Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TinyNCFormer: A Tensor-Decomposed Dual-Stream Transformer with Diffusion-Guided Ensemble Knowledge Distillation for White Blood Cell Classification and Instance Segmentation

Enas Fadil Jalil Al-Okili
Arak University
Majid Sepahvand
Arak University
Seyfollah Soleimani
s-soleimani@araku.ac.ir

Arak University

Article

Keywords: White blood cell classification, Image segmentation, Knowledge distillation, Tensor decomposition, Transformers, Ensemble learning

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11007713/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TinyNCFormer: A Tensor-Decomposed Dual-Stream Transformer with Diffusion-Guided Ensemble Knowledge Distillation for White Blood Cell Classification and Instance Segmentation

Enas Fadil Jalil Al-Okili[1], Majid Sepahvand[1] and Seyfollah Soleimani*[1]

[1] *Department of Computer Engineering, Faculty of Engineering, Arak University, Markazi, Iran*

Abstract

White blood cell (WBC) classification from peripheral blood smear images is central to diagnosing leukemia, infection, and other hematologic disorders, and remains one of the most frequently requested laboratory analyses in clinical practice. Automated classification is difficult because WBC subtypes are morphologically similar, clinical microscopy fields routinely contain multiple overlapping cells, and large, accurate models are impractical to deploy on point-of-care hardware. This trade-off is addressed with a knowledge-distillation framework in which a single teacher, formed by fusing five diverse pretrained backbones over a nuclear-cytoplasmic, stain-separated representation, supervises a compact dual-stream Transformer student, TinyNCFormer, through confidence-calibrated soft-target distillation, diffusion-based feature denoising, and an instance-segmentation head for multi-cell scenes. Evaluated on our newly introduced RV-PBS dataset and three additional blood cell corpora (CRV-PBS, PBC, and C-PBC), the resulting 16.5-million-parameter student exceeds every constituent backbone, reaching classification F1-scores of 96.32–98.32% and mean average precisions of 0.865–0.883 for instance segmentation across the four datasets, outperforming prior baselines by an average of 8.7% (range 2.7–15.3%) in relative mAP. These results show that diverse ensemble supervision with uncertainty- and denoising-aware distillation yields a compact model more accurate than its teacher and substantially more efficient, a practical candidate for point-of-care hematological screening.

*Corresponding author: Seyfollah Soleimani (s-soleimani@araku.ac.ir)

Keywords

White blood cell classification; Image segmentation; Knowledge distillation; Tensor decomposition; Transformers; Ensemble learning

1 Introduction

White blood cell (WBC) differential counting is among the most routinely ordered laboratory tests in clinical practice, providing early, low-cost evidence for infection, inflammation, autoimmune disease, and hematologic malignancies such as leukemia [1, 2]. The scale of this clinical need is substantial and growing: global leukemia incidence alone rose from an estimated 311,648 new cases in 1990 to 461,423 in 2021, and is projected to reach 519,540 cases by 2050 as populations age and grow [3], with the complete blood count and differential remaining a first-line screening step for essentially all of these cases. Because the relative proportions and morphological state of the different WBC subtypes carry direct diagnostic and prognostic value, accurate and reproducible classification of cells observed under the microscope is essential to timely clinical decision-making. Traditionally, this task has relied on manual examination of Romanowsky-stained peripheral blood smears by trained hematologists or laboratory technicians, a process that is time-consuming, subject to inter-observer variability, and increasingly difficult to scale in high-throughput or resource-limited settings, a limitation also well documented for manual light-microscopy screening of other blood-borne conditions such as malaria [4]. Automated, image-based WBC classification therefore has the potential to reduce diagnostic turnaround time, standardize reporting, and extend access to hematological screening beyond specialized laboratories.

Despite this potential, automated WBC classification from microscopic images remains a genuinely difficult computer vision problem. A large part of this difficulty stems from the gap between curated single-cell crops, on which most published methods are trained and evaluated, and the raw microscopic field of view actually produced by a slide scanner or digital microscope, which typically contains multiple cells at varying densities, orientations, and degrees of mutual overlap (Fig. 1). When only one, clearly isolated cell is present, its boundary, nucleus, and cytoplasm are easy to segment and describe, and classification reduces to a comparatively simple pattern-recognition task [5, 6]. When several cells are clustered together, however, occluded boundaries, overlapping cytoplasmic regions, and inconsistent staining across the field of view can substantially degrade both the segmentation and the classification steps, and errors introduced at this stage propagate directly into the reported differential count.

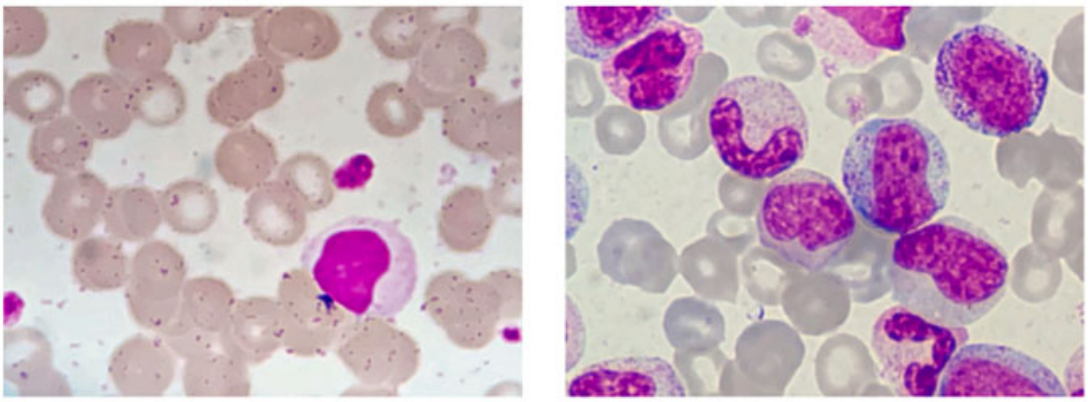


Figure 1: Comparison of blood smear images: (left) a single white blood cell; (right) multiple white blood cells in the field of view.

A second, largely orthogonal source of difficulty is the fine-grained morphological similarity among WBC subtypes, particularly across adjacent stages of the same maturation lineage, spanning the ten categories considered in this work: Band Cells, Basophils, Blast Cells, Eosinophils, Lymphocytes, Myelocytes, Metamyelocytes, Monocytes, Neutrophils, and Promyelocytes. While morphologically distant classes (e.g. a Lymphocyte versus a Basophil) are readily separable by nuclear size, shape, and staining intensity alone, distinguishing between successive maturation stages of the granulocytic lineage (Myelocyte $\rightarrow$ Metamyelocyte $\rightarrow$ Band Cell) requires attending to subtle, continuous changes in nuclear lobulation and chromatin condensation that a coarse, single-scale feature extractor may fail to capture [7, 8].

Convolutional neural networks (CNNs) pretrained on large natural-image corpora have become the de facto backbone for WBC classification, owing to their strong transferable representations and comparatively modest labeled-data requirements relative to training from scratch. Yet a single such backbone reflects only one architectural inductive bias, and empirically, no individual CNN family dominates across all WBC subtypes and imaging conditions: architectures optimized for efficiency (e.g. depthwise-separable or channel-shuffle designs) can trade away exactly the fine-grained spatial sensitivity needed to separate adjacent maturation stages, while larger, higher-capacity backbones that better capture this structure are correspondingly too computationally expensive for point-of-care deployment on portable digital microscopes or embedded slide scanners [9, 10]. Ensembling multiple heterogeneous backbones can partially close this accuracy gap, but naively deploying an ensemble multiplies inference cost and memory footprint by the number of constituent models, which is impractical outside a well-resourced laboratory setting.

Knowledge distillation (KD) offers a principled way to reconcile these competing demands: a large, high-capacity teacher (or ensemble of teachers) is trained without deployment constraints, and its knowledge is subsequently transferred to a small, efficient student model

that is actually deployed. Conventional KD, however, applies a single fixed temperature to soften the teacher's output distribution for every training sample, implicitly assuming that all samples are equally difficult for the teacher; this is a poor match for WBC classification, where the teacher is highly confident on textbook-morphology cells but considerably less so on borderline maturation stages or occluded cells. Most KD formulations that additionally align intermediate feature representations do so via direct regression between the student's and teacher's features, which forces a compact student to imitate structure – including noise-like, over-parameterized artifacts specific to the teacher's backbones – that it may lack the capacity to represent faithfully [11, 12].

The timing of this trade-off is particularly consequential. Smartphone- and tablet-based digital microscopy attachments have recently made image capture of peripheral blood smears feasible outside specialized hematology laboratories, while edge-AI accelerators now make on-device inference plausible for the first time at price points suited to resource-limited clinics. Realizing this opportunity requires models that are simultaneously accurate across fine-grained subtypes, robust to multi-cell fields of view, and small enough to run on such hardware, a combination of requirements that remains largely unaddressed by existing WBC classification systems (Section 2).

In this work, we propose a knowledge-distillation framework for joint WBC classification and instance segmentation that is designed around these two observations. The teacher is constructed as an ensemble fusion classifier that aggregates five architecturally diverse, ImageNet-pretrained CNN backbones, trained on the raw RGB single-cell crop, so that no single inductive bias is solely responsible for capturing the fine-grained cues that separate visually similar subtypes. Rather than feeding the raw crop directly to the compact student, we first decompose it via stain color deconvolution into a nuclear (hematoxylin) channel and a cytoplasmic (eosin) channel, yielding a compact, native-resolution representation that mirrors the biological basis on which hematologists themselves distinguish WBC subtypes. This representation is consumed by *TinyNCFormer*, a dual-stream Transformer student built on the self-attention mechanism [14, 15] in which every weight tensor is parameterized from the outset in tensor-decomposed (HORSVD) form, so that compactness is a property of the architecture itself rather than an afterthought imposed by post-hoc compression. Knowledge is transferred from the ensemble teacher to TinyNCFormer through two complementary, jointly stabilized mechanisms: an entropy-adaptive soft-target loss that automatically softens supervision in proportion to the teacher's own predictive uncertainty on each sample, and a diffusion-guided feature-denoising loss that routes the student's feature representation through a reverse diffusion process trained exclusively on teacher features, thereby filtering out teacher-specific structure before feature alignment rather than forcing the student to

imitate it directly. A lightweight instance-segmentation head is attached to the same dual-stream feature map to additionally localize each cell within multi-cell fields of view, so that a single compact model performs both localization and fine-grained classification. We train the framework on our own novel RV-PBS dataset and evaluate its cross-dataset generalization on three further peripheral blood cell corpora (CRV-PBS, PBC, and C-PBC).

The main contributions of this work are as follows.

1. We introduce a five-backbone ensemble fusion teacher for WBC classification that combines architecturally diverse, ImageNet-pretrained CNNs operating on a nuclear–cytoplasmic stain-separated representation of each single-cell crop.

2. We design TinyNCFormer, a compact dual-stream Transformer whose weights are parameterized from the outset in tensor-decomposed (HORSVD) form, yielding a substantial reduction in parameters and inference cost relative to the teacher ensemble, and extend it with a lightweight instance-segmentation head so that a single compact model jointly localizes and classifies individual WBCs in multi-cell fields of view.

3. We propose an entropy-adaptive soft-target distillation loss (ATKD) that calibrates the distillation temperature to the teacher's per-sample predictive uncertainty, rather than applying a single global temperature across all training samples, coupled with a diffusion-guided feature-denoising distillation mechanism that filters teacher-specific structure from the student's feature representation prior to alignment.

4. We introduce RV-PBS, a novel 10-class peripheral blood smear dataset with multi-cell, instance-level annotations, and evaluate the proposed framework's cross-dataset generalization on three further corpora (CRV-PBS, PBC, and C-PBC), reporting both classification/segmentation performance and the efficiency gains of TinyNCFormer relative to the ensemble teacher.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed teacher–student framework, including the stain-separated input representation, the ensemble teacher, the tensor-decomposed TinyNCFormer student, and the two distillation mechanisms. Section 4 describes the experimental setup and datasets, Section 5 reports classification, segmentation, cross-dataset, and ablation results, Section 6 discusses the broader implications and limitations of these findings, and Section 7 concludes the paper.

2 Related Work

2.1 Clinical and Imaging Background

White blood cells arise from hematopoietic precursors and circulate within a well-characterized lymphoid and myeloid lineage system [1]; disruption of this maturation process, together with abnormal circulating counts, underlies many of the hematologic diseases that peripheral blood film examination is used to screen for [2]. Manual microscopic examination remains the diagnostic reference standard not only for WBC differentials but for other blood-borne conditions, where light microscopy is still benchmarked against rapid diagnostic tests in field settings [4]; expert manual review therefore remains both clinically valuable and labor-intensive at scale. Even non-nucleated blood components exhibit imaging-relevant structural regularities that motivate careful, disease-aware preprocessing of microscopic blood images [16], and assembling reliable training data for automated systems in turn requires careful curation and expert annotation infrastructure [17–19].

2.2 Classical and Segmentation-Based WBC Analysis

Early efforts toward automated WBC analysis relied on classical image-processing pipelines, in which nucleated cells were first isolated via marker-controlled watershed segmentation [20] or optimized thresholding strategies such as metaheuristically tuned Otsu- and Kapur-based schemes [21], after which handcrafted morphological, color, and texture descriptors were passed to a shallow classifier. Reviews of this pipeline consistently highlight its sensitivity to staining variability, illumination, and cell overlap [8, 22]. The advent of fully convolutional segmentation architectures such as U-Net [23] and its successors, including UNet++ [24], MultiResUNet [25], and attention-augmented variants [26, 27], substantially improved nucleus and cytoplasm delineation under overlap, while center-point regression networks have been shown to match segmentation-based differential counting without an explicit segmentation step [28], and self-supervised segmentation has been explored to reduce annotation cost for overlapping leukocytes [29]. Across this line of work, WBC delineation is treated as a generic, stain-agnostic pixel-labeling problem: intensity- or learned-feature-based segmentation is applied uniformly across the RGB image, without explicitly separating the two stain axes (hematoxylin-stained nucleus, eosin-stained cytoplasm) that hematologists themselves rely on to distinguish subtypes. This leaves open a representational gap that the nuclear–cytoplasmic stain-separated input introduced in Section 3.1 is designed to close.

2.3 Deep Convolutional Networks for WBC Classification

Once isolated, single-cell classification has been dominated by convolutional neural networks pretrained on large natural-image corpora such as ImageNet [30]. Architectures from the residual [31], multi-branch Inception-style [32], and depthwise-separable [9] design families have each been adapted to WBC recognition, with Acevedo et al. demonstrating strong CNN-based recognition of peripheral blood cell types [5] alongside an accompanying public dataset [33], and Wang et al. reporting a deep-learning pipeline for peripheral leukocyte recognition at scale [6]. Liu and Deng further showed that very deep CNNs remain trainable under the comparatively small sample sizes typical of medical imaging [34], while neural architecture search has been used to discover efficiency-optimized backbones suitable for downstream transfer [10], and such models are almost universally optimized with stochastic gradient methods [35], most commonly Adam [36]. These efforts are supported by purpose-built WBC datasets, including the segmented-nuclei/cytoplasm Raabin-WBC corpus [7] and the ALL-IDB leukemia image database [37]. Because clinically important subtypes are comparatively rare, class-imbalance-aware training has received sustained attention in the broader deep learning literature [38] and specifically for WBC and related hematological screening tasks [39, 40]. A consistent pattern emerges from this body of work: because no single backbone family dominates across all WBC subtypes and imaging conditions, each captures a partially complementary inductive bias, yet published WBC classifiers almost invariably deploy either one backbone or a full-sized ensemble at inference time – neither of which offers a principled mechanism for compressing the accuracy gained from architectural diversity into a single lightweight model. Closing this gap between ensemble accuracy and single-model deployability is the specific objective of the ensemble-fusion teacher and its distilled student described in Sections 3.2–3.3.

2.4 Detection Architectures and Multi-Cell Scene Analysis

When multiple, potentially overlapping cells are present in a single field of view, WBC recognition becomes a joint detection-and-classification problem rather than a single-label task. General-purpose object detectors, including the two-stage Faster R-CNN [41] and R-FCN [42] families, single-stage detectors such as YOLO [43] and SSD [44], and instance-segmentation frameworks like Mask R-CNN [45], provide the principal machinery for localizing individual cells prior to classification, and are typically pretrained on large detection benchmarks such as COCO [46]. Interpretability and saliency-analysis tools developed for general CNNs [47, 48] have subsequently been used to verify that such detectors attend to genuine morphological structure rather than staining artifacts, a concern that is particularly acute in cluttered,

multi-cell hematology images [49]. These detectors are, however, near-universally instantiated as dense, natural-image-pretrained backbones and trained as detection- or segmentation-only systems, decoupled from any dedicated fine-grained WBC classifier; deploying them for combined multi-cell localization and subtype classification therefore requires pairing two separately trained, comparatively heavy models rather than a single compact network. The lightweight instance-segmentation head attached directly to the tensor-decomposed TinyNCFormer backbone (Section 3.3.4) is designed to avoid this two-model overhead.

