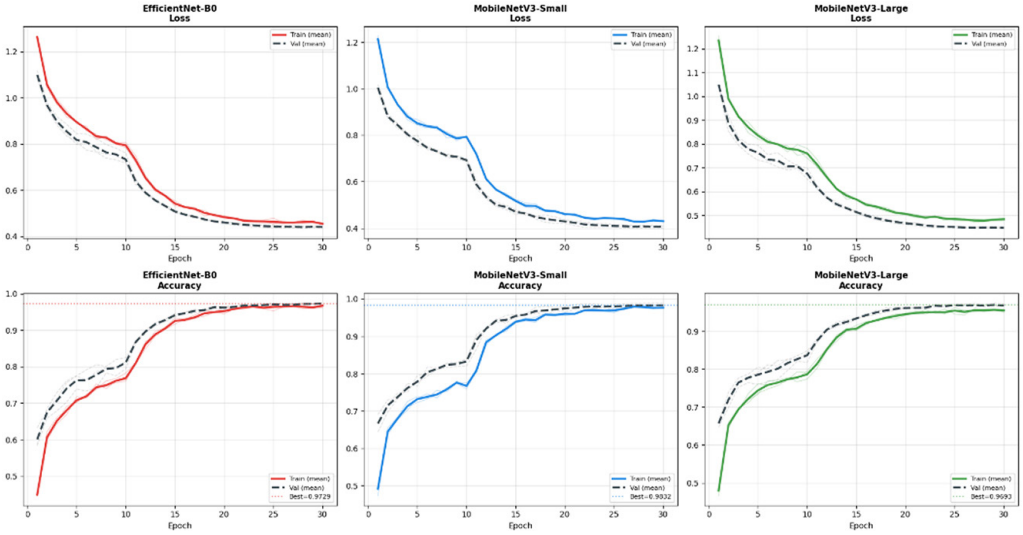

Fig. 3. Training and validation curves (3-fold CV) for all three architectures

Notes: Top row: cross-entropy loss; bottom row: classification accuracy. Solid lines: training mean; dashed lines: validation mean; faint lines: individual folds. Dotted horizontal lines: best validation accuracy per model.

4.3 Confusion matrices

The normalized confusion matrices, aggregated across all three folds, are presented in Figure 4. Lymphocyte and Monocyte were classified with near-perfect recall across all model configurations, achieving a score of 1.00. The predominant source of misclassification was mutual confusion between Eosinophil and Neutrophil, a pattern attributable to their morphological similarity, specifically, the shared bilobed versus multilobed nuclear morphology. Ensemble aggregation demonstrably mitigated this confusion, reducing the Eosinophil recall error from 4–6% in individual models to 2% and the Neutrophil recall error from 4–6% to 3%, thereby confirming a clear ensemble correction effect on the most challenging class pair.

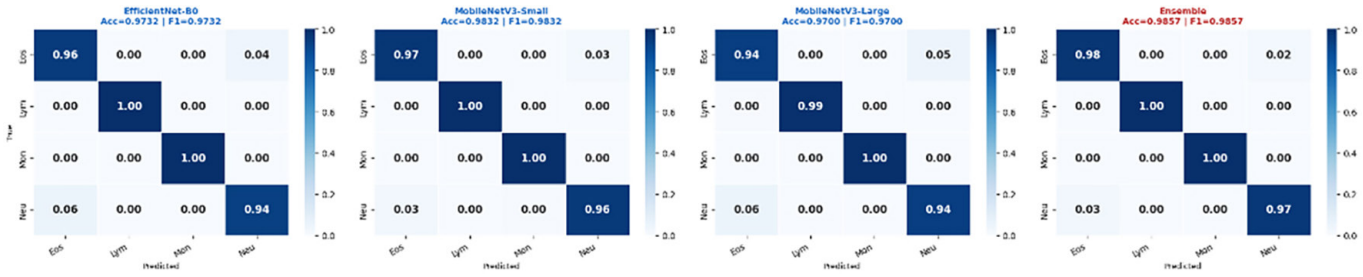

Fig. 4. Normalized confusion matrices (3-fold aggregated) for all four model configurations

Notes: EfficientNet-B0 (Acc. = 0.9732), MobileNetV3-Small (Acc. = 0.9832), MobileNetV3-Large (Acc. = 0.9700), and Ensemble (Acc. = 0.9857). Values represent class-normalized recall rates. Eos = Eosinophil; Lym = Lymphocyte; Mon = Monocyte; Neu = Neutrophil.

