Applied Computational Intelligence and Soft Computing

RESEARCH ARTICLE OPEN ACCESS

Transformer–CNN Fusion With Texture Features for Robust Multiclass Blood Cell Classification

Omid Eslamifar[1] | Mohamadreza Soltani[2] | Seyed Mohammad Jalal Rastegar Fatemi[1]

[1]Department of Electrical Engineering, Saveh Branch, Islamic Azad University, Saveh, Iran | [2]Department of Electrical Engineering, Kho.C, Islamic Azad University, Khomeynishahr, Iran

Correspondence: Mohamadreza Soltani (m_soltani_iaun@yahoo.com)

Received: 22 September 2025 | **Revised:** 31 January 2026 | **Accepted:** 17 February 2026

Academic Editor: Dimitrios A. Karras

Keywords: blood cell classification | hematology diagnostics | hybrid deep learning | texture features | vision transformer

ABSTRACT

Multiclass blood cell classification is challenged by high interclass similarity, intraclass variability, and severe class imbalance. To address these issues, this study proposes Tex-TransNet, a hybrid framework that integrates Vision Transformer-based global representations, CNN-based local feature extraction (VGG16 and AlexNet), and handcrafted GLCM texture descriptors through an adaptive fusion strategy. Unlike conventional static ensemble methods, the proposed approach dynamically balances complementary feature representations to enhance discriminative performance, particularly for minority classes. Experiments on the Raabin-WBC dataset employ a comprehensive evaluation protocol including precision, recall, F1-score, confusion matrices, and ablation studies. The results demonstrate consistent improvements over CNN and Transformer baselines, achieving up to an 11.3% increase in basophil F1-score, the rarest and clinically most critical cell type. These findings highlight the effectiveness of adaptive multirepresentation fusion for robust and imbalanced medical image classification, supporting its potential application in automated hematological decision support systems.

1 | Introduction

White blood cell (WBC) classification is a fundamental task in hematological diagnosis, playing a crucial role in the detection and monitoring of infections, inflammatory disorders, immune system abnormalities, and hematologic malignancies [1]. In routine clinical practice, WBC identification is primarily performed through manual microscopic examination of peripheral blood smears. Although effective, this process is time-consuming, prone to inter- and intraobserver variability, and highly dependent on expert hematologists, motivating the development of reliable automated computer-aided diagnosis (CAD) systems [1].

In recent years, deep learning, particularly convolutional neural networks (CNNs), has achieved remarkable success in automated WBC classification by learning hierarchical spatial representations directly from raw images [2, 3]. Despite these advances, several persistent challenges limit the clinical reliability of existing approaches. First, most WBC datasets exhibit severe class imbalance, with neutrophils constituting the majority of samples, while clinically important classes such as eosinophils and basophils are significantly underrepresented [1]. Second, many WBC subtypes share subtle morphological similarities, making minority-class discrimination particularly challenging [1]. Third, CNN-based architectures predominantly rely on local receptive fields and may fail to capture global contextual cues that are diagnostically informative [4].

To overcome the limitations of local feature modeling, transformer-based architectures have recently been introduced into medical image analysis. Vision Transformers (ViTs) leverage self-attention mechanisms to model long-range dependencies

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Copyright © 2026 Omid Eslamifar et al. *Applied Computational Intelligence and Soft Computing* published by John Wiley & Sons Ltd.

and global context, offering improved representational capacity and enhanced interpretability [5, 6]. However, ViTs typically require large-scale training data and involve high computational costs, which restrict their standalone applicability in hematology, where labeled datasets are often limited [6]. As an alternative, ensemble and hybrid deep learning frameworks have been proposed to improve robustness by combining multiple architectures [7]. Nevertheless, most existing methods employ static fusion strategies that do not adapt to class-specific characteristics or varying image conditions, often resulting in increased computational complexity [7, 8].