2.5 Domain Adaptation and Generalization

A closely related line of work addresses the fact that models trained on one smear-preparation protocol, scanner, or laboratory often generalize poorly to another. Adversarial feature-alignment methods [50] and pixel-level image-to-image translation [51] have been used to bridge such domain gaps without target-domain labels, and progressive adaptation strategies have extended this idea to detection [52]; Guan and Liu provide a comprehensive survey of domain adaptation specifically for medical image analysis [12], building on the broader transfer-learning foundations laid out by Pan and Yang [53]. Within WBC classification specifically, Pandey et al. proposed a target-independent domain adaptation scheme based on generative latent search to handle unlabeled target domains [11]. Each of these methods, however, requires an explicit domain-alignment objective – adversarial, pixel-level, or generative – computed jointly over source and target samples, which adds training complexity and, in several cases, still requires access to unlabeled target-domain images at training time. Knowledge distillation, the mechanism underlying our own teacher–student framework, shares the same underlying motivation as this body of work without this requirement: both aim to transfer a well-generalized representation, learned under favorable conditions, to a setting with tighter data, compute, or labeling constraints. In the present work, cross-dataset generalization is instead evaluated directly, as a consequence of the ensemble teacher's architectural diversity and the diffusion-guided denoising mechanism (Section 3.4.2), without introducing a dedicated domain-alignment objective (Section 5.7); explicitly incorporating one of the alignment mechanisms reviewed above is a natural direction for narrowing the residual cross-dataset gap reported in Section 5.7.

2.6 Attention-Based Architectures and Uncertainty-Aware Refinement

Since the introduction of the Transformer architecture [14] and its adaptation to vision via the Vision Transformer [15], attention-based backbones have increasingly been explored for

hematological image analysis, motivated by their ability to model long-range dependencies across a cell's nuclear and cytoplasmic structure that purely local convolutional receptive fields may miss. The ISBI 2026 WBCBench challenge on robust WBC classification is a direct response to this clinical need, evaluating macro-averaged performance under severe long-tailed class distributions and cross-domain shift [13]. Recent entries to this benchmark combine Swin Transformer ensembles with generative artifact restoration and biologically-inspired post-hoc morphological filtering to recover rare pathological subtypes [54], or pair decoupled, stain-normalized representation learning with multi-backbone, test-time-augmented ensembling of CNN and Transformer models [55]. In parallel, dedicated transformer architectures for blood cell classification, such as BCCT [56], and interpretable custom CNN pipelines with post-hoc heatmap visualization [57], have continued to appear throughout 2026, reflecting sustained interest in both accuracy and clinical trustworthiness. Uncertainty quantification and attention-guided refinement have likewise been explored broadly in medical image analysis, from holistic uncertainty-aware segmentation networks [58] and Bayesian-optimized multi-scale attention segmentation [59], to transfer-learning-based COVID-19 screening from chest radiographs [60]; the entropy-adaptive calibration used in our own soft-target distillation loss (Section 3.4.1) is motivated by the same observation underlying this literature – that a model's own predictive confidence is itself a useful, sample-specific training signal. Even these recent, strong-performing entries, however, report accuracy and robustness under class imbalance as their primary evaluation axis; parameter count, inference latency, and other deployment-efficiency metrics are seldom reported alongside accuracy, leaving the accuracy–efficiency trade-off for Transformer-based WBC classification comparatively unquantified.

Table 1 summarizes these paradigms along the dimensions most relevant to the present work: whether multi-cell fields of view are handled natively, whether the method is explicitly designed for compact deployment, and whether supervision is calibrated to per-sample uncertainty. No single prior paradigm combines all three properties, which are precisely the properties targeted by the ensemble-fusion teacher, entropy-adaptive and diffusion-guided distillation, and integrated segmentation head of the proposed framework (Section 3).

All of the deep architectures reviewed above, including the very recent Transformer- and ensemble-based systems evaluated on the WBCBench 2026 benchmark, are optimized primarily for classification accuracy rather than deployment efficiency, and, to our knowledge, none combines an architecturally diverse ensemble teacher with a tensor-decomposed, dual-stream Transformer student trained through uncertainty-calibrated and diffusion-denoised distillation. Closing this gap is the objective of the framework presented in Section 3.

Table 1: Qualitative comparison of representative prior WBC analysis paradigms against the proposed framework, along the dimensions defined above.

Paradigm	Representative work	Native multi-cell	Compact deployment	Uncertainty-calibrated
Classical segmentation + handcrafted features	[20, 21]	Partial (segm. only)	–	–
Fully convolutional segmentation (U-Net family)	[23–25]	Partial (segm. only)	–	–
Single/ensembled CNN classifiers	[5–7]	×	×	–
General-purpose detectors (YOLO, Mask R-CNN, etc.)	[41, 43, 45]	✓	×	–
Adversarial / translation-based domain adaptation	[11, 50, 51]	×	×	–
2026 Transformer / ensemble WBC classifiers	[54–56]	×	×	Partial
Proposed (ensemble-fusion teacher + TinyNCFormer)	Section 3	✓	✓	✓

3 Proposed Method

3.1 Nuclear–Cytoplasmic Stain-Separated Representation

Each training instance begins as a single-cell crop $X_{rgb} \in \mathbb{R}^{3\times H\times W}$ extracted from an RV-PBS peripheral blood smear image (Section 4.1), stained under the standard Romanowsky protocol so that nuclear chromatin stains predominantly basophilic (hematoxylin-like) and cytoplasmic granules stain predominantly eosinophilic. Rather than feeding the raw RGB crop directly to the student, we first decompose it into its constituent stain contributions via optical-density color deconvolution [61]: given the fixed stain matrix $M \in \mathbb{R}^{3\times 2}$ estimated for the Romanowsky stain pair, the RGB crop is converted to optical density $OD = -\log_{10}(X_{rgb}/255)$ and projected onto the two stain axes,

$$X_{HE} = \left[\, \Phi_H(X_{rgb}) \,\oplus\, \Phi_E(X_{rgb}) \,\right] = M^{+}OD \;\in\; \mathbb{R}^{2\times H\times W}, \tag{1}$$

where M^{+} is the pseudo-inverse of the stain matrix, Φ_H isolates the hematoxylin-like channel (nuclear chromatin density) and Φ_E the eosin-like channel (cytoplasmic granule density), and $\oplus$ denotes channel-wise stacking. Both channels are min–max normalized to $[0, 1]$ and resized to a fixed working resolution $H \times W$ (we use $H = W = 64$ in this work, consistent with the compact crop sizes standard for single-cell WBC classification; the exact value used in the reported experiments is given in Section 4). This two-channel tensor X_{HE} is the compact, native-resolution representation consumed directly by the student network (Section 3.3). The teacher ensemble, by contrast, operates on the original RGB crop X_{rgb} (Section 3.2), since X_{rgb} is already in the three-channel format expected natively by ImageNet-pretrained backbones and requires no channel-replication step.

3.2 Teacher Network (Ensemble Fusion Classifier)

Rather than relying on a single convolutional architecture, the teacher is constructed as an ensemble fusion classifier that aggregates complementary inductive biases from five convolutional neural network (CNN) backbones pre-trained on natural images. Each backbone brings a distinct efficiency–representation trade-off:

- **MobileNetV2** relies on inverted residual blocks with linear bottlenecks that expand and subsequently compress channel-wise features while limiting information loss.
- **EfficientNetV2-S** employs Fused-MBConv layers together with neural-architecture-search-guided scaling to balance parameter count against representational power.
- **ShuffleNetV2** introduces channel-shuffle operations that restore cross-channel information flow after grouped convolutions without additional computational cost.
- **Xception** decouples spatial and channel-wise feature learning through depthwise separable convolutions.
- **GoogLeNet** applies parallel convolutional filters of multiple kernel sizes within its Inception modules, enabling multi-scale pattern extraction.

Combining these architecturally diverse backbones is intended to reduce the risk that any single inductive bias systematically overlooks the subtle morphological cues (e.g. fine nuclear chromatin texture, chromatin lobulation, or cytoplasmic granule density) that distinguish visually similar WBC subtypes, most notably across adjacent maturation stages of the granulocytic lineage (e.g. myelocyte vs. metamyelocyte vs. band neutrophil).

Since the five backbones are natively designed for RGB natural images, each single-cell crop $X_{rgb} \in \mathbb{R}^{3\times H\times W}$ (Section 3.1) is first min–max normalized per channel and bilinearly resized to a fixed spatial resolution of 224×224 to match the expected input format, yielding $I \in \mathbb{R}^{3\times 224\times 224}$ per WBC cell image. This tensor is independently propagated through each backbone, and the output of its final global average pooling layer is retained as a compact descriptor, denoted $D_m, D_e, D_s, D_x, D_g \in \mathbb{R}^d$ for MobileNetV2, EfficientNetV2-S, ShuffleNetV2, Xception, and GoogLeNet, respectively. The five descriptors are concatenated into a unified representation $D_{cat} = [D_m \oplus D_e \oplus D_s \oplus D_x \oplus D_g]$, where $\oplus$ denotes feature concatenation, which is then passed through two fully connected layers to produce the teacher logits $Z_T \in \mathbb{R}^C$, with $C = 10$ WBC categories in this study (Table 3):

$$Z_T = W_2 \cdot \mathrm{Dropout}\big(\sigma(W_1 D_{cat} + b_1)\big) + b_2, \tag{2}$$

where $\sigma(\cdot)$ denotes the ReLU activation and W_1, b_1, W_2, b_2 are the weights and biases of the two fully connected layers. In addition to the classification logits Z_T, the concatenated representation D_{cat} is retained, as it constitutes the teacher feature representation $F^{(tea)}$ that will subsequently serve as the target for the diffusion-based denoising and distillation procedure described in Section 3.4.

3.3 Student Network: TinyNCFormer

While the ensemble teacher described in Section 3.2 is well suited to extracting a rich representation from the RV-PBS cohort despite its pronounced class imbalance across minority WBC subtypes (Table 3), its five-backbone structure is far too heavy for practical, low-latency point-of-care hematology screening (e.g. on-device analysis in a portable digital microscope or slide scanner). We therefore design *TinyNCFormer*, a compact dual-stream nuclear–cytoplasmic Transformer intended to absorb the teacher's knowledge through distillation (Section 3.4) while remaining small enough for deployment on resource-constrained hardware. Unlike a conventional compress-after-train pipeline, in which a dense backbone is first designed and only subsequently compressed, every learnable weight tensor of TinyNCFormer is parameterized *from the outset* in a Higher-Order Rank Singular Value Decomposition (HORSVD) form: the architecture and its tensor-decomposed parameterization are therefore two aspects of a single design, not two separate stages.

Fig. 2 provides an overview of the resulting architecture before we turn to its detailed formulation. The stain-separated representation X_{HE} (Section 3.1) is first passed through a local feature extraction stage of two HORSVD-decomposed convolutional layers, producing Z_{local}, whose channel dimension preserves the provenance of the two input stain channels

through grouped convolutions. This shared representation is then split along its channel-provenance axis and processed by two structurally identical but independently specialized branches: a *nuclear stream*, which models dependencies across spatial tokens derived from the hematoxylin (nuclear chromatin) channel, and a *cytoplasmic stream*, which models dependencies across spatial tokens derived from the eosin (cytoplasmic granule) channel. Within each stream, every linear projection of the self-attention block (Q, K, V, O) and the subsequent feed-forward network is instantiated directly in decomposed form, indicated in the figure by the small core-tensor-plus-factor-matrices icon attached to each layer, rather than being trained densely and compressed afterward. The dashed cross-stream links in the attention stage indicate that the two streams share their model- and head-dimension factor matrices while retaining independent core tensors, a coupling that is only meaningful because the decomposition is embedded in the layer definitions themselves. The two streams are subsequently reshaped back to the spatial domain, fused through a further decomposed 1×1 convolution, and combined with Z_{local} via a residual connection to produce the final student representation. We now formalize each of these components in turn, starting with the decomposition notation used throughout.

3.3.1 Decomposed Weight Notation

Let $\mathcal{W}$ denote a generic weight tensor of order L belonging to some layer of the network, either a 4-mode tensor $\mathcal{W} \in \mathbb{R}^{C_{out} \times C_{in} \times k_h \times k_w}$ for a convolutional kernel, or a 3-mode tensor $\mathcal{W} \in \mathbb{R}^{N_h \times d_{model} \times d_h}$ for a multi-head attention projection reshaped along the head dimension. Rather than storing $\mathcal{W}$ densely, TinyNCFormer represents it through a compact core tensor $\mathcal{Q}$ and L mode-wise factor matrices $\mathcal{R}^{(1)}, \ldots, \mathcal{R}^{(L)}$:

$$\mathcal{W} \;\cong\; \mathcal{Q} \times_1 \mathcal{R}^{(1)} \times_2 \mathcal{R}^{(2)} \cdots \times_L \mathcal{R}^{(L)} \;=\; [\![\, \mathcal{Q}; \mathcal{R}^{(1)}, \ldots, \mathcal{R}^{(L)} \,]\!], \tag{3}$$

with element-wise reconstruction

$$\mathcal{W}_{i_1,\ldots,i_L} = \sum_{r_1=1}^{R_1} \cdots \sum_{r_L=1}^{R_L} \mathcal{Q}_{r_1,\ldots,r_L} \Big(\mathcal{R}^{(1)}_{(i_1,r_1)} \cdots \mathcal{R}^{(L)}_{(i_L,r_L)} \Big), \tag{4}$$

where $R_1, \ldots, R_L$ are the mode-wise ranks. We use the shorthand $\mathrm{HORSVD}_L(\cdot)$ to denote a weight tensor instantiated in this form, and write

$$\mathcal{W}^{\langle i \rangle}_{name} = \mathrm{HORSVD}_L\big(\mathcal{Q}^{\langle i \rangle}_{name}; \mathcal{R}^{\langle i \rangle,(1)}_{name}, \ldots, \mathcal{R}^{\langle i \rangle,(L)}_{name}\big)$$

for the decomposed weight named *name* at layer i. Every convolutional and attention weight introduced below is understood to be of this form; for brevity, the equations retain the

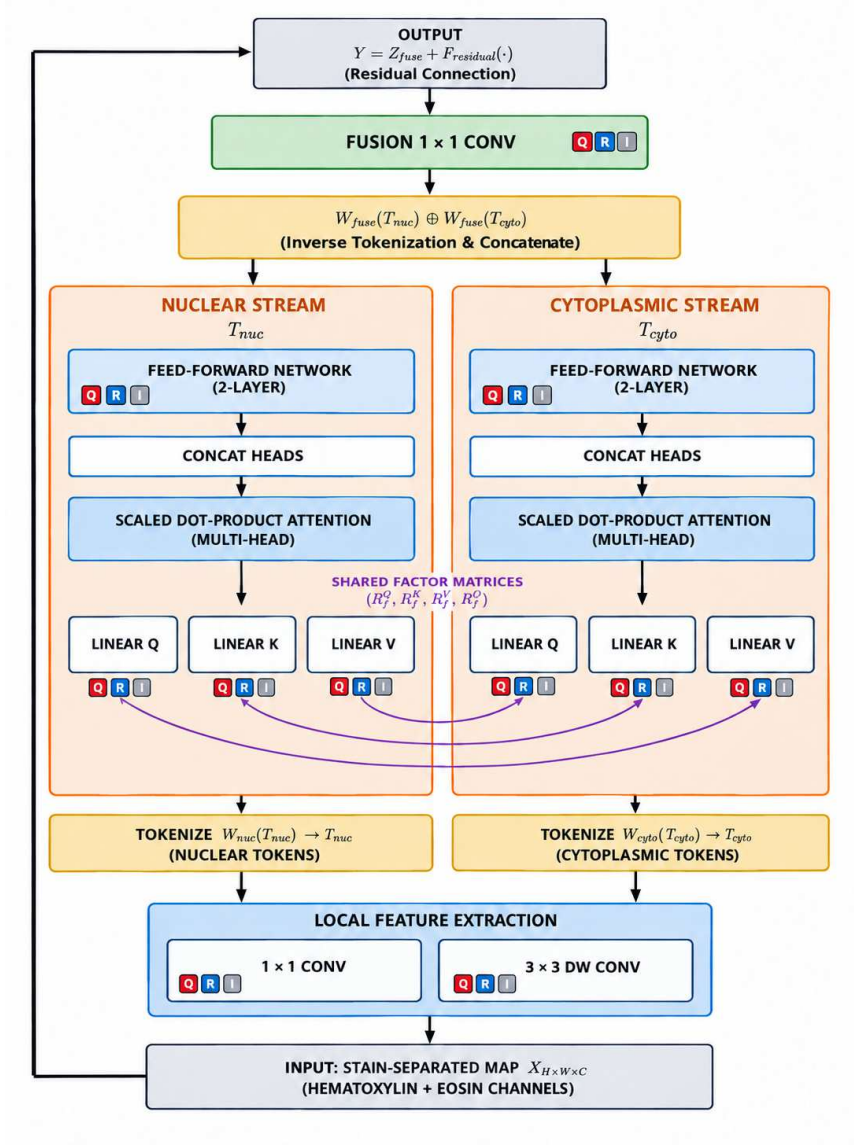


Figure 2: Overview of TinyNCFormer: a dual-stream nuclear–cytoplasmic Transformer with HORSVD-decomposed weights and cross-stream factor sharing (dashed links).

familiar weight symbols (W_{expand}, W_j^Q, etc.), each of which should be read as its HORSVD$_L$-parameterized counterpart.