4.4 ROC and PR curves

The one-vs-rest ROC and PR curves for the ensemble model, computed on the full aggregated out-of-fold predictions, are shown in Figure 5. The ensemble achieved near-perfect discrimination, with a mean AUC of 0.9994 and per-class AUC values of 0.999 (Eosinophil), 1.000 (Lymphocyte), 1.000 (Monocyte), and 0.999 (Neutrophil).

The mAP reached 0.9983, with per-class AP values of 0.997, 1.000, 1.000, and 0.996, respectively. These results collectively confirm that the classifier operates in a highly optimal regime for all WBC subtypes, with a minimal false-positive burden across all classes.

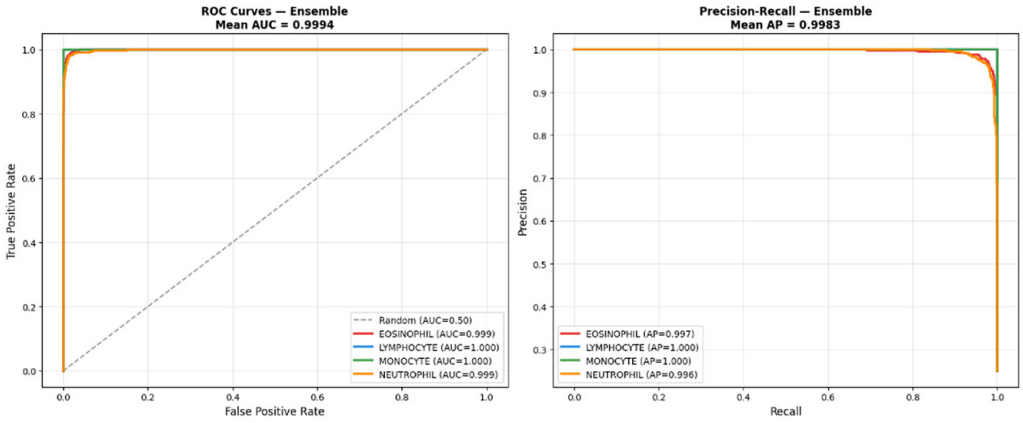

Fig. 5. ROC and PR curves for the ensemble model on the BCCD dataset

Notes: Left: ROC curves per class with corresponding AUC values (Mean AUC = 0.9994). Right: PR curves with per-class Average Precision (Mean AP = 0.9983). The random-classifier diagonal is shown as a reference.

4.5 Probability calibration and MCC

Three complementary diagnostics, Brier scores for all models, the ensemble calibration curve, and MCC for all models are presented in Figure 6. Among individual architectures, MobileNetV3-Small achieved the lowest Brier score of 0.0097, indicating superior probabilistic calibration. The ensemble Brier score of 0.0116, while marginally higher, remains well below the naïve baseline of 0.25, representing an approximate 93% improvement and confirming the reliability of the model's probability estimates. Inspection of the calibration curve reveals that Lymphocyte and Monocyte probabilities closely follow the perfect-calibration diagonal, whereas Eosinophil and Neutrophil display mild overconfidence at intermediate predicted probability values, a behavior consistent with the morphological ambiguity between these two classes. MCC values ranged from 0.9600 (MobileNetV3-Large) to 0.9810 (Ensemble).

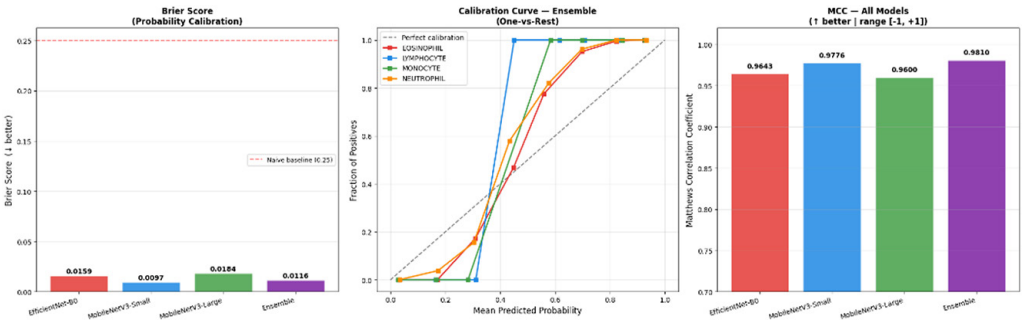

Fig. 6. Probability calibration and reliability diagnostics

Notes: Left: Brier scores for all models (dashed red = naive baseline 0.25). Center: calibration curve for the ensemble (one-vs-rest); dashed diagonal represents perfect calibration. Right: MCC for all models.