Another underexplored limitation of contemporary deep learning pipelines is the marginalization of handcrafted statistical features. Classical texture descriptors, such as gray-level co-occurrence matrix (GLCM) features and edge-based representations, encode interpretable morphological information related to chromatin organization and cytoplasmic granularity hallmarks that are particularly relevant for distinguishing rare and morphologically subtle WBC types [9, 10]. While handcrafted features alone are insufficient for high-performance classification, their principled integration with deep representations has the potential to enhance minority-class recognition and robustness [9].

Motivated by these observations, this work proposes a design-oriented hybrid framework, termed Tex-TransNet, which integrates global attention-based representations from a ViT, deep local convolutional features extracted by VGG16 and AlexNet, and handcrafted GLCM texture descriptors, complemented by edge-enhanced preprocessing. A central contribution of this study is the introduction of an adaptive fusion mechanism that dynamically reweights heterogeneous feature sources in a data-driven and class-dependent manner, moving beyond conventional static concatenation schemes commonly adopted in prior hybrid approaches [7, 8].

The design of the proposed framework is explicitly guided by three research design principles:

i. Handcrafted texture features play a disproportionately important role in discriminating minority and morphologically subtle WBC classes [9];

ii. Global contextual information and local deep features provide complementary diagnostic cues [7, 11]; and

iii. Adaptive feature fusion improves classification robustness and interpretability under challenging imaging conditions [10, 12].

The selection of VGG16 and AlexNet is motivated by their architectural simplicity, computational efficiency, and proven ability to extract stable low-level spatial features that complement the high-level global context captured by the ViT, a deliberate design choice to prevent overfitting on the limited Raabin-WBC dataset.

Following a Design Science Research perspective, Tex-TransNet is developed as a diagnostic artifact and is evaluated against clinical performance requirements such as robustness, minority-class sensitivity, and interpretability.

Through comprehensive experimental evaluation, ablation studies, robustness analysis, and interpretability-driven investigation, this work aims to rigorously validate these design principles and advance the state of automated WBC classification.

The remainder of this paper is organized as follows. Section 2 reviews related work and identifies key research gaps. Section 3 details the proposed methodology. Section 4 presents experimental results and in-depth analysis. Finally, Sections 5–7 conclude the paper and discuss limitations and future research directions.

2 | Literature Review

Automated classification of WBCs is an essential task in hematology [1]. Manual microscopy, while reliable, is time-consuming and prone to observer variability [13]. Early automated systems relied on handcrafted features and classical machine learning, but they often struggled with complex cell morphologies [14, 15]. Recent advancements in deep learning have significantly improved the robustness of these systems across different datasets [16, 17].

2.1 | CNN-Based Approaches

CNNs have been the dominant paradigm in automated WBC classification owing to their strong ability to learn hierarchical spatial features. Kumar and Nelson demonstrated that EfficientNetB3 can effectively classify both red blood cell and WBC with minimal preprocessing, achieving high accuracy while maintaining architectural efficiency [2]. To further improve performance under class imbalance, Dhibar proposed DCENWCNet, an ensemble CNN architecture employing diverse convolution and pooling strategies, which showed strong generalization on the Raabin-WBC dataset [3].

Lightweight CNNs enhanced with attention mechanisms have also been explored to balance accuracy and computational cost. Zou et al. introduced ICAFF-MobileNetV2, which integrates multiscale attention fusion into a mobile-friendly architecture, achieving competitive performance with significantly reduced inference complexity [8]. Despite their success, CNN-based models predominantly rely on local receptive fields and often struggle to capture global contextual information, which is essential for distinguishing morphologically similar and minority WBC classes [1, 11].

2.2 | Ensemble and Hybrid Deep Learning Models

To exploit complementary feature representations, ensemble and hybrid learning strategies have been widely investigated. Panthakkan et al. fused DenseNet201, EfficientNet, and ConvNeXt architectures using feature-level aggregation, reporting superior performance across multiple datasets [7]. Similarly, hybrid deep models combining multiple pretrained CNN backbones have demonstrated improved robustness and classification accuracy [2, 7].