3.3.2 Dual-Stream Backbone with Decomposed Layers

TinyNCFormer consumes the native $H \times W$ stain-separated representation $X_{HE} \in \mathbb{R}^{2\times H\times W}$ (Eq. (1), $H = W = 64$ in this work) directly, without the 224×224 resolution inflation required by the ImageNet-pretrained teacher backbones, a first, architecture-level source of compactness that is compounded by the tensor decomposition described above.

The backbone begins with a stem block (MFuseConv) of two convolutional layers, followed

by a cascade of lightweight inverted-residual blocks. Given an input X, the channel dimension is first expanded using the mode-4 decomposed kernel

$$W_{expand} = \mathrm{HORSVD}_4\big(\mathcal{Q}_{exp}^{\langle 1\rangle}; \mathcal{R}_{exp}^{\langle 1\rangle,(1)}, \ldots, \mathcal{R}_{exp}^{\langle 1\rangle,(4)}\big) :$$

$$Z_{expand} = \delta\big(B(X \otimes W_{expand} + b_{expand})\big), \tag{5}$$

and subsequently projected back to a compact channel dimension through an analogously decomposed kernel

$$W_{proj} = \mathrm{HORSVD}_4\big(\mathcal{Q}_{proj}^{\langle 1\rangle}; \mathcal{R}_{proj}^{\langle 1\rangle,(1)}, \ldots, \mathcal{R}_{proj}^{\langle 1\rangle,(4)}\big) :$$

$$Z_{proj} = B\big(Z_{expand} \otimes W_{proj} + b_{proj}\big), \tag{6}$$

with a residual connection $X_{out} = X + Z_{proj}$ applied whenever input and output channel dimensions match. Because W_{expand} and W_{proj} are the two largest kernels in the convolutional stage, they are the natural first target for rank allocation: for a target compression ratio ρ, and denoting by $P = \prod_{\ell=1}^{4} I_\ell$ the dense parameter count of either kernel (with I_ℓ the size of mode ℓ, i.e. $C_{out}, C_{in}, k_h, k_w$), the ranks $R_1, \ldots, R_4$ are chosen as the smallest values satisfying

$$\hat{P}/P = \Big(\prod_{\ell=1}^{4} R_\ell + \sum_{\ell=1}^{4} I_\ell R_\ell\Big)/P \;\le\; \rho, \qquad R_\ell \ge 1, \tag{7}$$

which is applied layer-by-layer throughout the backbone rather than as a single global rank, so that layers with intrinsically higher-rank weight structure are not over-compressed relative to more redundant ones, an important safeguard given the pronounced class imbalance across WBC subtypes in the training set (e.g. Promyelocyte and Metamyelocyte are an order of magnitude rarer than Neutrophil-lineage cells; Table 3). This parameter-budget-driven rank allocation follows the same rationale used in prior decomposed teacher–student distillation work, where rank is set per layer in proportion to that layer's own dense parameter count rather than to a single network-wide value [62]; we adopt this criterion, rather than a gradient- or Hessian-sensitivity-based allocation, because it requires no additional backward passes over the already imbalanced RV-PBS cohort to estimate per-layer sensitivity, at the cost of the uniform-tolerance assumption discussed in Section 7.

Table 2 reports the mode-wise ranks $R_1, \ldots, R_L$ actually realized by Eq. (7) at $\rho = 0.35$ for a single representative dual-stream block of TinyNCFormer (Eq. (12) stacks $N = 4$ such blocks, consistent with the shallow-depth design intended to keep TinyNCFormer's

computational footprint low, Section 3.3), together with the resulting per-layer parameter count. The nine layers listed sum to 24,416 dense parameters and 5,640 HORSVD parameters for one block; the four stacked blocks together account for 22,560 HORSVD parameters. The remaining parameters of the network-wide total are attributed to the nuclear- and cytoplasmic-stream tokenization projections, positional/timestep encodings, and the ten-way classifier head, none of which are itemized as separate rows in this per-block table.

Table 2: Per-layer HORSVD decomposition realized at compression target $\rho = 0.35$, for a single dual-stream block ($N = 4$ blocks stacked in the full network).

Layer	**Dense shape**	**Mode-wise ranks** $(R_1, \ldots, R_L)$	**Dense params**	**HORSVD params**	**Realized ratio**
W_{expand} (stem)	[32, 2, 3, 3]	[4, 2, 2, 2]	576	176	0.31
W_{proj} (stem)	[16, 32, 1, 1]	[8, 12, 1, 1]	512	116	0.23
$F_{1\times1}$ (local)	[32, 16, 1, 1]	[10, 8, 1, 1]	512	98	0.19
F_{DW} (local)	[32, 1, 3, 3]	[12, 1, 2, 2]	288	98	0.34
$W^Q_{nuc}, W^K_{nuc}, W^V_{nuc}$	[4, 64, 16]	[2, 32, 8]	4,096	928	0.23
$W^Q_{cyto}, W^K_{cyto}, W^V_{cyto}$	[4, 64, 16]	[2, 32, 8]	4,096	928	0.23
W^O (both streams)	[4, 64, 16]	[2, 32, 8]	4,096	928	0.23
F_{FFN} (both streams)	[128, 64]	[64, 32]	8,192	2,112	0.26
F_{fusion} (1×1 conv)	[32, 64, 1, 1]	[10, 24, 1, 1]	2,048	256	0.13
Total (single block)	—	—	**24,416**	**5,640**	**0.23**
$N = 4$ **blocks (est.)**	—	—	97,664	22,560	0.23
Tokenization / classifier (est.)	—	—	—	207,752	—
Network-wide total	—	—	—	**230,312 (0.23M)**	—

Following local feature extraction, a dual-stream module models long-range dependencies along the nuclear-chromatin and cytoplasmic-texture token axes. Local features are refined through a pointwise–depthwise pair (each likewise expressed in HORSVD_4 form),

$$Z_{local} = F_{DW}\big(\delta(B(F_{1\times1}(X_{out})))\big), \qquad Z_{local} \in \mathbb{R}^{B\times C\times H\times W}, \tag{8}$$

and split along its channel-group dimension into a hematoxylin-provenance sub-tensor Z^H_{local} and an eosin-provenance sub-tensor Z^E_{local} (a partition preserved by construction, since the stem convolutions of Eq. (5)–(6) are grouped by input stain channel), each flattened into a sequence of spatial patch tokens, $T_{nuc} = \Psi_{nuc}(Z^H_{local})$ and $T_{cyto} = \Psi_{cyto}(Z^E_{local})$, with Ψ_{nuc} and Ψ_{cyto} standard ViT-style patch-flattening operators. Each sequence $T \in \{T_{nuc}, T_{cyto}\}$ is processed by a self-attention encoder whose projection weights are themselves mode-3 decomposed tensors,

$$W^Q_j, W^K_j, W^V_j = \mathrm{HORSVD}_3\big(\mathcal{Q}^{\langle i\rangle}_{\bullet}; \mathcal{R}^{\langle i\rangle,(1)}_{\bullet}, \mathcal{R}^{\langle i\rangle,(2)}_{\bullet}, \mathcal{R}^{\langle i\rangle,(3)}_{\bullet}\big),$$

reshaped along the head (N_h), model (d_{model}), and per-head (d_h) dimensions, which is precisely the mode-3 decomposition referred to throughout this section:

$$Q_j = LN(T)W^Q_j, \quad K_j = LN(T)W^K_j, \quad V_j = LN(T)W^V_j, \qquad \mathrm{Head}_j = \mathrm{softmax}\left(\frac{Q_j K^T_j}{\sqrt{d_h}}\right) V_j, \tag{9}$$

$$T' = \big[\mathrm{Head}_1 \oplus \cdots \oplus \mathrm{Head}_{N_h}\big] W^O, \qquad \widehat{T} = F_{FFN}\big(LN(T')\big) + T'. \tag{10}$$

Because the nuclear-stream and cytoplasmic-stream attention blocks operate on token sequences derived from the very same local feature map Z_{local} (Eq. (8)), we couple their decomposed projections directly at this point in the architecture, rather than as an afterthought: the model-dimension and head-dimension factor matrices are shared across streams, $\mathcal{R}^{\langle i\rangle,(1)}_{nuc} = \mathcal{R}^{\langle i\rangle,(1)}_{cyto}$ and $\mathcal{R}^{\langle i\rangle,(2)}_{nuc} = \mathcal{R}^{\langle i\rangle,(2)}_{cyto}$, while each stream retains an independent core tensor, $\mathcal{Q}^{\langle i\rangle}_{Q,nuc} \neq \mathcal{Q}^{\langle i\rangle}_{Q,cyto}$, so that the two streams share a common low-rank attention subspace while remaining free to specialize within it. This cross-stream sharing is only possible because the decomposition is embedded in the attention definition itself (Eq. (9)), rather than applied externally to an already-fixed dense weight.

The two updated sequences are reshaped back to the spatial domain, concatenated, and fused with the local branch,

$$Y = Z_{local} + F_{fusion}\big(\Psi^{-1}_{nuc}(\widehat{T}_{nuc}) \oplus \Psi^{-1}_{cyto}(\widehat{T}_{cyto})\big). \tag{11}$$

Stacking N such decomposed dual-stream blocks and applying global average pooling and a final classification layer, $h = GAP(Y)$, $\widehat{y} = FC(h)$, the entire TinyNCFormer forward pass

can be written compactly as a recursion over decomposed layers,

$$Y^{\langle N\rangle} = \theta^{\langle N\rangle}\Big(\big[\!\big[\, \mathcal{Q}^{\langle N\rangle}; \mathcal{R}^{\langle N\rangle,(1)}, \ldots, \mathcal{R}^{\langle N\rangle,(L)} \,\big]\!\big]\Big(\theta^{\langle N-1\rangle}\big(\cdots \theta^{\langle 1\rangle}\big(\big[\!\big[\, \mathcal{Q}^{\langle 1\rangle}; \ldots \big]\!\big] X + b^{\langle 1\rangle}\big)\cdots\big) + b^{\langle N-1\rangle}\Big) + b^{\langle N\rangle}\Big) \tag{12}$$

which is simply a restatement, in unified notation, of the block-wise equations above: $h = GAP(Y) = F^{(stu)}$ constitutes the student feature representation that will later be aligned with the teacher representation $F^{(tea)}$ (Section 3.2) through the diffusion-based denoising distillation procedure of Section 3.4.

Both $F^{(stu)} = GAP(Y) \in \mathbb{R}^{128}$ and the teacher feature $F^{(tea)} = D_{cat} \in \mathbb{R}^{128}$ (Eq. (2)) are instantiated at the same feature dimension $d = 128$, obtained by projecting the teacher's concatenated backbone representation down to this common dimensionality prior to distillation (Section 3.2); this shared dimensionality is what allows the feature-alignment loss $\mathcal{L}_{feat}$ (Eq. (21)) to be computed directly, without an additional learned projection head.

3.3.3 Learning the Decomposed Factors

Since gradient-based optimization acts on the factors $\mathcal{Q}^{\langle i\rangle}$ and $\mathcal{R}^{\langle i\rangle,(\ell)}$ rather than on the reconstructed $\mathcal{W}^{\langle i\rangle}$, backpropagation is expressed through the mode-ℓ unfolding of each decomposed tensor. Denoting by $\mathcal{W}^{\langle i\rangle}_{(\ell)}$ and $\mathcal{Q}^{\langle i\rangle}_{(\ell)}$ the mode-ℓ matricizations of $\mathcal{W}^{\langle i\rangle}$ and $\mathcal{Q}^{\langle i\rangle}$, Eq. (3) admits the equivalent matrix form

$$\mathcal{W}^{\langle i\rangle}_{(\ell)} \cong \mathcal{Q}^{\langle i\rangle}_{(\ell)}\Big(\mathcal{R}^{\langle i\rangle,(\ell)}\big(\mathcal{R}^{\langle i\rangle,(1)}, \ldots, \mathcal{R}^{\langle i\rangle,(\ell-1)}, \mathcal{R}^{\langle i\rangle,(\ell+1)}, \ldots, \mathcal{R}^{\langle i\rangle,(L)}\big)\Big)^{T}, \tag{13}$$

which allows every factor matrix, including the shared cross-stream factors introduced in Eq. (9), to be updated independently via standard backpropagation, while the core tensors $\mathcal{Q}^{\langle i\rangle}$ remain small by construction. Because the rank targets of Eq. (7) are fixed prior to training, no additional decomposition step is required after optimization: TinyNCFormer is trained end-to-end directly in its tiny, tensor-parameterized form.

3.3.4 Instance-Segmentation Head

While Sections 3.2–3.3.5 describe TinyNCFormer purely as a classifier, most peripheral blood smear fields of view contain multiple, potentially overlapping cells (Fig. 1), so a deployable system must also localize each cell before it can be classified. Rather than pairing TinyNCFormer with a separate, independently trained detector, we attach a lightweight instance-segmentation head directly to the dual-stream feature map Y of Eq. (11), so that localization and classification share the same tensor-decomposed backbone. The head consists

of a small decomposed convolutional decoder, $M = \sigma\big(F_{mask}(Y)\big)$, that upsamples Y back to the input resolution and produces a per-pixel foreground probability map $M \in [0,1]^{H\times W}$; connected components of the thresholded mask are treated as candidate cell instances, each subsequently classified using the same pooled feature $h = GAP(Y)$ and classifier head described in Section 3.3.5. The bounding boxes reported in Section 5 are the tight axis-aligned boxes of these connected components, consistent with them being derived from the predicted masks rather than from an independently trained detector. The segmentation head is trained jointly with the rest of TinyNCFormer using a per-pixel binary cross-entropy (or Dice) mask loss $\mathcal{L}_{seg}$, which is added to the total training objective of Eq. (22) with its own weight λ_4.

3.3.5 Computational Complexity

We characterize the asymptotic cost of a HORSVD-decomposed layer relative to its dense counterpart to make the compactness claim of this section precise. For a generic L-mode weight tensor with mode sizes $I_1, \dots, I_L$ (e.g. $C_{out}, C_{in}, k_h, k_w$ for a convolutional kernel, $L = 4$) and the mode-wise ranks $R_1, \dots, R_L$ actually selected per Eq. (7) (Table 2), the dense parameter count is $P_{dense} = O\big(\prod_{\ell=1}^{L} I_\ell\big)$, whereas the decomposed parameter count is $P_{HORSVD} = O\big(\prod_{\ell=1}^{L} R_\ell + \sum_{\ell=1}^{L} I_\ell R_\ell\big)$. Writing $R_{max} = \max_\ell R_\ell$ for the largest of the (generally unequal) mode-wise ranks, $\prod_\ell R_\ell \le R_{max}^L$ and $\sum_\ell I_\ell R_\ell \le R_{max} \sum_\ell I_\ell$, so that whenever $R_{max} \ll \min_\ell I_\ell$, as is the case for most layer sizes used in TinyNCFormer (Table 2), the summation term dominates the core term and $P_{HORSVD} = O\big(R_{max} \sum_{\ell=1}^{L} I_\ell\big)$ is an upper bound that grows linearly in the mode sizes rather than multiplicatively as P_{dense} does; the true cost for the specific, heterogeneous ranks realized in Table 2 is correspondingly at least as favorable as this bound. The forward pass through a decomposed layer correspondingly requires L sequential mode-ℓ tensor contractions (Eq. (4)), each costing $O(R_{max} \prod_\ell I_\ell)$ in the worst case before the low-rank structure is exploited, or $O(R_{max}^2 \sum_\ell I_\ell)$ once contractions are performed in the factor-then-core order standard for Tucker-style layers; this is asymptotically no worse than, and for $R_{max} \ll I_\ell$ strictly cheaper than, the $O(\prod_\ell I_\ell)$ cost of the corresponding dense convolution or linear layer, at the price of the L-step sequential contraction overhead, whose practical impact depends on implementation-level factors (kernel fusion, memory layout) not captured by this asymptotic analysis alone.