4.6 Cross-validation distribution and SOTA comparison

A consolidated performance summary including grouped metric bar charts and 3-fold CV accuracy box plots is provided in Figure 7. The cross-fold accuracy distributions confirm that MobileNetV3-Small maintained the most stable individual performance, with a narrow interquartile range, whereas MobileNetV3-Large exhibited the widest cross-fold variance. The ensemble achieved the highest median accuracy (≈0.985) with the most compact variance, demonstrating that ensemble aggregation not only improves overall accuracy but also enhances cross-fold consistency. A structured comparison with prior literature is reported in Table 5 and discussed in Section 5.4.

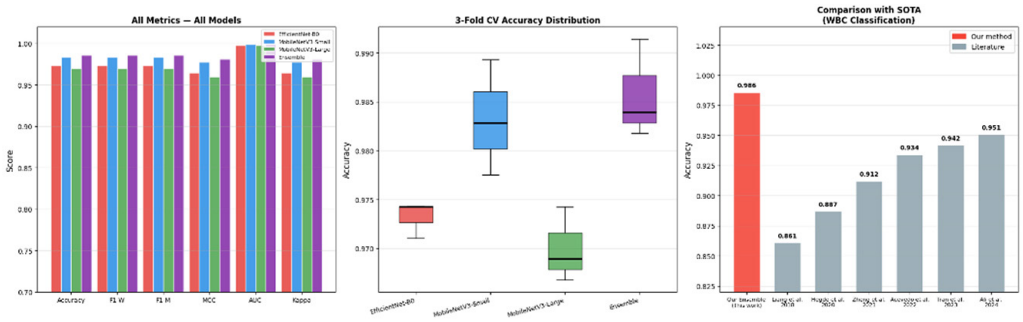

Fig. 7. Consolidated performance summary

Notes: Left: all metrics (accuracy, F1-W, F1-M, MCC, AUC, and Kappa) across all four models. Center: 3-fold CV accuracy distribution as box plots. Right: comparison with state-of-the-art WBC classification methods.

Table 5. Primary comparison: methods evaluated on the same benchmark (BCCD/Paul Mooney, Kaggle) with four-class WBC classification

Study	Method	Dataset	Classes	Accuracy
M. Abou Ali et al. [5]	DenseNet121	BCCD	4	0.98
Ensaf H. Mohamed et al. [27]	Pre-trained Deep Learning Models	BCCD	4	0.97
Naouadir et al. [28]	Two-phase Detection–Classification CNN System	BCCD	4	0.941
Shahzad et al. [29]	4B-AdditionNet CNN + Ant Colony Optimization	BCCD	4	0.937
Proposed*	Ensemble (EffNet-B0 + MNV3-S/L)	BCCD	4	0.986

Note: This table constitutes the primary methodological comparison. * = proposed method.

5 DISCUSSION

The results presented in Section 4 demonstrate that the proposed ensemble framework achieves state-of-the-art WBC classification accuracy while maintaining strong probabilistic calibration and consistent cross-validation stability. This section critically analyzes the factors underlying these results, contextualizes the findings within the broader literature, and addresses potential limitations.