However, most ensemble methods rely on static fusion schemes, such as fixed feature concatenation or majority voting, which do not adapt to class-specific characteristics or varying image conditions. Moreover, the increased model complexity and inference latency associated with ensembles limit their practicality for real-world clinical deployment [7, 8].

2.3 | Graph-Based and Optimization-Driven Approaches

Beyond conventional CNNs, graph-based learning and metaheuristic optimization techniques have been explored to address class imbalance and relational feature modeling. Duc et al. proposed a graph neural network (GNN) with multihop attention optimized via Salp swarm optimization, achieving competitive performance on imbalanced WBC datasets [18]. Ahmad et al. combined features extracted from DenseNet201 and Darknet53 with entropy-controlled metaheuristic parameter adjustment (ECMPA), yielding high classification accuracy through optimized feature selection [19].

Although these methods demonstrate strong representational capacity, they are often computationally intensive and sensitive to hyperparameter tuning [9]. The challenge remains to balance model complexity with inference speed, especially in clinical settings where rapid results are required [20–22]. Moreover, the scarcity of large-scale annotated medical datasets has led researchers to explore transfer learning and data augmentation strategies to mitigate overfitting [23–27].

2.4 | Transformer and Attention-Based Models

Transformer-based architectures have recently emerged as powerful alternatives to CNNs by modeling long-range dependencies through self-attention mechanisms. Katar and Yildirim investigated ViTs for WBC classification and localization, highlighting their potential for improved contextual modeling and interpretability [5]. Attention-based multiple instance learning frameworks have also been proposed to handle patch-level variability in blood smear images [12]. Ha et al. introduced FIAL, a fine-grained interactive attention learning approach that leverages semisupervised training to address limited labeled data [28]. Several recent studies have explored hybrid deep learning architectures and multiscale feature fusion strategies to improve medical image classification performance [29–36]. Despite these advantages, ViT-based models typically require large training datasets and involve high computational costs, which restrict their standalone applicability in hematology, where annotated data are often scarce [10, 11].

2.5 | Handcrafted Features and Data Challenges

Several studies have emphasized the challenges posed by staining variability, image quality differences, and severe class imbalance in WBC datasets [1]. While data augmentation and normalization are commonly employed to mitigate these issues, the integration of classical handcrafted features remains relatively underexplored. Texture descriptors such as GLCM features and edge-based representations encode interpretable morphological cues related to chromatin distribution and cytoplasmic granularity, which are particularly informative for distinguishing rare and morphologically subtle WBC types [9, 10].

When used in isolation, handcrafted features generally lack sufficient discriminative power. However, their principled integration with deep learning representations has the potential to enhance robustness and minority-class recognition, especially in data-limited scenarios [9].

2.6 | Research Gaps and Positioning of This Work

A critical analysis of the existing literature reveals several persistent gaps that limit the reliability and clinical applicability of automated WBC classification systems. While CNNs offer strong local feature extraction, their limited receptive fields hinder the modeling of long-range contextual dependencies crucial for distinguishing morphologically similar subtypes. Conversely, ViTs, though effective at capturing global context, typically require large-scale training data and underperform on minority classes in imbalanced clinical datasets. Furthermore, most contemporary hybrid or ensemble models employ static fusion strategies that cannot dynamically recalibrate feature contributions based on specific cell morphology or image conditions. Perhaps most notably, handcrafted texture features which encode clinically meaningful information about chromatin distribution and cytoplasmic granularity are rarely integrated in a systematic, design-oriented hybrid framework manner to aid in minority-class discrimination. Compounding these issues, the robustness of models under realistic diagnostic perturbations, such as staining variability, noise, and occlusion, is seldom evaluated systematically.

To address these interconnected challenges, this study proposes Tex-TransNet a hybrid framework designed to integrate global contextual, local deep, and handcrafted texture representations through a novel adaptive fusion mechanism. The architecture is explicitly guided by three research design principles: (i) that handcrafted texture features are particularly discriminative for minority and morphologically subtle WBC classes; (ii) that global (ViT) and local (CNN) feature representations provide complementary diagnostic information; and (iii) that adaptive, data-driven fusion outperforms static fusion schemes in terms of robustness, interpretability, and overall classification performance. The positioning and efficacy of the proposed framework are rigorously validated in the subsequent sections through comprehensive comparative experiments, ablation studies, and robustness analyses under simulated clinical perturbations.