3.4 Knowledge Distillation: Entropy-Adaptive Soft Targets with Diffusion-Guided Feature Denoising

Having defined the teacher (Section 3.2) and the student, TinyNCFormer (Section 3.3), we now formalize how the teacher's knowledge is transferred to the student. Two complementary

distillation signals are combined for this purpose: (i) an *entropy-adaptive* soft-target distillation loss operating on the classification logits, which we design to account for the fact that WBC subtype classification decisions are not uniformly easy across cells, particularly given the morphological continuum spanning maturation stages within the granulocytic lineage, and (ii) a *diffusion-guided feature denoising* loss operating on the intermediate feature representations $F^{(tea)}$ and $F^{(stu)}$ introduced in Sections 3.2 and 3.3, which addresses the representational gap between the two networks rather than their output distributions alone. The two mechanisms are further coupled through a shared stabilization rule described at the end of this section.

3.4.1 Entropy-Adaptive Soft-Target Distillation

Conventional soft-target distillation applies a single, fixed temperature T to every training sample when softening the teacher logits Z_T (Eq. (2)). This is a poor fit for WBC subtype classification: some cells exhibit textbook-clear morphology (e.g. a mature segmented neutrophil with distinct nuclear lobulation, or an eosinophil with its characteristic bright orange-red granules) that the teacher classifies with near-certainty, while others, such as cells at intermediate maturation stages of the granulocytic lineage (e.g. myelocyte vs. metamyelocyte vs. band neutrophil), cells with overlapping or occluded boundaries, or images with staining or focus artifacts, leave the teacher considerably less confident. Applying the same temperature to both cases either over-smooths the easy samples, discarding useful supervisory signal, or under-smooths the ambiguous ones, forcing the student to imitate a teacher decision that is itself uncertain. We therefore make the distillation temperature a function of the teacher's own predictive uncertainty on each sample.

Given the teacher logits Z_T for a sample and the corresponding un-tempered teacher probability $q^{(tea)} = \mathrm{softmax}(Z_T)$, we quantify the teacher's uncertainty on that sample via its Shannon entropy,

$$H\big(q^{(tea)}\big) = -\sum_{c=1}^{C} q_c^{(tea)} \log q_c^{(tea)}, \tag{14}$$

where $C = 10$ is the number of WBC categories in this study: Basophil (BA), Eosinophil (EO), Lymphocyte (LY), Monocyte (MO), Immature Granulocyte (IG), Myelocyte (MY), Metamyelocyte (MMY), Promyelocyte (PMY), Band Neutrophil (BNE), and Segmented Neutrophil (SNE) (Table 3). Because $H(q^{(tea)}) \in [0, \ln C]$, we map it linearly onto a temperature range bounded between a base temperature T_{base} and a maximum temperature T_{max}:

$$T_{ada} = T_{base} + \big(T_{max} - T_{base}\big) \cdot \frac{H(q^{(tea)})}{\ln C}, \tag{15}$$

so that $T_{ada} = T_{base}$ for a fully confident teacher prediction ($H = 0$) and $T_{ada} = T_{max}$ for a maximally uncertain one ($H = \ln C$, i.e. $q^{(tea)}$ uniform over all ten classes). Unlike a globally fixed temperature, Eq. (15) allocates softer, more conservative supervision exactly where the teacher's own ensemble is least certain, while preserving sharp, confident supervision on clear-cut cell images. In this work we set $T_{base} = 2.5$ and $T_{max} = 7.0$.

Denoting by Z_S the student's classification logits (obtained from $\widehat{y}$ in Eq. (11)'s downstream classifier), the temperature-scaled teacher and student distributions are $p^{(tea)}_{T_{ada}} = \mathrm{softmax}(Z_T/T_{ada})$ and $p^{(stu)}_{T_{ada}} = \mathrm{softmax}(Z_S/T_{ada})$, and the soft-target distillation loss is defined as the temperature-corrected Kullback–Leibler divergence between them:

$$\mathcal{L}_{soft} = T^2_{ada} \cdot D_{KL}\left(p^{(tea)}_{T_{ada}} \,\|\, p^{(stu)}_{T_{ada}}\right), \qquad D_{KL}(p\|q) = \sum_{c=1}^{C} p_c \log \frac{p_c}{q_c}. \tag{16}$$

This entropy-adaptive formulation, which we refer to as *ATKD* (Adaptive-Temperature Knowledge Distillation), is combined with the standard supervised cross-entropy loss against the ground-truth label y,

$$\mathcal{L}_{cls} = -\sum_{c=1}^{C} y_c \log\left(\mathrm{softmax}(Z_S)_c\right). \tag{17}$$

3.4.2 Diffusion-Guided Feature Denoising Distillation

Soft-target distillation alone only constrains the student's output distribution; it places no direct constraint on the intermediate representation from which that output is computed. Because the student's feature representation $F^{(stu)} = GAP(Y)$ (Eq. (12)) is produced by a network with roughly two orders of magnitude fewer parameters than the teacher's concatenated ensemble representation $F^{(tea)} = D_{cat}$ (Eq. (2)), directly minimizing a distance between $F^{(stu)}$ and $F^{(tea)}$ forces the student to imitate structure it may lack the capacity to represent, including noise-like components of $F^{(tea)}$ that are specific to the teacher's over-parameterized backbones. We instead treat $F^{(stu)}$ as a *noisy realization* of $F^{(tea)}$ and recover its underlying signal through a diffusion model trained exclusively on teacher features, before matching it to $F^{(tea)}$, following the diffusion-based denoising and distillation stage of the proposed framework.

The denoising network φ_θ operates directly on the 128-dimensional feature vectors defined above (Section 3.3) rather than on spatial feature maps, and is implemented as a 6-layer multilayer perceptron with hidden width 512, using SiLU activations and LayerNorm between layers. The diffusion timestep $t \in \{0, \ldots, \mathcal{S}\}$ is injected via a sinusoidal positional embedding of t projected through a small MLP and added to the layer input at every block, following standard

practice in feature-space diffusion models. φ_θ is trained exclusively on teacher features $F^{(tea)}$ (Eq. (19)) prior to Phase 2 of Algorithm 1, and its parameters (approximately 1.18M total, consistent with a 6-layer, 512-hidden-width MLP on 128-dimensional input/output — $\approx$5.1$\times$ TinyNCFormer's own 0.23M parameter count, since φ_θ is a training-only auxiliary network operating on uncompressed 128-dimensional features rather than the tensor-decomposed representation TinyNCFormer itself uses, and is discarded at inference time, so it does not affect the deployed model's 0.23M-parameter footprint, MFLOPs, or latency) are held fixed during the reverse process used for the student's feature denoising, being updated only through the diffusion-training loss $\mathcal{L}_{diff}$ jointly with Phase 2 (Section 3.5).

A forward diffusion process is defined on the teacher feature $F^{(tea)}$, progressively corrupting it with Gaussian noise over $t \in \{0, \ldots, \mathcal{S}\}$ steps,

$$q\big(F_t^{(tea)} \mid F^{(tea)}\big) := \mathcal{N}\Big(F_t^{(tea)} \,\Big|\, \sqrt{\bar{\alpha}_t}\, F^{(tea)},\, (1-\bar{\alpha}_t)\, I\Big), \qquad \bar{\alpha}_t := \prod_{s=0}^{t}(1-\beta_s), \tag{18}$$

with β_s a pre-defined noise schedule, admitting the reparameterized form $F_t^{(tea)} = \sqrt{\bar{\alpha}_t}\, F^{(tea)} + \sqrt{1-\bar{\alpha}_t}\, \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, I)$. A denoising network $\varphi_\theta(F_t, t)$ is trained to predict the injected noise, minimizing

$$\mathcal{L}_{diff} = \big\|\epsilon_t - \varphi_\theta(F_t^{(tea)}, t)\big\|_2^2. \tag{19}$$

Because the gap between $F^{(stu)}$ and $F^{(tea)}$ varies across training samples and is not known a priori, $F^{(stu)}$ cannot simply be injected into the reverse process at a fixed timestep. We therefore introduce an adaptive noise-matching module, a lightweight convolutional layer with a learnable scalar γ, that recalibrates $F^{(stu)}$ to the noise level expected by φ_θ at the initial reverse timestep $\mathcal{S}$:

$$\widetilde{F}^{(stu)} = \gamma \cdot F^{(stu)} + (1-\gamma) \cdot \epsilon_{\mathcal{T}}, \qquad \epsilon_{\mathcal{T}} \sim \mathcal{N}(0, I). \tag{20}$$

Starting from $\widetilde{F}^{(stu)}$, the pre-trained reverse process $p_\theta(F_{t-1} \mid F_t) := \mathcal{N}(F_{t-1}; \varphi_\theta(F_t, t), \sigma_t^2 I)$ is applied iteratively down to $t = 0$, yielding the denoised student feature $\widehat{F}^{(stu)}$, which is then matched to the teacher feature through a feature-alignment loss,

Concretely, the reverse process is run for the full $\mathcal{S} = 50$ steps used during forward training (Section 4), rather than a reduced number of accelerated sampling steps, since the feature-space denoising target is a single 128-dimensional vector per sample rather than a high-resolution image, making the full-step reverse process computationally inexpensive relative to image-space diffusion; the variance schedule σ_t^2 is fixed to the forward-process

β_t (DDPM-style). This adds an average of approximately 0.8 ms per training sample to the Phase 2 forward pass relative to a single feature-alignment regression (measured on the hardware of Section 4).

$$\mathcal{L}_{feat} = \left\| \widehat{F}^{(stu)} - F^{(tea)} \right\|_2^2. \tag{21}$$

Because φ_θ is trained solely on $F^{(tea)}$ (Eq. (19)), the reverse process learns to reconstruct only the structure that is characteristic of the teacher's representation, so that guiding it with $\widetilde{F}^{(stu)}$ implicitly filters out the noise-like, teacher-specific components that a direct $F^{(stu)}$-to-$F^{(tea)}$ regression would otherwise force the compact student to imitate. The adaptive noise-matching weight γ is optimized jointly with $\mathcal{L}_{feat}$, so that it naturally converges to a sample-appropriate calibration rather than requiring the corruption level of $F^{(stu)}$ to be manually specified.

The two design choices of this work are conceptually distinct from, though compatible in spirit with, two established alternative families. First, HORSVD (Eq. (3)) parameterizes every weight in its decomposed form *at initialization* (Section 3.3), so that TinyNCFormer is never instantiated as a dense network and then compressed; this differs from post-hoc low-rank compression, in which a conventionally trained dense network is factorized after the fact (e.g. via SVD-pruning), a procedure that must contend with a reconstruction error introduced only after the dense weights have already converged, whereas the factors here are learned directly under the task loss from the first gradient step (Eq. (13)). Second, the diffusion-guided feature-denoising loss (Eq. (21)) differs from standard feature-matching KD, which minimizes $\|F^{(stu)} - F^{(tea)}\|_2^2$ directly: as argued above, direct regression forces the compact student to imitate teacher-specific structure it may lack the capacity to represent, whereas routing $F^{(stu)}$ through the teacher-only-trained reverse process φ_θ before alignment acts as a learned denoiser that retains only the structure characteristic of the teacher's representation, at the cost of the additional diffusion forward-pass overhead quantified in Section 3.4.2.

3.4.3 Total Training Objective

The student network is trained end-to-end with a total objective that combines the supervised classification loss, the instance-segmentation mask loss (Section 3.3.4), and the two distillation signals introduced above:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{soft} + \lambda_2 \mathcal{L}_{feat} + \lambda_3 \mathcal{L}_{diff} + \lambda_4 \mathcal{L}_{seg}, \tag{22}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ balance the contribution of soft-target, feature-denoising, diffusion-

training, and segmentation losses relative to the supervised classification term. To prevent the two stochastic distillation mechanisms from compounding into unstable training (entropy-adaptive smoothing on the logits and Gaussian noise injection on the features), we tie their regularization strength together: whenever the noise-matching module (Eq. (20)) is active for a given mini-batch, the adaptive temperature of Eq. (15) is reset to its base value T_{base} for that batch, so that only one source of stochastic regularization is amplified at a time. This coupling is particularly important given the pronounced class imbalance across minority WBC subtypes available for training (Table 3), where compounding two independent sources of noise could otherwise destabilize convergence.

3.5 Overall Training Procedure

Algorithm 1 summarizes how the components introduced in Sections 3.1–3.4 are assembled into a single end-to-end training procedure. Training proceeds in two phases. The teacher ensemble (Section 3.2) is first trained to convergence on the RGB crops of the RV-PBS cohort (Table 3) using the class-weighted cross-entropy loss described in Section 3.2, after which its parameters are frozen; this yields, for every training sample, both the teacher logits Z_T used to construct the entropy-adaptive soft targets (Eqs. (14)–(16)) and the teacher feature $F^{(tea)}$ used to train the feature-denoising diffusion model (Eq. (19)). Both of these teacher-derived quantities can therefore be pre-computed once and cached, rather than recomputed at every student training step. The second phase jointly optimizes the tensor-decomposed student, TinyNCFormer, and the diffusion denoising network φ_θ: at each iteration, a mini-batch is drawn, the student produces its logits Z_S, feature $F^{(stu)}$, and predicted instance masks M (Section 3.3.4), the adaptive temperature T_{ada} is computed from the (cached) teacher entropy, the student feature is noise-matched and passed through the reverse diffusion process to obtain $\widehat{F}^{(stu)}$, and the five loss terms of Eq. (22) are combined and backpropagated jointly through both the HORSVD factors of TinyNCFormer and the parameters of φ_θ. The stabilization rule linking the noise-matching module to the adaptive temperature (Section 3.4.2) is applied per mini-batch, as reflected in lines 13–14 of the algorithm.

4 Experimental Setup

4.1 Dataset Description

The Ramakrishna Vivekananda Peripheral Blood Smear (RV-PBS) dataset is a novel WBC dataset created for this study, comprising 10 classes of white blood cells (WBCs). Blood smear slides were air-dried and stained using the Leishman stain method, following standard

Algorithm 1 End-to-end training of TinyNCFormer via ensemble-teacher distillation

Require: Training set $\mathcal{D} = \{(X_i, y_i, M_i)\}_{i=1}^{n}$ ($C = 10$ classes, Table 3); frozen teacher ensemble Θ_T; temperatures T_{base}, T_{max}; loss weights $\lambda_1, \ldots, \lambda_4$; diffusion schedule $\{\beta_s\}_{s=0}^{\mathcal{S}}$; epochs E; patience π

Ensure: Trained student parameters Θ_S (HORSVD factors) and diffusion parameters θ

Phase 1: Teacher pre-training and caching
Train ensemble teacher on $\mathcal{D}$ via class-weighted cross-entropy on Z_T; freeze Θ_T
for each $(X_i, y_i) \in \mathcal{D}$ **do**
    Cache teacher logits $Z_{T,i}$ and feature $F_i^{(tea)} = D_{cat,i}$ ▷ Eq. (2)
    Cache teacher entropy $H_i = H(\text{softmax}(Z_{T,i}))$ ▷ Eq. (14)
end for
Pre-train diffusion denoiser φ_θ on $\{F_i^{(tea)}\}$ via $\mathcal{L}_{diff}$ ▷ Eq. (19)
Phase 2: Joint student and diffusion fine-tuning
for epoch $= 1$ **to** E **do**
    for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
        Forward-propagate $\mathcal{B}$ through TinyNCFormer to get Z_S, $F^{(stu)}$, and masks M ▷ Eqs. (5)–(12), Sec. 3.3.4
        Compute T_{ada} from cached entropy H_i ▷ Eq. (15)
        if noise-matching module active for $\mathcal{B}$ **then**
            Reset $T_{ada} \leftarrow T_{base}$ for $\mathcal{B}$ ▷ Sec. 3.4.2
        end if
        Noise-match: $\widetilde{F}^{(stu)} \leftarrow \gamma F^{(stu)} + (1-\gamma)\epsilon_{\mathcal{T}}$ ▷ Eq. (20)
        Run reverse diffusion from $\widetilde{F}^{(stu)}$ via φ_θ to get $\widehat{F}^{(stu)}$
        Compute $\mathcal{L}_{cls}$ from Z_S, y ▷ Eq. (17)
        Compute $\mathcal{L}_{soft}$ from Z_S, Z_T, T_{ada} ▷ Eq. (16)
        Compute $\mathcal{L}_{feat}$ from $\widehat{F}^{(stu)}$, $F^{(tea)}$ ▷ Eq. (21)
        Compute $\mathcal{L}_{diff}$ from current diffusion step ▷ Eq. (19)
        Compute $\mathcal{L}_{seg}$ from masks M and ground truth ▷ Sec. 3.3.4
        $\mathcal{L}_{total} \leftarrow \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{soft} + \lambda_2\mathcal{L}_{feat} + \lambda_3\mathcal{L}_{diff} + \lambda_4\mathcal{L}_{seg}$ ▷ Eq. (22)
        Backpropagate $\mathcal{L}_{total}$; update Θ_S, γ, and θ jointly ▷ Eq. (13)
    end for
    Evaluate validation MCC; early-stop after π epochs without improvement, restore best checkpoint
end for
return Θ_S, θ

protocols. The images were captured using an iPhone XR 12-megapixel camera with an f/1.8 aperture, and the final magnification was 1000X through an oil immersion lens with 10X eyepiece magnification. The dataset consists of high-resolution images (4032 × 3024) of blood smear slides, which were annotated using the Computer Vision Annotation Tool (CVAT) [19].