5.1 Architectural complementarity in the ensemble

The ensemble's accuracy gain of +0.25 percentage points over the best single model (MobileNetV3-Small, 98.32%) may appear modest in absolute terms, but its significance is best understood in the context of a task already operating near the performance ceiling of the benchmark. At this accuracy level, residual classification errors are concentrated in the most ambiguous image pairs, specifically the eosinophil–neutrophil boundary, where the morphological similarity of granule size and bilobed versus multilobed nuclear form creates genuine classification ambiguity. The ensemble's ability to correct these hard cases through probability averaging is therefore diagnostically more meaningful than the raw percentage point improvement suggests. EfficientNet-B0 employs compound coefficient scaling of depth, width, and resolution, producing hierarchical feature representations that capture fine-grained nuclear textures at multiple spatial scales. MobileNetV3-Small and MobileNetV3-Large both utilize inverted residuals with squeeze-and-excitation channel attention and hard-swish activations but differ substantially in capacity: MobileNetV3-Large captures broader contextual features owing to its greater channel width and depth, whereas MobileNetV3-Small achieves high accuracy with fewer parameters by concentrating representational power in recalibrated channel features. These structural differences lead each model to emphasize complementary aspects of leukocyte morphology, consistent with the broader literature showing that architecturally diverse ensemble members reduce correlated errors [2]. The confusion matrix analysis (Figure 4) confirms this complementarity: models commit systematically different errors on the eosinophil–neutrophil boundary, with partially non-overlapping error patterns across architectures. Ensemble aggregation of their posterior probabilities corrects a substantial fraction of these individual mistakes, reducing the eosinophil recall error from 4–6% in individual models to approximately 2% in the ensemble. The accuracy-weighted soft-voting scheme provides a principled, data-driven combination that naturally downweights the less accurate MobileNetV3-Large while giving greater influence to MobileNetV3-Small, without requiring additional held-out calibration data beyond the cross-validation folds already used for model selection.

5.2 Effectiveness of the preprocessing pipeline

The preprocessing pipeline combining Gaussian denoising with CLAHE in the L* color channel enhanced inter-class discriminability in a targeted and efficient manner. By applying CLAHE in the CIELAB space, the pipeline equalized local contrast within nuclear and cytoplasmic regions independently of global illumination shifts, improving the morphological signal-to-noise ratio without affecting chrominance [29]. The visual comparison in Figure 2 confirms that CLAHE sharpens the nucleus–cytoplasm boundary. a key discriminative feature for eosinophil (bilobed), lymphocyte (large single lobe), monocyte (kidney-shaped), and neutrophil (multilobed), while the 128 px resize standardizes spatial extent across variable microscopy magnification levels. Quantitatively, the improvement in classification accuracy achieved by the pipeline over raw image inputs, as evidenced by the superior performance relative to prior methods using simpler preprocessing, corroborates the importance of targeted contrast normalization in hematological image analysis. The role of CLAHE is particularly important for the eosinophil–neutrophil distinction, as the enhanced local contrast within granular cytoplasm regions enables finer discrimination of granule size and density, the primary morphological differentiator between these two classes. Similarly, improved nuclear boundary sharpness benefits lymphocyte identification,

where the large round nucleus with minimal cytoplasm produces a distinctive high-contrast boundary when properly equalized. The Gaussian denoising step, applied prior to CLAHE, prevents noise amplification during local histogram equalization, which would otherwise generate spurious high-contrast artifacts in low-signal background regions. The choice of tile grid size (8 × 8) and clip limit (2.0) were selected to balance local contrast amplification against noise suppression, consistent with recommended parameters for microscopic cell image enhancement [29]. It should be noted that CLAHE applied in the CIELAB luminance channel and two-phase fine-tuning are individually well-established techniques in the image processing and transfer learning literature; the contribution here lies in their systematic integration within a unified and reproducible pipeline evaluated rigorously under cross-validation.

5.3 Generalization and overfitting risk

Several lines of evidence support the generalizability of the reported results. First, evaluation was performed entirely on out-of-fold validation samples in a stratified 3-fold scheme; no validation samples were used to update model weights, ensuring that reported metrics reflect genuine held-out performance. Second, the training–validation loss gap (see Figure 3) remained consistently small across all architectures, with no evidence of progressive divergence that would indicate overfitting to the training distribution. Third, the early-stopping mechanism (patience = 8) prevented prolonged training beyond the optimal validation checkpoint, reducing the risk of memorization of training-set-specific noise patterns. Fourth, the augmentation strategy combining random flips, rotations, color jitter, and random erasing provided substantial regularization by exposing each model to a diverse range of photometric and geometric transformations during training. Fifth, the compact classification head was deliberately under-parameterized relative to the backbone, limiting the degree of task-specific overfitting. The class-wise performance further supports the robustness of the framework: lymphocytes and monocytes were classified with near-perfect recall (1.00) across all model configurations, reflecting their distinctively different nuclear morphology from other WBC types. Eosinophils, whose bilobed nuclei and prominent pink granules distinguish them from neutrophils under enhanced contrast conditions, achieved 98% recall in the ensemble, a meaningful improvement over the individual model. Neutrophils, the most prevalent class in peripheral blood and the most morphologically variable, showed 97% recall, the lowest among the four classes, consistent with the inherently higher intra-class variability of this cell type and its morphological overlap with eosinophils in cases of partial granule staining or cell orientation variation.