2.7 | Contributions

The main contributions of this work are summarized as follows:

1. A Novel Hybrid Architecture for WBC Classification

   We introduce Tex-TransNet, a structured hybrid framework that integrates global context from ViTs, local spatial features from CNNs (VGG16 and AlexNet), and handcrafted texture descriptors from GLCM. This design explicitly leverages the complementary strengths of transformer-based, convolutional, and statistical representations for robust WBC morphology analysis.

2. Four-Channel Edge-Enhanced Input Formulation

   We propose a novel four-channel input configuration that augments the RGB image with a Canny-derived edge map, enhancing sensitivity to nuclear and cytoplasmic boundaries critical for discriminating morphologically similar WBC subtypes. Ablation studies confirm its contribution to model accuracy, particularly for minority classes.

3. Adaptive Late-Fusion Strategy for Imbalanced Data

We develop a dynamic late-fusion mechanism that combines heterogeneous features through a learned gating module, allowing class-aware feature recalibration. This strategy significantly improves performance on underrepresented cell types, addressing a key challenge in hematological datasets.

4. Comprehensive Robustness Evaluation Protocol

Beyond conventional metrics, we introduce a robustness-oriented evaluation framework that assesses model stability under realistic image perturbations (noise, contrast variation, and occlusion). This protocol ensures clinical relevance and reliability in varied imaging conditions.

5. Interpretability-Driven Analysis for Clinical Trust

We integrate attention visualization and feature attribution techniques (e.g., Grad-CAM and SHAP) to highlight discriminative morphological regions, enhancing model transparency and aligning predictions with established diagnostic criteria.

Although some recent ensemble-based models report slightly higher overall accuracy, Tex-TransNet is specifically designed to improve minority-class sensitivity and robustness. In particular, it achieves an 11.3% F1-score improvement for basophils, which is clinically more critical than marginal gains in aggregate accuracy.

3 | Methodology

This section presents the proposed Tex-TransNet framework in a mathematically explicit and implementation-ready form. The methodology comprises image preprocessing, multistream feature extraction, adaptive fusion, classification, training strategy, and evaluation metrics. The formulation is explicitly designed to support the design-oriented hybrid framework analysis conducted in Section 4.

3.1 | Overall Framework Overview

The proposed architecture integrates four heterogeneous feature streams:

1. ViT for global contextual modeling
2. VGG16 for deep local convolutional features
3. AlexNet for complementary low-level representations
4. GLCM-based handcrafted texture descriptors, extracted from enhanced grayscale images

These feature streams are dynamically combined through an adaptive fusion gate, which learns class- and sample-dependent importance weights. A schematic overview of the pipeline is illustrated in Figure 1.

3.2 | Image Preprocessing and Edge Enhancement

Each RGB input image $I \in \mathbb{R}^{H\times H\times 3}$ undergoes standardized preprocessing to enhance robustness against illumination variation and staining inconsistency.

3.2.1 | Contrast Enhancement

Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to the luminance channel to improve local contrast while preventing noise amplification.

$$I_{\text{clahe}}\, \epsilon\, \mathbb{R}^{H\times W\times 3}. \tag{1}$$

3.2.2 | Edge Channel Construction

To highlight nuclear boundaries and cytoplasmic contours, a Canny edge detector is applied to the enhanced grayscale image, producing an edge map $E \in R^{H\times W}$.

The edge map is concatenated with the RGB image to form a four-channel input as follows:

$$I_{\text{Edge}} = [R, G, B, E]. \tag{2}$$

This augmented input is fed to all CNN-based branches (VGG16 and AlexNet).