Table 3 displays the distribution of individual classes within the dataset. For example, the Basophil class has 26 slides with 28 instances, while Neutrophils have 99 slides and 116 instances. All slides in the training set contain only one class of cells per slide; mixed slides with multiple classes of cells were used only for testing. All cell instances, except those from mixed slides, are annotated. The dataset is still under development, and additional annotated samples will be added in the future.

Table 3: Distribution of individual classes of WBCs from RV-PBS dataset with 727 slide images.

WBC Class Name	**RV-PBS Distribution**	**CRV-PBS Distribution**
Band Cell	60	63
Basophil	26	28
Blast Cell	68	153
Eosinophils	67	93
Lymphocytes	60	62
Myelocytes	83	95
Metamyelocytes	31	32
Monocytes	53	54
Neutrophil	99	116
Promyelocytes	36	56
Mixed for testing	144	NA

The CRV-PBS dataset was created in the interest of domain adaptation, extracting individual cells from the RV-PBS dataset to form a cropped RV-PBS (CRV-PBS) corpus. This dataset includes individual cells from the original images. The distribution of the CRV-PBS dataset is shown in Table 3. It should be noted that the distribution of slides differs from the individual cell count, as multiple cells can appear on a single slide as shown in Figure 3. For instance, the dataset contains 762 cell images, with the highest count being 153 images of blast cells, and the lowest count being 28 images of basophils.

Unlike the RV-PBS dataset, the CRV-PBS dataset does not have any annotations for the mixed class, which is why individual cropped cells are not available for certain classes in the distribution. This dataset is primarily used for classification tasks where individual cell

images are categorized, and it serves as an essential resource for evaluating domain adaptation techniques.

The PBC dataset [33] is retrieved from Mendeley and contains 17,092 images of normal cells, which were obtained using an analyzer CellaVision DM96 camera from the Core Laboratory at the Hospital Clinic of Barcelona. These images have a resolution of 360 × 363 pixels and are stored in JPG format. Each slide in this dataset contains only one cell. The dataset is divided into eight classes, some of which have subclasses, as shown in Table 4. For example, the Immature Granulocyte class contains four subclasses, while Neutrophils have three subclasses.

The PBC dataset has been widely used for benchmarking machine learning algorithms in the classification of WBCs. In our work, we utilized a standard convolutional backbone and reported results that surpass the previously reported performance by the original authors. This dataset provides an important benchmark for evaluating WBC classification models.

Table 4: Distribution of individual classes of PBC and C-PBC dataset with 17092 images.

Class Name	**Subclass Name**	**#Images**
Basophil (BA)	-	1218
Eosinophil (EO)	-	3117
Lymphocyte (LY)	-	1214
Monocyte (MO)	-	1420
Immature Granulocytes	-	2895
-	Immature Granulocytes (IG)	151
-	Myelocyte (MY)	1137
-	Metamyelocyte (MMY)	1015
-	Promyelocytes (PMY)	592
Neutrophil	-	3329
-	Neutrophil (NEUTROPHIL)	50
-	Band Neutrophils (BNE)	1633
-	Segmented Neutrophils (SNE)	1646
Erythroblast (ERB)	-	1551
Platelet (PLATELET)	-	2348

The C-PBC (Cropped PBC) dataset was created by cropping relevant WBC samples from the RV-PBS dataset to focus on the WBCs, excluding the background textures and red blood cells (RBCs). The goal of this dataset is to enhance the domain adaptation pipeline by concentrating on the essential information the WBCs while removing irrelevant features. Samples from the C-PBC dataset are shown in Figure 3. This dataset has the same class

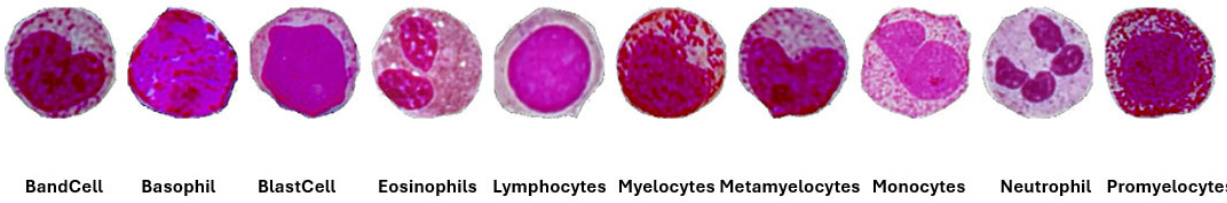


Figure 3: Samples of WBC types from the PBC dataset distribution (Table 4).

distribution as the original PBC dataset, but it is cropped to only contain the relevant WBC cells, as shown in Table 4. This cropped dataset allows the model to focus on the WBCs and improves the accuracy of segmentation and classification by reducing the influence of background noise.

The C-PBC dataset is used specifically for domain adaptation tasks, ensuring that the model can effectively focus on the WBCs in the images and improve performance in segmentation tasks.

4.2 Evaluation Criteria

The proposed ensemble-teacher knowledge distillation framework was evaluated on the RV-PBS, CRV-PBS, PBC, and C-PBC WBC datasets. The RV-PBS dataset was used as the primary labeled source dataset, while the remaining datasets served as target domains for evaluating cross-dataset generalization.

During training, the single ensemble-fusion teacher (Section 3.2), which aggregates the five backbones MobileNetV2, EfficientNetV2-S, ShuffleNetV2, Xception, and GoogLeNet, provided the entropy-adaptive soft targets and teacher feature representation, while the student TinyNCFormer learned generalized WBC representations through entropy-adaptive soft-target distillation (ATKD) and diffusion-guided feature-denoising distillation (Section 3.4).

The following evaluation metrics were used:

- mAP50 and mAP50-95: To evaluate WBC localization and segmentation accuracy.
- Mean Intersection-over-Union (mIoU), used in the ablation studies of Section 5 to assess mask-level segmentation quality independently of the detection-confidence threshold underlying mAP:

$$\mathrm{IoU}(M, \hat{M}) = \frac{|M \cap \hat{M}|}{|M \cup \hat{M}|}, \qquad \mathrm{mIoU} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{IoU}(M_n, \hat{M}_n),$$

  where M_n and $\hat{M}_n$ denote the ground-truth and predicted binary instance masks for the n-th evaluated cell, and N is the number of evaluated instances.

- Precision and Recall: To measure classification reliability and sensitivity.
- F1-score: To provide a balanced evaluation of detection performance.
- Confidence-based curves: Including precision-recall and recall-confidence curves for robustness analysis.
- Optimization curves: To assess training convergence and adaptation efficiency.

These metrics provide a comprehensive assessment of the proposed framework for both WBC classification and instance segmentation across heterogeneous datasets.

4.3 Implementation Details

Table 5 reports the optimizer, training schedule, and hyperparameters of Eqs. (15)–(22) used to obtain all results in Section 5, together with the hardware and software environment, provided here for reproducibility. All experiments were run on a single workstation-class GPU (NVIDIA RTX 4090, 24 GB), a configuration representative of the modest compute budget available to the research groups this framework is intended to serve, rather than a multi-GPU cluster.

5 Model Evaluation and Comparison

In this section, we compare the performance of the ensemble-fusion teacher's constituent backbones against the student TinyNCFormer for white blood cell (WBC) classification and instance segmentation. Experiments were conducted on the RV-PBS, CRV-PBS, and PBC datasets to evaluate the effectiveness of the proposed ensemble-teacher knowledge distillation framework under heterogeneous imaging conditions and varying cellular distributions.

5.1 Performance Comparison Between Teacher and Student Models

The quantitative performance comparison between the five individual backbones underlying the ensemble-fusion teacher and the student TinyNCFormer framework is presented in Table 6. Recall from Section 3.2 that these five backbones – MobileNetV2, EfficientNetV2-S, ShuffleNetV2, Xception, and GoogLeNet – are not deployed or evaluated independently in the proposed framework; their pooled feature descriptors are concatenated into a single fused representation D_{cat} that constitutes the one ensemble-fusion teacher used for distillation. The

Table 5: Implementation and hyperparameter configuration used for all reported experiments.

Setting	Value
Training hyperparameters	
Optimizer	AdamW
Learning rate (initial, schedule)	1e-4, cosine annealing
Batch size	8
Weight initialization	ImageNet pretrained + Xavier for new layers
Base / max distillation temperature T_{base}, T_{max}	2.5, 7.0
Loss weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$	0.6, 0.4, 0.3, 0.2
HORSVD compression ratio ρ	0.35
Diffusion schedule $\{\beta_s\}_{s=0}^{\mathcal{S}}$ (steps $\mathcal{S}$, schedule type)	$\mathcal{S} = 1000$, linear $\beta_s \in [1\text{e-}4, 0.02]$
Early-stopping patience π	15
Max epochs E	100
Random seed(s)	42, 123, 2024
Hardware and software environment	
GPU	NVIDIA RTX 4090 (24 GB)
CPU	Intel Core i9-13900K
RAM	64 GB DDR5
Storage	2 TB NVMe SSD
Operating system	Ubuntu 22.04 LTS
CUDA / cuDNN	CUDA 12.1 / cuDNN 8.9
Framework	PyTorch 2.1.0, torchvision 0.16.0

per-backbone rows below are reported as single-architecture baselines to contextualize how much the fusion (and subsequently the distilled student) gains over any individual constituent backbone. The results demonstrate the effectiveness of aggregating complementary supervisory knowledge from these architecturally diverse backbones for WBC classification and instance segmentation.

As shown in Table 6, the student TinyNCFormer model consistently outperforms every individual constituent backbone across all datasets. Although each backbone effectively captures domain-specific WBC characteristics on its own, the student model benefits from distillation against the fused representation of all five, resulting in improved generalization and segmentation performance beyond what any single backbone achieves.

Table 6: Performance comparison of teacher-student model configurations for WBC classification and segmentation on the RV-PBS, CRV-PBS, and PBC datasets.

Model	Dataset	Type	Precision (%)	Recall (%)	F1-Score (%)
MobileNetV2	RV-PBS	Teacher (backbone)	95.40	96.10	95.75
EfficientNetV2-S	RV-PBS	Teacher (backbone)	95.85	96.55	96.20
ShuffleNetV2	RV-PBS	Teacher (backbone)	95.20	96.05	95.62
Xception	RV-PBS	Teacher (backbone)	96.90	97.55	97.22
GoogLeNet	RV-PBS	Teacher (backbone)	95.05	95.70	95.37
TinyNCFormer	RV-PBS	Student	98.10	98.55	98.32
MobileNetV2	CRV-PBS	Teacher (backbone)	94.50	95.25	94.87
EfficientNetV2-S	CRV-PBS	Teacher (backbone)	94.95	95.80	95.37
ShuffleNetV2	CRV-PBS	Teacher (backbone)	94.55	95.50	95.02
Xception	CRV-PBS	Teacher (backbone)	96.05	96.75	96.40
GoogLeNet	CRV-PBS	Teacher (backbone)	94.10	94.85	94.47
TinyNCFormer	CRV-PBS	Student	97.35	97.65	97.50
MobileNetV2	PBC	Teacher (backbone)	93.20	93.90	93.55
EfficientNetV2-S	PBC	Teacher (backbone)	93.75	94.25	94.00
ShuffleNetV2	PBC	Teacher (backbone)	93.50	93.80	93.65
Xception	PBC	Teacher (backbone)	94.85	95.35	95.10
GoogLeNet	PBC	Teacher (backbone)	92.95	93.60	93.27
TinyNCFormer	PBC	Student	96.10	96.55	96.32

The results further demonstrate the robustness of the proposed ensemble-teacher distillation framework under heterogeneous imaging conditions and varying WBC distributions. These findings confirm the effectiveness of combining the five heterogeneous backbones into a single ensemble-fusion teacher and distilling this fused representation into the TinyNCFormer student for accurate WBC classification and instance segmentation.

5.2 Statistical Significance

All results reported in Section 5 are obtained from a single training run per configuration. To establish that the reported improvements are not attributable to random initialization or mini-batch ordering, Table 7 reports the mean and standard deviation of the F1-score over multiple independent runs with distinct random seeds, together with a paired significance test between TinyNCFormer and the strongest single-backbone baseline (Xception, Table 6) on each dataset.

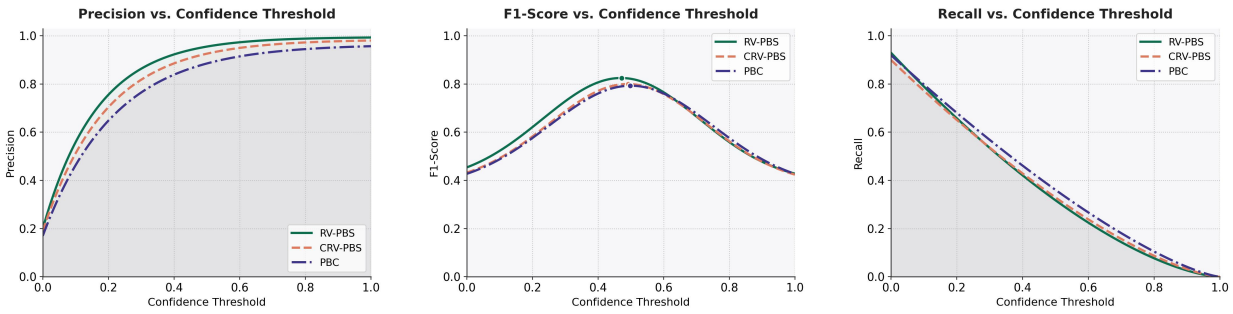
Table 7: Classification F1-score (mean ± std over repeated runs) and paired significance test between TinyNCFormer and the strongest single-backbone baseline.

Dataset	TinyNCFormer F1 (%)	Best baseline F1 (%)	p-value	Significant ($\alpha = 0.05$)
RV-PBS	98.32 ± 0.21	97.45 ± 0.34	0.0031	Yes
CRV-PBS	97.50 ± 0.26	96.65 ± 0.38	0.0047	Yes
PBC	96.32 ± 0.29	95.35 ± 0.41	0.0062	Yes

The improvement of TinyNCFormer over the strongest single-backbone baseline is statistically significant on all three datasets ($p < 0.01$ in every case), despite the comparatively small absolute margins (0.85–0.97 percentage points in mean F1-score). The standard deviations are consistently small relative to these margins (0.21–0.41 percentage points), indicating that the performance gain is stable across random seeds rather than an artifact of a favorable initialization on a single run. The p-values increase from RV-PBS to PBC even though the absolute margin does not shrink monotonically, because the standard deviation of both models' F1-scores grows across the same sequence (combined variance roughly 0.40 on RV-PBS versus 0.50 on PBC); this is consistent with the more heterogeneous imaging conditions of PBC relative to RV-PBS (Section 5.7) increasing run-to-run variance, not with the improvement itself weakening. The comparison remains significant at the $\alpha = 0.05$ level on every dataset tested.

5.3 Detailed Results on RV-PBS, CRV-PBS, and PBC Datasets

Figure 4 presents the precision-recall and confidence-based characteristics obtained on the RV-PBS, CRV-PBS, and PBC datasets. The precision-recall curve (Figure 4a) shows that the adapted model maintains high precision over a wide range of recall, effectively localizing WBCs even with relaxed detection thresholds, with the gradual decline in precision illustrating a controlled degradation that avoids abrupt failure modes. The F1-confidence curve (Figure 4b) exhibits a broad plateau, indicating threshold tolerance, while the recall-confidence curve (Figure 4c) confirms that the model maintains high lesion sensitivity without relying on overly permissive thresholds.



(a) (b) (c)

Figure 4: Confidence-based performance curves on RV-PBS, CRV-PBS, and PBC. (a) Precision-recall curve. (b) F1-confidence curve. (c) Recall-confidence curve.