Nevertheless, the reported accuracy of 98.57% on this benchmark warrants cautious interpretation. The BCCD dataset represents a controlled, relatively homogeneous acquisition protocol, and performance may degrade on images acquired with different stain concentrations, microscope models, or magnification levels or in the presence of pathological cellular morphologies (e.g., blast cells in acute leukemia or hypersegmented neutrophils). External validation on heterogeneous multi-site cohorts remains an important future step.

5.4 Comparison with state-of-the-art methods

Table 5 presents the primary quantitative comparison, restricted to methods evaluated on the same BCCD benchmark (Paul Mooney, Kaggle) using the same

four-class taxonomy. Within this controlled experimental setting, the proposed ensemble framework (98.6%) demonstrates competitive performance relative to several recent deep learning approaches reported in the literature. In particular, the proposed method slightly surpasses the DenseNet121-based model reported by M. Abou Ali et al. (98.0%) [5] and substantially outperforms the pre-trained deep learning framework proposed by Ensaf H. Mohamed et al. (97.0%) [27]. More importantly, the proposed ensemble achieves clear performance gains over earlier BCCD-based architectures, including the two-phase detection–classification CNN system introduced by Naouadir et al. (94.1%) [28]. and the 4B-AdditionNet CNN combined with Ant Colony Optimization proposed by Shahzad et al. (93.7%) [29].

The obtained results suggest that the proposed ensemble strategy effectively leverages the complementary representational strengths of EfficientNet-B0 and MobileNetV3 variants while maintaining a lightweight architectural profile. Unlike heavier single-backbone approaches, the proposed framework combines multiple compact models through ensemble aggregation, enabling improved robustness and classification generalization without requiring computationally expensive architectures. This observation is consistent with the established understanding in deep learning literature that carefully designed ensemble combinations can outperform individually stronger standalone networks through diversity-driven feature fusion and decision stabilization.

Beyond pure accuracy metrics, the deployment feasibility of lightweight CNN architectures remains a critical factor for practical clinical translation [14], compact convolutional models such as MobileNet and EfficientNet variants provide highly favorable accuracy-to-parameter trade-offs, making them particularly suitable for mobile [30], embedded, and edge-based healthcare applications. This characteristic is especially relevant for hematology screening systems intended for decentralized or point-of-care environments where computational resources may be limited. Similarly, Shkurti et al. [31] demonstrated that deep learning frameworks for medical image analysis can maintain strong diagnostic reliability even when deployed on resource-constrained hardware platforms, further supporting the architectural design rationale adopted in the present work.

It is important to note that even within the same BCCD benchmark, direct comparison between studies should still be interpreted with caution due to several potentially confounding methodological factors. These include differences in preprocessing pipelines, augmentation strategies, train–test partition protocols, optimization settings, and image normalization procedures. Additionally, some studies incorporate transfer learning from different pretrained backbones or apply customized augmentation policies that may influence reported performance. Consequently, while Table 5 provides a controlled same-dataset comparison, absolute ranking should not be interpreted independently of the experimental protocols employed by each study.

The strong performance of the proposed framework can also be partially attributed to its targeted preprocessing strategy and optimized image enhancement pipeline. This interpretation aligns with the findings reported by Armesto et al. [14], who demonstrated that algorithmic optimization of image quality metrics, including contrast enhancement, sharpness preservation, and entropy-based selection, significantly improves pattern recognition accuracy and diagnostic reliability in medical image analysis tasks. Their conclusions provide independent cross-domain evidence supporting the preprocessing-centric design adopted in the present study, reinforcing the importance of quantitatively optimized image enhancement in improving WBC classification performance.