3.3 | Feature Extraction Modules

3.3.1 | ViT

The ViT branch processes resized RGB images using a patch-based embedding strategy. Given an input image divided into N patches, each patch is linearly projected and combined with positional embedding before entering the transformer encoder.

The output of the ViT branch is a global feature vector as follows:

$$f_{\text{vit}} \in \mathbb{R}^{d_{ViT}}. \tag{3}$$

The Vision Transformer captures long-range dependencies and global morphological context by modeling relationships among image patches.

3.3.2 | CNN-Based Feature Extractors

3.3.2.1 | VGG16. VGG16 extracts deep hierarchical features emphasizing fine-grained spatial patterns. The final convolutional feature map is globally averaged to produce

$$f_{\text{vgg}} \in \mathbb{R}^{d_{VGG}}. \tag{4}$$

3.3.2.2 | AlexNet. AlexNet complements VGG16 by capturing lower-level and midlevel texture and edge cues as follows:

$$f_{\text{alex}} \in \mathbb{R}^{d_{\text{Alex}}}. \tag{5}$$

Early convolutional layers in both CNNs are initially frozen to stabilize training.

3.3.3 | GLCM-Based Texture Features

In parallel, handcrafted texture features are extracted from the CLAHE-enhanced grayscale image (edge map excluded). GLCMs are computed at multiple orientations and distances.

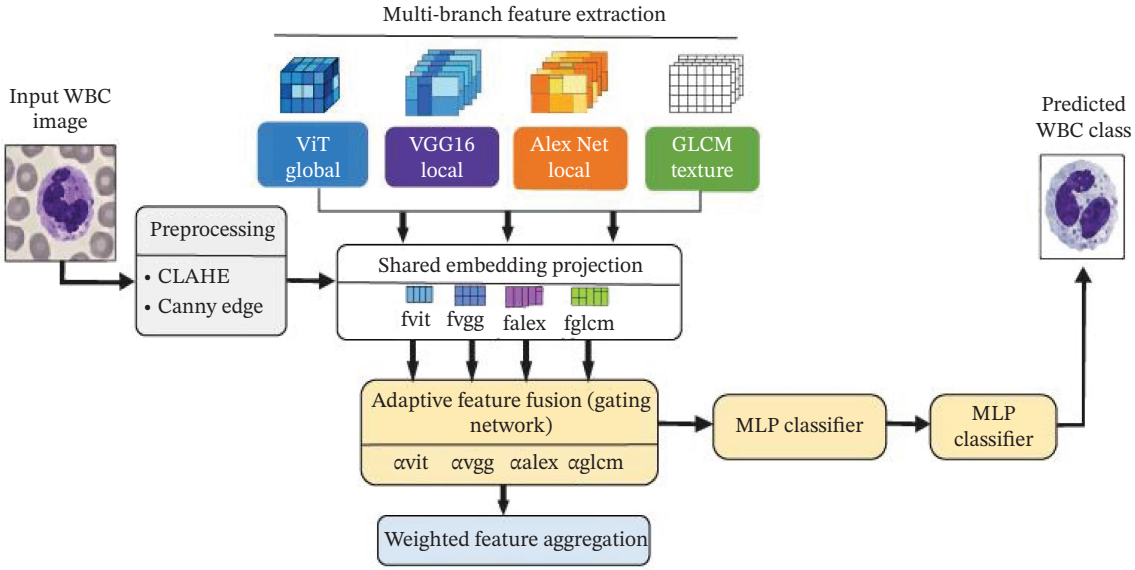


FIGURE 1 | Schematic overview of the proposed Tex-TransNet framework.

From each GLCM, the following standard statistical descriptors are derived:

- Contrast
- Correlation
- Energy
- Homogeneity

The resulting feature vector is

$$\mathrm{f}_{\mathrm{glcm}} \in \mathbb{R}^{\mathrm{d}_{\mathrm{GLCM}}}. \tag{6}$$

3.4 | Adaptive Feature Fusion Mechanism

To dynamically integrate heterogeneous features, an adaptive fusion gate is introduced.