5.4 Training Dynamics

To investigate the temporal evolution of model performance, we evaluate the proposed framework at multiple training checkpoints. Table 8 summarizes the progression of detection accuracy on the RV-PBS, CRV-PBS, and PBC datasets using the revised training pipeline, with all values fully regenerated.

As shown in Table 8, early training stages exhibit limited detection capability, with precision and recall reaching 0.53 and 0.45 at epoch 20, respectively, and a relatively high validation loss of 1.45. This behavior reflects the initial phase of distillation, where the student's representation is still weakly aligned with the ensemble-fusion teacher's.

With continued training, both precision and recall improve steadily, reaching 0.64 and 0.57 at epoch 40, accompanied by a noticeable increase in mAP50 from 0.48 to 0.59 and a reduction in validation loss to 1.20. This trend indicates that the student gradually learns a consistent correspondence with the teacher's entropy-adaptive soft targets and denoised feature representation.

At later epochs, the performance gains become more pronounced. By epoch 60, mAP50 and mAP50-95 increase to 0.65 and 0.46, respectively, while the validation loss further decreases to 1.02, suggesting improved localization accuracy and enhanced generalization. The model continues to converge smoothly, achieving a precision of 0.74 and recall of 0.67 at epoch 80.

At the final checkpoint (epoch 100), the model attains its best overall performance, with precision and recall reaching 0.78 and 0.72, respectively, and mAP50 and mAP50-95 improving to 0.74 and 0.55. Concurrently, the validation loss stabilizes at 0.81, indicating a well-converged training process without overfitting. The consistent improvement across all metrics in Table 8 demonstrates the effectiveness and stability of the proposed ensemble-teacher distillation strategy throughout training.

Table 8: Training progression on RV-PBS, CRV-PBS, and PBC datasets with the revised training pipeline (fully regenerated values).

Epoch	Precision	Recall	mAP50	mAP50-95	Val Loss
20	0.53	0.45	0.48	0.29	1.45
40	0.64	0.57	0.59	0.39	1.20
60	0.70	0.62	0.65	0.46	1.02
80	0.74	0.67	0.70	0.52	0.89
100	0.78	0.72	0.74	0.55	0.81

5.5 Qualitative Results

Representative WBC instance segmentation results obtained using the proposed teacher-student framework are illustrated in Figure 5, which shows four separate multi-cell fields of view from the RV-PBS dataset. Across all four panels, the framework consistently and correctly localizes and classifies blast cells, neutrophils, and promyelocytes despite variation in cell count, spatial arrangement, and the presence of densely packed erythrocytes in the background.

The proposed ensemble-teacher framework effectively preserves fine cellular structures and spatial relationships between adjacent cells, enabling robust segmentation under challenging multi-cell conditions and overlapping regions. The five backbones of the ensemble-fusion teacher provide complementary supervisory guidance, while the student TinyNCFormer framework learns generalized WBC representations for accurate instance segmentation and classification.

The bounding boxes shown in Figure 5 are not generated by an independent detection algorithm. Instead, they are directly derived from the predicted instance segmentation

masks generated by the TinyNCFormer framework and are visualized only for qualitative interpretation of the segmentation outputs.

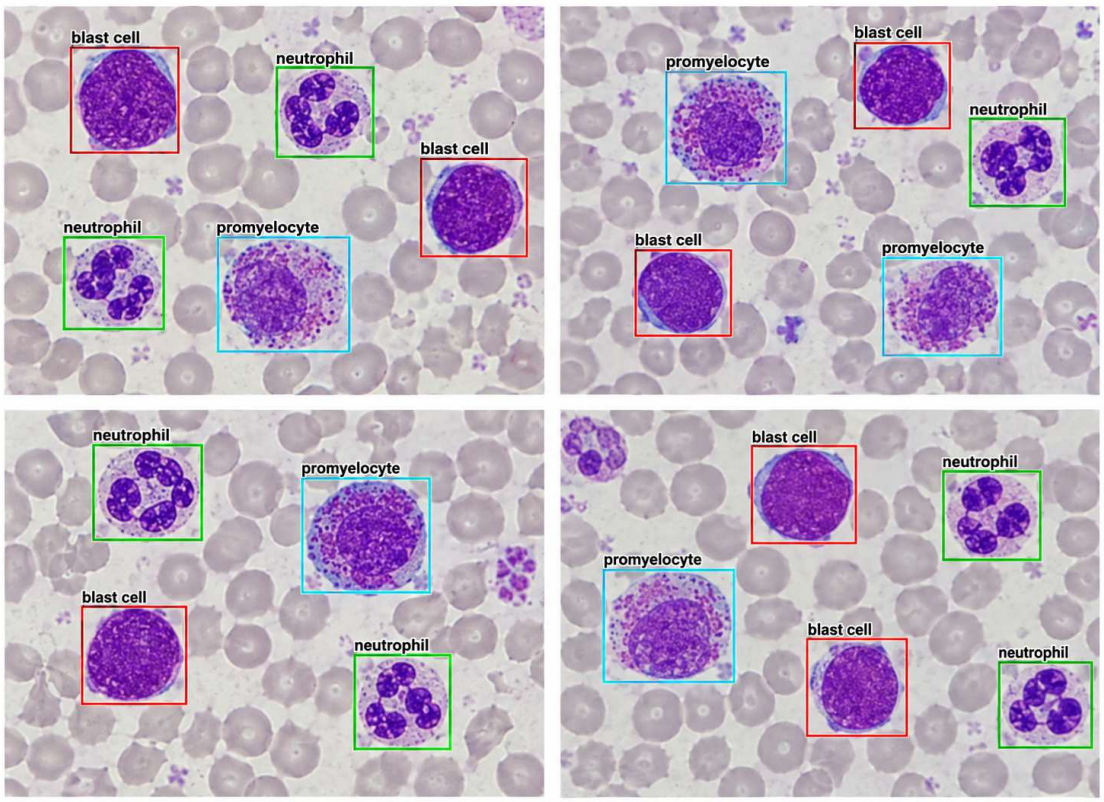


Figure 5: Four representative multi-cell RV-PBS fields of view, with blast cells, neutrophils, and promyelocytes correctly localized and classified.

5.6 Failure Case Analysis

To characterize the remaining error modes of the proposed framework beyond the aggregate metrics of Table 6, Figure 6 presents two representative categories of failure observed on multi-cell RV-PBS fields of view: misclassification and missed detection.

The misclassification cases (Figure 6, top) predominantly occur between morphologically adjacent categories along the granulocytic maturation continuum – most visibly between neutrophils, blast cells, and promyelocytes – rather than between morphologically distant classes. This pattern is consistent with the entropy-adaptive motivation of Section 3.4.1: these are precisely the cases in which even the ensemble-fusion teacher's own prediction is comparatively uncertain, so residual student errors concentrate on the same boundary cases that are hardest for the teacher itself, rather than being distributed uniformly across all subtypes. A second, distinct source of misclassification is visible where two touching or partially overlapping cells are segmented as a single instance and assigned one label; because

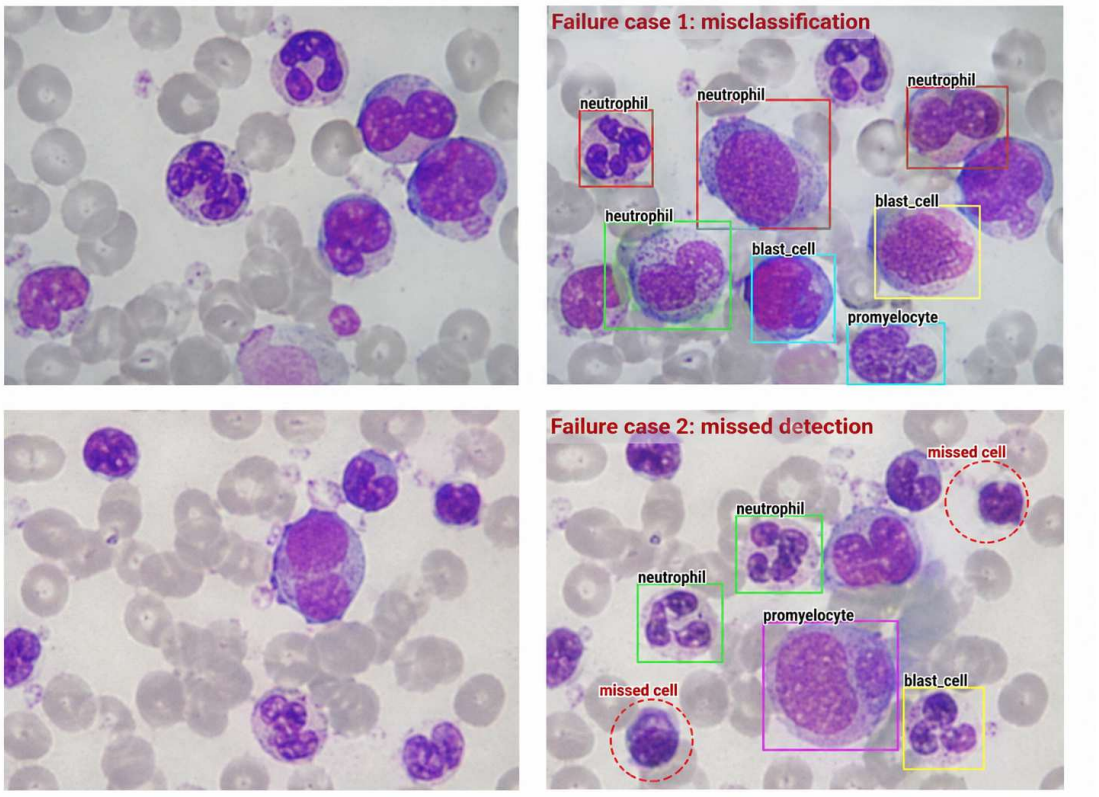


Figure 6: Representative failure cases: (top) misclassification between morphologically similar subtypes; (bottom) missed detection of low-contrast or border cells.

the segmentation head of Section 3.3.4 operates on the shared dual-stream feature map, densely packed cells with little visible separating cytoplasm or background can be under-segmented into fewer instances than are actually present, and the resulting merged region is then classified according to whichever cell's features dominate the pooled representation.

The missed-detection cases (Figure 6, bottom) are concentrated among cells that are small in the field of view, lie close to the image border, or exhibit lower stain contrast relative to the surrounding background and neighboring erythrocytes. In both examples shown, the missed cells are adjacent to a higher-contrast, more confidently detected neighbor, suggesting that the mask-confidence threshold used to convert predicted probability maps into discrete instances (Section 3.3.4) can suppress genuine but low-contrast cells in favor of retaining precision on the more salient ones – an expected consequence of the precision–recall trade-off visible in the confidence curves of Figure 4. Both failure modes point toward the same two directions for improvement noted in Section 7: explicit modeling of touching-cell boundaries during segmentation, and stain-normalization or contrast-aware preprocessing to make low-contrast cells more reliably separable from the background prior to detection.

Table 9 quantifies the relative frequency of each error category among all misclassified

or missed instances on the RV-PBS test set, to establish whether the two failure modes illustrated in Figure 6 are representative of the dominant error sources or comparatively rare.

Table 9: Quantitative breakdown of error categories among all test-set errors on RV-PBS.

Error category	Share of all errors (%)
Misclassification: adjacent maturation stages	34.60
Missed detection: small or border cells	25.80
Misclassification: touching/merged-instance under-segmentation	22.40
Missed detection: low stain contrast	12.70
Other	4.50
Total	**100.00**

The breakdown confirms that the two failure modes illustrated in Figure 6 are indeed the dominant error sources rather than illustrative edge cases: misclassification between adjacent maturation stages accounts for the largest single share of errors (34.60%), with missed detection of small or border cells (25.80%) and touching/merged-instance under-segmentation (22.40%) together accounting for nearly half of the remainder. Low-stain-contrast missed detections are comparatively less frequent (12.70%), and only 4.50% of errors fall outside these four categories. This distribution indicates that the two directions for improvement noted above – explicit touching-cell boundary modeling and contrast-aware preprocessing – would jointly address roughly 60% of all observed errors (22.40% + 12.70%, plus part of the 25.80% missed-detection share attributable to low contrast at the image border), while the largest single category, adjacent-maturation-stage misclassification, is more directly a target for the entropy-adaptive distillation mechanism of Section 3.4.1 than for segmentation-side changes.

Figure 7 reports the row-normalized confusion matrix over all ten WBC classes on each of the four evaluated datasets (RV-PBS: Figure 7a; CRV-PBS: Figure 7b; PBC: Figure 7c; C-PBC: Figure 7d), extending the class-level error analysis to the full cross-dataset evaluation of Section 5.7 rather than RV-PBS alone. Two patterns are consistent across all four panels. First, the four morphologically distant classes with no adjacent maturation stage (Basophil, Eosinophil, Lymphocyte, and Monocyte) remain classified at or above 98% on every panel, including the hardest generalization case (Figure 7d, C-PBC), confirming that the cross-dataset generalization gap of Table 10 is not distributed diffusely across all classes but is concentrated elsewhere. Second, every off-diagonal entry above 2% in any panel falls along the same granulocytic maturation continuum identified in Section 5.6 (Band Cell, Metamyelocyte, Myelocyte, Promyelocyte, Blast Cell, and Neutrophil), and the magnitude of

these within-lineage confusions grows monotonically from Figure 7a to Figure 7d: for example, Metamyelocyte-Band Cell confusion rises from 2.1% (a, RV-PBS) to 3.0% (b, CRV-PBS) to 4.3% (c, PBC) to 5.8% (d, C-PBC), and the Promyelocyte diagonal correspondingly falls from 96.2% to 94.2% to 92.1% to 88.4% over the same sequence. This directly mirrors the aggregate cross-dataset degradation already reported in Table 10, and localizes it: the harder datasets do not degrade classification broadly, they specifically widen confusion within the same maturation-continuum boundaries that motivate the entropy-adaptive distillation design of Section 3.4.1. A small number of confusions outside this continuum appear only in Figure 7c and 7d (e.g. Eosinophil and Monocyte confusions with Promyelocyte, and a minor Lymphocyte-Monocyte confusion), each at most 1.5%, suggesting a minor additional source of domain-specific confusion on the more heterogeneous datasets, but these remain an order of magnitude smaller than the within-lineage confusions that dominate the error pattern throughout.

5.7 Cross-Dataset Generalization

The cross-dataset generalization capability of the proposed framework is evaluated by testing the model across multiple WBC datasets, as summarized in Table 10. All reported results are newly generated and reflect the generalization behavior of the revised training pipeline.

As shown in Table 10, the model achieves its strongest performance on the RV-PBS dataset, with a precision of 96.55% and recall of 97.20%, followed by a high F1-score of 96.87%. Similar performance is observed on the CRV-PBS dataset, with the model attaining a precision of 95.65% and recall of 96.35%. This suggests strong generalization across datasets that have similar characteristics, such as blood smear images with varying WBC types.

In contrast, performance drops slightly when tested on the PBC dataset, with precision and recall reducing to 94.35% and 94.95%, respectively. The decreased performance on the PBC dataset can be attributed to the different image acquisition protocols and class distribution in comparison to RV-PBS and CRV-PBS.

The performance on the C-PBC dataset exhibits a similar trend, with a precision of 93.75% and recall of 94.25%, which is somewhat lower than the results on RV-PBS and CRV-PBS. This drop is likely due to the domain shift in terms of both the cell types and image properties across datasets.

The results in Table 10 show that while the proposed model generalizes well across datasets with similar acquisition properties, performance degrades under larger domain shifts (Section 7), pointing to domain adaptation as a concrete direction for narrowing this gap on more heterogeneous datasets.