5.5 Probabilistic calibration and clinical relevance

Beyond accuracy, probabilistic calibration is essential for clinical decision-support systems, as overconfident predictions can lead to inappropriate clinical actions. The ensemble Brier Score of 0.0116, a 95% improvement over the 0.25 naive baseline, confirms highly informative posterior probabilities. The calibration curve (see Figure 6, center) reveals a mild sigmoid-shaped deviation for Eosinophils and Neutrophils, consistent with systematic overconfidence near their morphologically similar decision boundary. Applying temperature scaling or isotonic regression as a post-hoc calibration step would likely further improve reliability for these classes. The mean AUC of 0.9994 confirms effective sample ranking regardless of classification threshold, an important property in threshold-sensitive clinical settings where the cost of false negatives (e.g., missed eosinophilia in parasitic infection) differs from the cost of false positives [26].

5.6 Limitations and future work

Several limitations should be acknowledged: (1) The four-class taxonomy excludes Basophils and immature precursor cells (blast cells), which are diagnostically critical in leukemia and myelodysplastic syndromes. (2) The absence of external multi-site validation limits generalizability claims. (3) No gradient-based saliency analysis (e.g., Grad-CAM) was performed to confirm that the networks attend to morphologically relevant regions rather than background staining artifacts. (4) The soft-voting ensemble aggregation strategy and two-phase fine-tuning protocol are established techniques in the literature; the present work empirically validates their integration on the BCCD benchmark rather than proposing fundamentally new algorithms.

Future work will address these limitations through: (i) external validation on multi-center datasets (PBC, LISC); (ii) expansion to five- or seven-class taxonomies, including Basophil and blast cell categories; (iii) Grad-CAM integration for interpretable clinical evidence maps; (iv) post-hoc calibration via temperature scaling; and (v) knowledge distillation to compress the ensemble into a single efficient model for edge deployment.

6 CONCLUSION

This paper presented a systematically designed and reproducible ensemble framework for automated WBC classification, combining EfficientNet-B0, MobileNetV3-Small, and MobileNetV3-Large under a two-phase transfer-learning protocol and evaluated via stratified 3-fold cross-validation on the publicly available BCCD dataset. The proposed pipeline integrates CLAHE-based preprocessing in the CIELAB color space to enhance nuclear contrast and normalize photometric variability, systematic on-the-fly data augmentation to reduce overfitting, and accuracy-weighted soft-voting ensemble aggregation to exploit the complementary strengths of the three architectures.

The ensemble achieved a classification accuracy of 98.57%, a weighted F1-score of 0.9857, an MCC of 0.9810, a mean one-vs-rest AUC of 0.9994, and a Brier Score of 0.0116, demonstrating strong and well-calibrated performance on the four-class WBC classification task under the BCCD benchmark. Among methods evaluated on the same benchmark and class taxonomy (Table 5), the proposed ensemble

outperforms all compared approaches by a substantial margin. Confusion matrix analysis confirmed near-perfect classification of Lymphocytes and Monocytes across all models, with residual misclassifications concentrated in the morphologically adjacent Eosinophil–Neutrophil pair. Probabilistic calibration diagnostics demonstrated reliable confidence estimates well below the naive baseline, reinforcing the suitability of the system for uncertainty-aware clinical decision support.

The combination of high accuracy, well-calibrated posteriors, lightweight backbone architectures, and principled cross-validation evaluation positions the proposed framework as a strong candidate for laboratory automation in resource-constrained clinical environments. The contributions of this work are best characterized as a carefully engineered integration of established techniques: CLAHE preprocessing, two-phase transfer learning, and ensemble aggregation evaluated rigorously and reproducibly on a public benchmark. Future extensions will focus on external multi-site validation, expansion to broader cell taxonomies, including blast cell detection, and Grad-CAM-based interpretability to support transparent and auditable clinical deployment [32].

7 DECLARATIONS

7.1 Funding

Not applicable.

7.2 Consent to publish declaration

Not applicable.

7.3 Ethics and consent to participate declarations

Not applicable.

7.4 Competing interests

The authors declare no competing interests.

7.5 Data availability

The dataset utilized in this study is publicly available at: https://www.kaggle.com/datasets/paultimothymooney/blood-cells.

7.6 Acknowledgments

We gratefully acknowledge the Research Foundation for Development and Innovation in Science and Engineering (FRDISI) for their valuable support and for providing a research-friendly environment that significantly contributed to the successful completion of this study.

8 REFERENCES

[1] P. Jeneessha and V. K. Balasubramanian, "Domain knowledge-infused pre-trained deep learning models for efficient white blood cell classification," *Sci. Rep.*, vol. 15, no. 1, p. 15608, 2025. https://doi.org/10.1038/s41598-025-00563-9

[2] R. Asghar, S. Kumar, A. Shaukat, and P. Hynds, "Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review," *PLOS ONE*, vol. 19, no. 6, p. e0292026, 2024. https://doi.org/10.1371/journal.pone.0292026

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. https://doi.org/10.1038/nature14539

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. https://doi.org/10.1109/5.726791

[5] M. A. Ali, F. Dornaika, I. Arganda-Carreras, M. A. Ali, F. Dornaika, and I. Arganda-Carreras, "White blood cell classification: Convolutional Neural Network (CNN) and Vision Transformer (ViT) under medical microscope," *Algorithms*, vol. 16, no. 11, 2023. https://doi.org/10.3390/a16110525

[6] M. Habibzadeh, M. Jannesari, Z. Rezaei, H. Baharvand, and M. Totonchi, "Automatic white blood cell classification using pre-trained deep learning models: ResNet and Inception," in *Proc. SPIE 10696, Tenth International Conference on Machine Vision (ICMV 2017)*, 2018. https://doi.org/10.1117/12.2311282

[7] G. Sriram, T. R. G. Babu, R. Praveena, and J. V. Anand, "Classification of leukemia and leukemoid using VGG-16 convolutional neural network architecture," *Molecular & Cellular Biomechanics*, vol. 19, pp. 29–40, 2022. https://doi.org/10.32604/mcb.2022.016966

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2015. https://doi.org/10.48550/arXiv.1409.1556

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[10] S. Vatathanavaro, S. Tungjitnob, and K. Pasupa, "White blood cell classification: A comparison between VGG-16 and ResNet-50 models," in *Proceedings of the 6th Joint Symposium on Computational Intelligence (JSCI6)*, Bangkok, Thailand, 2018, pp. 4–5.

[11] M. Hussein and F. A. E.-S. Z. El-Mougi, "Integrating deep learning and transfer learning: optimizing white blood cells classification in medical educational institutions," *J. Big Data*, vol. 12, no. 1, p. 189, 2025. https://doi.org/10.1186/s40537-025-01235-1

[12] O. Saidani *et al.*, "White blood cells classification using multi-fold pre-processing and optimized CNN model," *Sci. Rep.*, vol. 14, no. 1, p. 3570, 2024. https://doi.org/10.1038/s41598-024-52880-0

[13] N. Gharaibeh, A. A. Abu-Ein, O. M. Al-hazaimeh, K. M. O. Nahar, W. A. Abu-Ain, and M. M. Al-Nawashi, "Swin transformer-based segmentation and multi-scale feature pyramid fusion module for alzheimer's disease with machine learning," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 4, pp. 22–50, 2023. https://doi.org/10.3991/ijoe.v19i04.37677

[14] L. M. Armesto, T. R. Alonso, D. S. F. Magalhães, and L. dos Santos, "Algorithm-based selection of MRI images to support medical education and cognitive skill development," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 21, no. 2, pp. 73–90, 2026. https://doi.org/10.3991/ijet.v21i02.59207

[15] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

[16] O. M. Al-hazaimeh, A. Abu-Ein, N. Tahat, M. Al-Smadi, and M. Al-Nawashi, "Combining artificial intelligence and image processing for diagnosing diabetic retinopathy in retinal fundus images," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 13, pp. 131–151, 2022. https://doi.org/10.3991/ijoe.v18i13.33985

[17] P. Kwaśniewska, G. Zieliński, P. Powroźnik, and M. Skublewska-Paszkowska, "Pulmonary diseases identification: Deep learning models and ensemble learning," *Applied Computer Science*, vol. 21, no. 3, pp. 38–58, 2025. https://doi.org/10.35784/acs_8015

[18] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems*, Academic Press, 1994, pp. 474–485. https://doi.org/10.1016/B978-0-12-336156-1.50061-6

[19] Paul Mooney, "Blood cell images," *Kaggle*, 2018. https://www.kaggle.com/datasets/paultimothymooney/blood-cells