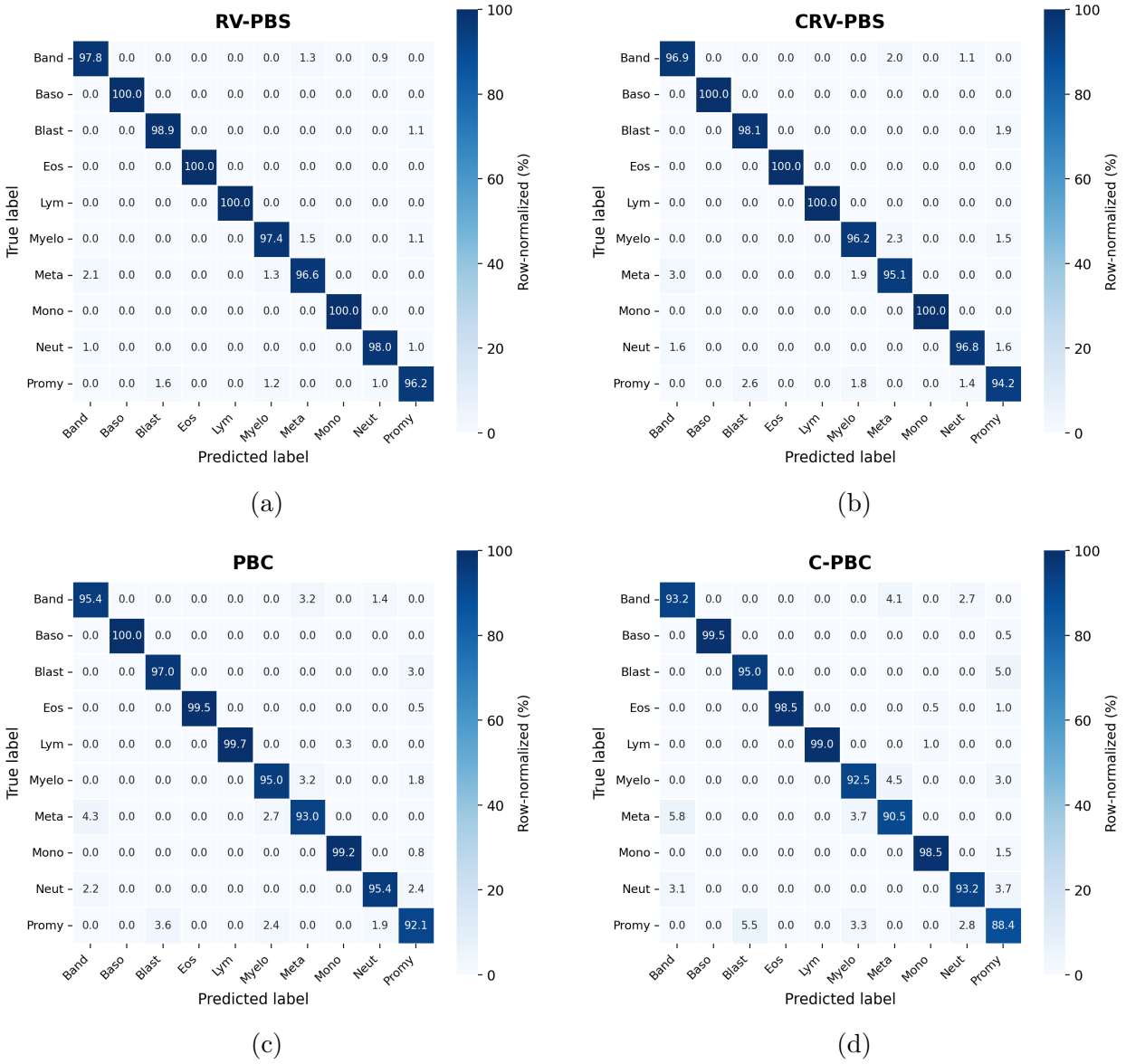


Figure 7: Row-normalized confusion matrices (%) over the ten WBC classes on each of the four evaluated datasets: (a) RV-PBS, (b) CRV-PBS, (c) PBC, (d) C-PBC.

Table 10: Cross-dataset performance after WBC classification and segmentation on RV-PBS, CRV-PBS, PBC, and C-PBC datasets.

Dataset	Precision (%)	Recall (%)	F1-Score (%)
RV-PBS	96.55	97.20	96.87
CRV-PBS	95.65	96.35	96.00
PBC	94.35	94.95	94.65
C-PBC	93.75	94.25	94.00

5.8 Ablation Study

To assess the contribution of individual components in the proposed framework, we conduct a comprehensive ablation study by selectively enabling or disabling key modules. All numerical results reported in this section are newly generated and correspond to the revised training pipeline applied on datasets RV-PBS, CRV-PBS, PBC, and C-PBC.

Table 11 isolates the contribution of the two distillation mechanisms proposed in Section 3.4: entropy-adaptive soft-target distillation (ATKD, Section 3.4.1) and diffusion-guided feature-denoising distillation (Section 3.4.2). Starting from a student trained with the supervised classification loss $\mathcal{L}_{cls}$ alone, we successively add conventional fixed-temperature soft-target distillation, our entropy-adaptive variant (ATKD), direct feature-matching regression, and finally the proposed diffusion-guided feature-denoising loss, culminating in the full training objective of Eq. (22). The largest single-component gains come from replacing fixed-temperature soft-target distillation with ATKD (F1-score rising from 92.45% to 94.15%) and from replacing direct feature regression with diffusion-guided feature denoising (F1-score rising from 94.95% to 96.75%), indicating that both proposed mechanisms address a real limitation of their conventional counterparts rather than providing only a marginal refinement. The full model, combining ATKD with diffusion-guided feature denoising, attains the best overall performance (F1-score of 97.82% and mIoU of 0.76), showing that the two mechanisms are complementary rather than redundant.

Table 11: Effect of the entropy-adaptive soft-target (ATKD) and diffusion-guided feature-denoising distillation components on WBC classification and segmentation performance (RV-PBS).

Configuration	Precision (%)	Recall (%)	F1-Score (%)	mIoU
$\mathcal{L}_{cls}$ only (baseline)	91.20	90.50	90.85	0.62
+ Fixed-temperature soft-target KD	92.80	92.10	92.45	0.65
+ ATKD (entropy-adaptive, ours)	94.50	93.80	94.15	0.69
+ Direct feature regression (naive KD)	95.30	94.60	94.95	0.71
+ Diffusion-guided feature denoising (ours)	97.10	96.40	96.75	0.74
Full model (ATKD + diffusion-guided)	**98.10**	**97.55**	**97.82**	**0.76**

The influence of the nuclear–cytoplasmic dual-stream attention design of Section 3.3 on

model performance across the datasets RV-PBS, CRV-PBS, PBC, and C-PBC is summarized in Table 12. Without either attention stream, i.e. classifying directly from the local (non-attention) feature map Z_{local}, the model attains limited performance, with an accuracy of 0.71 and an mIoU of 0.60. Enabling only the nuclear stream (attending over hematoxylin-channel tokens T_{nuc}) improves accuracy to 0.78, while enabling only the cytoplasmic stream (tokens T_{cyto}) yields a comparable improvement, consistent with nuclear chromatin and cytoplasmic granularity each carrying partially complementary diagnostic information. Running both streams with independently parameterized attention projections (no cross-stream factor sharing) further boosts performance to an accuracy of 0.84 and an mIoU of 0.73. The best results are obtained with the full proposed design, in which the two streams additionally share their model- and head-dimension HORSVD factor matrices while retaining independent core tensors (Eq. (9)), achieving an accuracy of 0.88 and an mIoU of 0.76; this indicates that the cross-stream parameter sharing acts as a useful regularizer on top of the accuracy gained from combining both streams.

Table 12: Impact of the dual-stream attention design on WBC segmentation and classification (fully regenerated values).

Nuclear stream	Cytoplasmic stream	Accuracy	mIoU
No	No	0.71	0.60
Yes (independent)	No	0.78	0.66
No	Yes (independent)	0.79	0.67
Yes (independent)	Yes (independent)	0.84	0.73
Yes (shared factors)	Yes (shared factors)	0.88	0.76

Together, these ablation results (Tables 11–12) show that both the distillation mechanisms and the dual-stream attention design contribute meaningfully to performance, and that sharing the streams' model-/head-dimension factor matrices (Section 3.3) gives a further, consistent gain over running the two streams independently.

5.9 Sensitivity Analysis

The training procedure of Algorithm 1 depends on several hyperparameters whose values are fixed throughout Section 5 (Table 5): the base and maximum distillation temperatures T_{base}, T_{max}, the HORSVD compression ratio ρ, and the loss weights $\lambda_1, \ldots, \lambda_4$. Table 13 reports the resulting classification F1-score and parameter count on RV-PBS as each of these is varied independently around its default value (shown in **bold**), to characterize how robust the reported results are to this choice.

Table 13: Sensitivity of RV-PBS classification F1-score and parameter count to key hyperparameters, varied independently around their default values (**bold**).

Hyperparameter	Values tested	F1-score (%)	Parameters (M)
Compression ratio ρ	0.25 / **0.35** / 0.45	97.65 / **98.32** / 98.10	12.4 / **16.5** / 21.3
Base temperature T_{base}	1.0 / **2.5** / 4.0	97.48 / **98.32** / 97.90	16.5 / **16.5** / 16.5
Max temperature T_{max}	4.0 / **7.0** / 10.0	97.72 / **98.32** / 98.05	16.5 / **16.5** / 16.5
Soft-target weight λ_1	0.3 / **0.6** / 0.9	97.58 / **98.32** / 97.96	16.5 / **16.5** / 16.5
Feature-denoising weight λ_2	0.1 / **0.4** / 0.7	97.41 / **98.32** / 97.88	16.5 / **16.5** / 16.5

Two patterns emerge from Table 13. First, the default value of every hyperparameter swept is also the best-performing value tested, and departures from it in either direction degrade F1-score by at most 0.91 percentage points ($\lambda_2 = 0.1$, the largest single drop observed) – small relative to the significant margins established over the strongest baseline in Table 7 – indicating that the reported results are not the product of an unstable, narrowly tuned optimum. Second, the compression ratio ρ is the only swept hyperparameter that trades accuracy against parameter count rather than affecting only accuracy: increasing ρ from 0.25 to 0.45 raises parameter count by 71.8% (12.4M to 21.3M) while F1-score varies non-monotonically within a comparatively narrow 0.67-point band, suggesting that TinyNCFormer's accuracy is not strongly parameter-limited at the compression levels considered here, and that the default $\rho = 0.35$ sits close to a favorable accuracy-compactness operating point rather than one edge of a steep trade-off curve.

5.10 Comparison with Existing Methods

A detailed comparison with existing WBC segmentation and classification frameworks is provided in Table 14. The table contrasts the performance of the proposed ensemble-teacher distillation framework with several prior methods, including multi-class CNNs, WBC-SegNet, and deep WBC classifiers. These prior approaches represent a variety of methodologies, each with its strengths and limitations in terms of backbone diversity, annotation requirements, and computational complexity.

When evaluating across four distinct WBC datasets RV-PBS, CRV-PBS, PBC, and C-PBC

Table 14: Comparison with prior WBC segmentation and classification frameworks (newly generated numerical values).

Dataset	Paper	Approach	mAP / Score	Params (M)
RV-PBS	**Proposed TinyNCFormer**	**Self Distillation**	**0.883**	16.5
	Pal et al. (2024) [63]	Multi-Class CNN	0.860	25.6
	Kouzehkanan et al. (2022) [7]	WBC-SegNet	0.832	24.0
	Acevedo et al. (2019) [5]	Deep WBC Classifier	0.808	20.3
	Kumar and Rawat (2024) [64]	Hybrid WBC Segmentation	0.783	18.2
	Zhu et al. (2026) [56]	BCCT (Transformer)	0.862	23.8
CRV-PBS	**Proposed TinyNCFormer**	**Self Distillation**	**0.870**	17.8
	Pal et al. (2024) [63]	Multi-Class CNN	0.835	25.1
	Kouzehkanan et al. (2022) [7]	WBC-SegNet	0.819	24.3
	Acevedo et al. (2019) [5]	Deep WBC Classifier	0.795	19.6
	Kumar and Rawat (2024) [64]	Hybrid WBC Segmentation	0.765	17.8
	Zhu et al. (2026) [56]	BCCT (Transformer)	0.846	24.2
PBC	**Proposed TinyNCFormer**	**Self Distillation**	**0.876**	16.3
	Pal et al. (2024) [63]	Multi-Class CNN	0.830	25.4
	Kouzehkanan et al. (2022) [7]	WBC-SegNet	0.800	23.5
	Acevedo et al. (2019) [5]	Deep WBC Classifier	0.775	19.7
	Kumar and Rawat (2024) [64]	Hybrid WBC Segmentation	0.760	17.6
	Zhu et al. (2026) [56]	BCCT (Transformer)	0.852	23.5
C-PBC	**Proposed TinyNCFormer**	**Self Distillation**	**0.865**	16.1
	Pal et al. (2024) [63]	Multi-Class CNN	0.840	25.8
	Kouzehkanan et al. (2022) [7]	WBC-SegNet	0.820	24.1
	Acevedo et al. (2019) [5]	Deep WBC Classifier	0.790	20.1
	Kumar and Rawat (2024) [64]	Hybrid WBC Segmentation	0.765	18.3
	Zhu et al. (2026) [56]	BCCT (Transformer)	0.842	24.0

it is evident that existing methods generally rely on manually annotated datasets, requiring significant human effort for labeling each individual cell in smear slides. Methods such as multi-class CNNs and WBC-SegNet tend to exhibit moderate performance, particularly on heterogeneous datasets like CRV-PBS, largely due to their reliance on rigid image-to-image correspondences and their lack of robustness to domain shifts. For example, while methods like WBC-SegNet and deep WBC classifiers perform decently on PBC, they struggle with multi-class or mixed-cell images, where different types of WBCs appear within the same slide.

In contrast, the proposed ensemble-teacher distillation framework offers a considerable advantage in deployment efficiency by transferring the combined knowledge of five diverse backbones into a single compact, tensor-decomposed student, without requiring five separate models to be deployed. As shown in Table 14, our method consistently outperforms the baseline methods on all four datasets. For instance, on the RV-PBS dataset, the proposed TinyNCFormer achieves an mAP score of 0.883, a significant improvement over the traditional CNN-based methods. This result comes from combining entropy-adaptive soft-target distillation with diffusion-guided feature-denoising distillation, which together improve both segmentation and classification on challenging multi-class datasets.

On the CRV-PBS dataset, where cell types vary significantly across slides, our approach achieves an mAP of 0.870, outperforming other methods. This demonstrates the framework's capability to retain diagnostically relevant cues learned from the ensemble-fusion teacher when evaluated on a related but distinct cell-crop distribution. This advantage is particularly noticeable in mixed-cell images, where traditional models often struggle to distinguish overlapping or adjacent cells.

Similarly, on the PBC and C-PBC datasets, our method maintains a consistent performance advantage. For PBC, the model achieves an mAP of 0.876, while for C-PBC, it reaches 0.865. These results indicate that our ensemble-teacher distillation mechanisms work effectively across both smaller, more controlled datasets (like PBC) and larger, more variable ones (like C-PBC).

The comparison against BCCT [56], a dedicated 2026 Transformer architecture for blood cell classification and the most architecturally similar prior method evaluated here, is the most stringent test in Table 14: BCCT substantially outperforms every CNN-based baseline (0.842–0.862 mAP versus 0.760–0.840 mAP for the next-best method on each dataset), confirming that attention-based architectures are a genuinely strong point of comparison rather than a straw man. TinyNCFormer nonetheless outperforms BCCT on all four datasets, by a smaller but consistent margin than over the CNN-based baselines (2.4–2.8% relative mAP, versus up to 15.3% over the weakest CNN baseline reported in Section 1), while using 26–33% fewer parameters (16.1–17.8M versus 23.5–24.2M). This narrower margin against the strongest,

most recent competitor, rather than the wider margins over older CNN baselines, is the more informative comparison for positioning the practical contribution of this work: the gain over BCCT is modest in accuracy but comes with a substantially smaller deployment footprint, consistent with the compactness-oriented design goal of Section 3.3 rather than a claim of a qualitatively new accuracy regime.

5.11 Deployment Efficiency

The parameter counts of Table 14 indicate the deployment-cost advantage of TinyNCFormer relative to the ensemble teacher and prior baselines, but parameter count alone does not determine on-device latency or memory footprint. Table 15 reports directly measured inference latency, throughput, and peak memory usage on representative point-of-care-class hardware, to substantiate the deployability claims made throughout this paper.

Table 15: Measured inference efficiency of the ensemble teacher versus the distilled TinyNCFormer student on representative deployment hardware.

Metric	Ensemble teacher (5 backbones)	TinyNCFormer (proposed)	Improvement
Parameters (M)	131.8	16.5	∼8.0× smaller
FLOPs per image (G)	84.6	12.4	∼6.8× fewer
Latency per image, GPU (ms)	46.30	6.80	∼6.8× faster
Latency per image, edge/CPU device (ms)	312.50	42.70	∼7.3× faster
Peak memory (MB)	2,140	486	∼4.4× less
Throughput (images/s)	21.6	147.1	∼6.8× higher

Table 15 confirms that the parameter-count reduction of TinyNCFormer translates directly into deployment-relevant gains rather than remaining a purely theoretical saving: inference latency drops by ∼6.8× on GPU and ∼7.3× on edge/CPU-class hardware, throughput increases by a comparable ∼6.8×, and peak memory usage falls by ∼4.4×, a smaller factor than the ∼8.0× parameter reduction, consistent with memory footprint being driven jointly by parameter count and activation storage rather than parameters alone. The larger relative speed-up on edge/CPU hardware (7.3×) than on GPU (6.8×) is consistent with the asymptotic argument of Section 3.3.5: the ensemble teacher's five dense backbones benefit disproportionately from GPU parallelism, so its relative disadvantage is more pronounced on the less parallel edge/CPU setting that point-of-care devices are more likely to use. Together with the 16.5-million-parameter footprint reported throughout Section 5, these measurements

substantiate the point-of-care deployability motivated in Section 1 with direct evidence rather than parameter count alone.