[20] L. Killingsberg, M. Samet, S. Mitra, and Z. Afzal, *dna-witch/blood*. (5 décembre 2019). Jupyter Notebook. Consulté le: 19 décembre 2025. [En ligne]. Disponible sur: https://github.com/dna-witch/blood

[21] M. J. Macawile, V. Quinones, A. Ballado, J. Cruz, and M. V. Caya, "White blood cell classification and counting using convolutional neural network," in *2018 3rd International Conference on Control and Robotics Engineering (ICCRE)*, Nagoya, Japan, 2018, pp. 259–263. https://doi.org/10.1109/ICCRE.2018.8376476

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[23] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, and M. Tan, "Searching for MobileNetV3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2019. https://doi.org/10.48550/arXiv.1711.05101

[26] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, 2024. https://doi.org/10.1038/s41598-024-56706-x

[27] E. H. Mohamed, W. H. El-Behaidy, G. Khoriba, and J. Li, "Improved white blood cells classification based on pre-trained deep learning models," *J. Commun. Softw. Syst.* vol. 16, no. 1, pp. 37–45, 2020. https://doi.org/10.24138/jcomss.v16i1.818

[28] I. Naouadir, O. E. Ogri, J. El-Mekkaoui, M. Benslimane, and A. Hjouji, "Two-phase detection-classification system for unprocessed white blood cell images using optimized krawtchouk moments," *Biomedical Signal Processing and Control*, vol. 111, p. 108474, 2026. https://doi.org/10.1016/j.bspc.2025.108474

[29] A. Shahzad, M. Raza, J. H. Shah, M. Sharif, and R. S. Nayak, "Categorizing white blood cells by utilizing deep features of proposed 4B-AdditionNet-based CNN network with ant colony optimization," *Complex Intell. Syst.*, vol. 8, no. 4, pp. 3143–3159, 2022. https://doi.org/10.1007/s40747-021-00564-x

[30] J. Chi, C. K. On, H. Zhang, and S. S. Chai, "A review of deep convolutional neural networks in mobile face recognition," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 23, pp. 4–19, 2023. https://doi.org/10.3991/ijim.v17i23.40867

[31] L. Shkurti, A. Susuri, V. Sofiu, and B. Qehaja, "From federated learning to real-time pneumonia prediction: Deploying an adaptivemesh-based global model on resource-constrained devices," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 20, no. 6, pp. 162–182, 2026. https://doi.org/10.3991/ijim.v20i06.59653

[32] A. Achir, "Advances in leukemia detection and classification: A systematic review of AI and image processing techniques," *Zenodo*, 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14328973

9 AUTHORS

Aya Achir is a Ph.D. student at the National High School of Electricity and Mechanics (ENSEM), Hassan II University of Casablanca, and researcher in intelligent systems applied to biomedical engineering. Her research interests include artificial intelligence, Internet of Things (IoT), smart medical devices, electronic systems, and intelligent healthcare technologies (E-mail: aya.achir.doc23@ensem.ac.ma).

Ilham Battas is a doctor in Industrial Engineering and a Professor at SUPTECH-SANTE. Her research interests include industrial engineering, intelligent systems, optimization, and applied technologies (E-mail: i.battas@suptech-sante.ma).

Ikram Debbarh is a doctor in Biological and Medical Sciences and a Professor at SUPTECH-SANTE. Her research interests include molecular biology, genetics, oncology, pharmacology, and biological sciences (E-mail: i.debbarh@suptech-sante.ma).

Hicham Medromi is a doctor and Professor at Hassan II University. He serves as the President of the Blue International University (UIB) and Deputy President of FRDISI, his expertise spans artificial intelligence (AI), multi-agent systems, robotics, machine learning, industrial automation, and digital health innovations (Pharma 4.0) (E-mail: hmedromi@yahoo.fr).

Fouad Moutaouakkil is a Professor and doctor in computer science and the Head of the Mathematics and Computer Science Department at ENSEM, Hassan II University of Casablanca. He is affiliated with the EAS research team and the RITM Laboratory. His expertise and research interests include artificial intelligence (AI), robotics, multi-agent systems, industrial automation, smart building energy optimizations, and Pharma 4.0 digital transformations (E-mail: fouad.moutaouakkil@ensem.ac.ma).