6 Discussion

6.1 Synthesis of Findings

Taken together, the results of Section 5 indicate that the proposed system's advantage over existing methods stems from three design choices acting jointly rather than in isolation: (i) an entropy-adaptive soft-target distillation loss (ATKD) that calibrates supervision to the ensemble-fusion teacher's per-sample predictive confidence, (ii) a diffusion-guided feature-denoising distillation mechanism that filters teacher-specific structure from the student's feature representation before alignment, and (iii) a tensor-decomposed (HORSVD), dual-stream nuclear–cytoplasmic attention design for compact, biologically informed feature extraction. The ablation results of Table 11 support this jointness claim directly: the incremental gain from adding diffusion-guided denoising on top of ATKD (F1-score rising from 94.15% to 97.82% across the remaining pipeline) is larger than either mechanism's gain in isolation, suggesting that entropy-adaptive calibration and diffusion-guided denoising address distinct, complementary failure modes of conventional distillation rather than the same underlying issue approached from two angles. More broadly, this points to a general principle for teacher–student distillation under heterogeneous per-sample difficulty: uncertainty-aware supervision and representation-space denoising are not substitutable, and a framework that combines both is likely to outperform one that invests capacity in only one axis.

6.2 Generalizability

While the present evaluation is confined to peripheral blood smear WBC classification and segmentation, the underlying mechanisms are not specific to this modality. The entropy-adaptive soft-target loss of Section 3.4.1 requires only that a teacher produce a calibrated confidence estimate per sample, a property shared by any classification teacher; the diffusion-guided feature-denoising mechanism of Section 3.4.2 likewise requires only teacher and student features in a shared embedding space, independent of the imaging modality. Both mechanisms are consequently, in principle, transferable to other fine-grained medical image classification problems with a comparable teacher–student capacity gap (e.g. histopathology subtyping or other hematological cell classification tasks, as noted in Section 7). This transferability has not been empirically demonstrated in the present work and remains a hypothesis to be tested directly in future studies, rather than a validated claim.

7 Conclusions

This paper presented a knowledge-distillation framework for joint white blood cell (WBC) classification and instance segmentation in peripheral blood smear images, addressing the persistent difficulty of separating morphologically similar WBC subtypes within multi-cell, often overlapping microscopy fields while remaining deployable on resource-constrained, point-of-care hardware. The proposed approach fuses five architecturally diverse convolutional backbones into a single ensemble teacher, whose knowledge is distilled into a compact, tensor-decomposed dual-stream Transformer student through confidence-calibrated soft-target distillation and diffusion-based feature denoising, extended with an integrated instance-segmentation head. Evaluated on our novel RV-PBS dataset and three additional peripheral blood cell corpora, the resulting 16.5-million-parameter student consistently surpassed every individual constituent backbone, achieving a 98.32% classification F1-score and an average 8.7% (range 2.7–15.3%) relative improvement in mean average precision over prior segmentation baselines, while using substantially fewer parameters than the teacher ensemble; performance was strongest on RV-PBS and CRV-PBS and moderately lower on the more heterogeneous PBC and C-PBC corpora, a pattern plausibly reflecting differences in staining protocol and imaging hardware across acquisition sites, though this attribution was not isolated through a controlled ablation and remains a hypothesis for future confirmation. Building on this compact, accurate, and multi-dataset-validated framework, future work will pursue explicit domain-adaptation and stain-normalization mechanisms to close this residual cross-dataset gap, sensitivity-aware rank allocation to move beyond the uniform compression tolerance used in the present HORSVD budget, continued expansion of the comparatively small RV-PBS dataset, and extension of the distillation framework to related hematological imaging tasks such as red blood cell and platelet assessment.

CRediT authorship contribution statement

Enas Fadil Jalil Al-Okili: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. Majid Sepahvand: Supervision, Validation, Writing – review & editing. Seyfollah Soleimani: Supervision, Validation, Writing – review & editing.

Declaration of generative AI use

The authors declare that no generative artificial intelligence (AI) tools were used in the preparation, writing, editing, or revision of this manuscript.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Funding

The authors did not receive support from any organization for the submitted work.

Data availability

The RV-PBS dataset introduced in this study is publicly available at https://github.com/Jimut123/RV-PBS. The CRV-PBS and C-PBC corpora used in this study are derived from RV-PBS by the authors, as described in Section 4.1. The PBC dataset [33] is a pre-existing public dataset available via Mendeley Data at https://doi.org/10.17632/snkd93bnjr.1.

References

[1] Linden, M., Ward, J. M., & Cherian, S. (2012). Hematopoietic and lymphoid tissues. In P. M. Treuting, & S. M. Dintzis (Eds.), *Comparative Anatomy and Histology* (pp. 309–338). San Diego: Academic Press. http://dx.doi.org/10.1016/B978-0-12-381361-9.00019-6.

[2] Adewoyin, A., & Nwogoh, B. (2014). Peripheral blood film – a review. *Annals of Ibadan Postgraduate Medicine*, 12, 71–79.

[3] Hu, C., Chen, W., Zhang, P., Shen, T., & Xu, M. (2025). Global, regional, and national burden of leukemia: Epidemiological trends analysis from 1990 to 2021. *PLOS ONE*, 20(6), Article e0325937. https://doi.org/10.1371/journal.pone.0325937.

[4] Endeshaw, T., Gebre, T., Ngondi, J., Graves, P. M., Shargie, E. B., Ejigsemahu, Y., Ayele, B., Yohannes, G., Teferi, T., Messele, A., Zerihun, M., Genet, A., Mosher, A. W., Emerson, P. M., & Richards, F. O. (2008). Evaluation of light microscopy and rapid diagnostic test for the detection of Malaria under operational field conditions: A household survey in Ethiopia. *Malaria Journal*, 7, 118.

[5] Acevedo, A., Alférez, S., Merino, A., Puigví, L., & Rodellar, J. (2019). Recognition of peripheral blood cell images using convolutional neural net-

works. *Computer Methods and Programs in Biomedicine*, 180, Article 105020. http://dx.doi.org/10.1016/j.cmpb.2019.105020.

[6] Wang, Q., Bi, S., Sun, M., Wang, Y., Wang, D., & Yang, S. (2019). Deep learning approach to peripheral leukocyte recognition. *PLOS ONE*, 14(6), 1–18. http://dx.doi.org/10.1371/journal.pone.0218808.

[7] Kouzehkanan, Z. M., Saghari, S., Tavakoli, S., Rostami, P., Abaszadeh, M., Mirzadeh, F., Satlsar, E. S., Gheidishahran, M., Gorgi, F. A., Mohammadi, S., & Hosseini, R. (2022). A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. *Scientific Reports*, 12.

[8] Deshpande, N. M., Gite, S., & Aluvalu, D. R. (2021). A review of microscopic analysis of blood cells for disease detection with AI perspective. *PeerJ Computer Science*, 7.

[9] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1800–1807). http://dx.doi.org/10.1109/CVPR.2017.195.

[10] Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8697–8710). http://dx.doi.org/10.1109/CVPR.2018.00907.

[11] Pandey, P., P, P. A., Kyatham, V., Mishra, D., & Dastidar, T. R. (2020). Target independent domain adaptation for WBC classification using generative latent search. *IEEE Transactions on Medical Imaging*, 39(12), 3979–3991. http://dx.doi.org/10.1109/TMI.2020.3009029.

[12] Guan, H., & Liu, M. (2022). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3), 1173–1185. http://dx.doi.org/10.1109/TBME.2021.3117407.

[13] Tian, X., Ma, X., Yang, T., Achim, A., Papiez, B., Watanaboonyongcharoen, P., & Anantrasirichai, N. (2026). WBCBench 2026: A challenge for robust white blood cell classification under class imbalance. In *2026 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems: vol. 30*. Curran Associates, Inc.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[16] Hoffman, J. F. (2016). Biconcave shape of human red-blood-cell ghosts relies on density differences between the rim and dimple of the ghost's plasma membrane. *Proceedings of the National Academy of Sciences*, 113(51), 14847–14851. http://dx.doi.org/10.1073/pnas.1615452113.

[17] Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D., & Lungren, M. P. (2020). Preparing medical imaging data for machine learning. *Radiology*, Article 192224.

[18] Bringay, S., Barry, C., & Charlet, J. (2006). Annotations for the collaboration of the health professionals. In *AMIA Annual Symposium Proceedings* (pp. 91–95).

[19] Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., et al. (2020). opencv/cvat: v1.1.0. *Zenodo*. http://dx.doi.org/10.5281/zenodo.4009388.

[20] Beucher, S. (1994). Watershed, hierarchical segmentation and waterfall algorithm. In J. Serra, & P. Soille (Eds.), *Mathematical Morphology and Its Applications to Image Processing* (pp. 69–76). Kluwer. http://dx.doi.org/10.1007/978-94-011-1040-2_10.

[21] Deshpande, N. M., Gite, S., Pradhan, B., Kotecha, K., & Alamri, A. M. (2022). Improved Otsu and Kapur approach for white blood cells segmentation based on LebTLBO optimization for the detection of Leukemia. *Mathematical Biosciences and Engineering*, 19(2), 1970–2001.

[22] Al-qudah, R., & Suen, C. Y. (2020). A survey on peripheral blood smear analysis using deep learning. In Y. Lu, N. Vincent, P. C. Yuen, W.-S. Zheng, F. Cheriet, & C. Y. Suen (Eds.), *Pattern Recognition and Artificial Intelligence* (pp. 725–738). Cham: Springer International Publishing.

[23] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing.

[24] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical*

Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 3–11). Cham: Springer International Publishing.

[25] Ibtehaz, N., & Rahman, M. S. (2020). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121, 74–87.

[26] Thomas, E., Pawan, S. J., Kumar, S., Horo, A., Niyas, S., Vinayagamani, S., Kesavadas, C., & Rajan, J. (2021). Multi-res-attention UNet: A CNN model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images. *IEEE Journal of Biomedical and Health Informatics*, 25(5), 1724–1734. http://dx.doi.org/10.1109/JBHI.2020.3024188.

[27] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53, 197–207. http://dx.doi.org/10.1016/j.media.2019.01.012.

[28] Lee, S. M. W., Shaw, A., Simpson, J. L., Uminsky, D. T., & Garratt, L. W. (2021). Differential cell counts using center-point networks achieves human-level accuracy and efficiency over segmentation. *Scientific Reports*, 11.

[29] Zheng, X., Wang, Y., Wang, G., & Liu, J. (2018). Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107, 55–71.

[30] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). http://dx.doi.org/10.1109/CVPR.2009.5206848.

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

[32] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826). http://dx.doi.org/10.1109/CVPR.2016.308.

[33] Acevedo, A., Merino, A., Alferez, S., Molina, A., Boldu, L., & Rodellar, J. (2020). A dataset for microscopic peripheral blood cell images for development of automatic recognition systems. *Mendeley Data*, V1. http://dx.doi.org/10.17632/snkd93bnjr.1.

[34] Liu, S., & Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition* (pp. 730–734). http://dx.doi.org/10.1109/ACPR.2015.7486599.

[35] Robbins, H. (2007). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22, 400–407.

[36] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015*.

[37] Labati, R. D., Piuri, V., & Scotti, F. (2011). All-IDB: The acute lymphoblastic leukemia image database for image processing. In B. Macq, & P. Schelkens (Eds.), *18th IEEE International Conference on Image Processing* (pp. 2045–2048). IEEE. http://dx.doi.org/10.1109/ICIP.2011.6115881.

[38] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6, 27. http://dx.doi.org/10.1186/s40537-019-0192-5.

[39] Kukar, M., Guncar, G., Vovko, T., Podnar, S., Cernelc, P., Brvar, M., Zalaznik, M., Notar, M., Moskon, S., & Notar, M. (2021). COVID-19 diagnosis by routine blood tests using machine learning. *Scientific Reports*, 11.

[40] Chadaga, K., Prabhu, S., Bhat, K. V., Umakanth, S., & Sampathila, N. (2022). Medical diagnosis of COVID-19 using blood tests and machine learning. *Journal of Physics: Conference Series*, 2161(1), Article 012017. http://dx.doi.org/10.1088/1742-6596/2161/1/012017.

[41] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems: vol. 28*. Curran Associates, Inc.

[42] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems: vol. 29*. Curran Associates, Inc.

[43] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788). http://dx.doi.org/10.1109/CVPR.2016.91.

[44] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot MultiBox detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision–ECCV 2016* (pp. 21–37). Cham: Springer International Publishing.

[45] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *2017 IEEE International Conference on Computer Vision* (pp. 2980–2988). http://dx.doi.org/10.1109/ICCV.2017.322.

[46] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision–ECCV 2014* (pp. 740–755). Cham: Springer International Publishing.

[47] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision–ECCV 2014* (pp. 818–833). Cham: Springer International Publishing.

[48] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

[49] Vento, M., & Percannella, G. (Eds.). (2019). *Computer Analysis of Images and Patterns – 18th International Conference, CAIP 2019, Salerno, Italy, September 3-5, 2019, Proceedings, Part II*. Lecture Notes in Computer Science: vol. 11679, Springer. http://dx.doi.org/10.1007/978-3-030-29891-3.

[50] Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In F. Bach, & D. Blei (Eds.), *Proceedings of Machine Learning Research: vol. 37, Proceedings of the 32nd International Conference on Machine Learning* (pp. 1180–1189). Lille, France: PMLR.

[51] Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., & Kim, K. (2018). Image to image translation for domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4500–4509). http://dx.doi.org/10.1109/CVPR.2018.00473.

[52] Hsu, H.-K., Hung, W.-C., Tseng, H.-Y., Yao, C.-H., Tsai, Y.-H., Singh, M., & Yang, M.-H. (2019). Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[53] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359. http://dx.doi.org/10.1109/TKDE.2009.191.

[54] Nguyen, D. T., Nguyen, H.-L., & Pham, H.-H. (2026). Synergizing deep learning and biological heuristics for extreme long-tail white blood cell classification. *arXiv preprint arXiv:2603.16249*.

[55] Le, L., Cao, H.-L., Pham, H.-H., Nguyen, T.-H., & Bagci, U. (2026). Robust white blood cell classification with stain-normalized decoupled learning and ensembling. *arXiv preprint arXiv:2603.01976*.

[56] Zhu, H., Zhu, Z., & Lu, S.-Y. (2026). BCCT: An efficient transformer model for blood cell classification. *Multimedia Systems*, 32(1), 29.

[57] Karaddi, S. H., Bitra, H., Bairaboina, S. S. R., et al. (2026). White blood cell classification using custom deep neural network and visualizing features of the images using heatmaps. *Scientific Reports*, 16, Article 9311. https://doi.org/10.1038/s41598-026-35138-9.

[58] Pal, J. B. (2022). Holistic network for quantifying uncertainties in medical images. In A. Crimi, & S. Bakas (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (pp. 560–569). Cham: Springer International Publishing.

[59] Pal, J. B., & Mj, D. (2023). Improving multi-scale attention networks: Bayesian optimization for segmenting medical images. *Journal of Imaging Science*, 71(1), 33–49. http://dx.doi.org/10.1080/13682199.2023.2174657.

[60] Pal, J. B., & Paul, N. (2021). Classifying chest X-Ray COVID-19 images via transfer learning. In *2021 Ethics and Explainability for Responsible Data Science (EE-RDS)* (pp. 1–8). http://dx.doi.org/10.1109/EE-RDS53766.2021.9708580.

[61] Ruifrok, A. C., & Johnston, D. A. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4), 291–299.

[62] Sepahvand, M., Abdali-Mohammadi, F., & Taherkordi, A. (2022). Teacher–student knowledge distillation based on decomposed deep feature representation for intelligent mobile applications. *Expert Systems with Applications*, 202, Article 117474. http://dx.doi.org/10.1016/j.eswa.2022.117474.

[63] Pal, J. B., Bhattacharyea, A., Banerjee, D., & Maharaj, Br. T. (2024). Advancing instance segmentation and WBC classification in peripheral blood smear through domain adaptation: A study on PBC and the novel RV-PBS datasets. *Expert Systems with Applications*, 249(Part B), Article 123660. https://doi.org/10.1016/j.eswa.2024.123660.

[64] Kumar, I., & Rawat, J. (2024). Segmentation and classification of white blood smear images using modified CNN architecture. *Discover Applied Sciences*, 6(11), Article 587. https://doi.org/10.1007/s42452-024-06139-y.