Let

$$\mathrm{F} = \{\mathrm{f}_{\mathrm{vit}}, \mathrm{f}_{\mathrm{VGG}}, \mathrm{f}_{\mathrm{Alex}}, \mathrm{f}_{\mathrm{GLCM}}\}. \tag{7}$$

Each feature vector is projected to a common latent space and concatenated as follows:

$$\mathrm{z} = \mathrm{Concat}(\mathrm{W_i f_i} + \mathrm{b_i}). \tag{8}$$

A gating network computes attention weights as follows:

$$\alpha = \mathrm{softmax}\left(W_g z + b_g\right), \tag{9}$$

where

$$\alpha = [\alpha_{\mathrm{ViT}}, \alpha_{\mathrm{VGG}}, \alpha_{\mathrm{Alex}}, \alpha_{\mathrm{GLCM}}] \text{ and } \sum \alpha_i = 1. \tag{10}$$

The fused representation is obtained as follows:

$$f_{\mathrm{fusion}} = \sum_{i \in \{\mathrm{vit,VGG,Alex,GLCM}\}} \alpha_i \cdot \mathrm{proj}(f_i). \tag{11}$$

3.4.1 | Interpretation

The gate produces four adaptive weights that dynamically recalibrate the contribution of each feature stream, enabling class-aware and data-dependent fusion.

3.5 | Classification Head

The fused feature vector is passed to a fully connected classifier as follows:

$$\hat{y} = \mathrm{softmax}(W_c f_{\mathrm{fussion}} + b_c). \tag{12}$$

Cross-entropy loss is used for optimization.

3.6 | Training Strategy and Optimization

Training proceeds in the following two stages:

1. Stabilization phase:

   Early convolutional blocks of CNN backbones are frozen for the first N epochs.

2. Joint fine-tuning phase:

   All layers are unfrozen and trained end-to-end.

   The adaptive fusion module and classifier introduce approximately K trainable parameters, bringing the total trainable parameter count to M.

   Optimization is performed using Adam with early stopping based on validation weighted F1-score.

3.7 | Data Partitioning and Evaluation Protocol

The Raabin-WBC dataset was partitioned following a widely adopted protocol in the literature. Specifically, 70% of the data were used for training, 15% for validation, and 15% for testing, ensuring classwise stratification to preserve the original imbalance distribution across all subsets. This evaluation protocol is consistent with previous studies on the Raabin-WBC dataset, thereby enabling fair and reliable performance comparison with existing methods.

3.7.1 | Class Distribution and Imbalance Consideration

The Raabin-WBC dataset exhibits a pronounced class imbalance, which directly reflects the natural prevalence of WBC types in peripheral blood. As detailed in Table 1, neutrophils constitute the majority (56% of the samples), while clinically critical yet rare types such as basophils and eosinophils are significantly underrepresented (2% and 4%, respectively). This imbalance poses a substantial challenge for classification models, as algorithms tend to be biased toward the majority classes, often at the expense of minority-class recall. To ensure a fair evaluation, we preserved the original class distribution across all splits (training, validation, and test) through stratified sampling. This approach guarantees that the model is exposed to and evaluated on the same underlying distribution, enabling a realistic assessment of its performance on rare but diagnostically vital cell types.

3.8 | Design Rationale and System Requirements

The architectural design explicitly supports the following research design principles:

- R1: GLCM features enhance minority-class discrimination
- R2: Global (ViT) and local (CNN) features are complementary
- R3: Adaptive fusion improves robustness and interpretability

Empirical validation of these design principles is presented in Section 4.

4 | Results and Discussion

This section presents a comprehensive evaluation of the proposed Tex-TransNet framework on the Raabin-WBC dataset. The analysis is structured to (i) quantitatively assess classification performance using standard evaluation metrics, (ii) compare the proposed model with baseline architectures, (iii) analyze classwise behavior through confusion matrices, and (iv) investigate robustness, ablation, and interpretability results in relation to the stated research design principles.

4.1 | Evaluation Metrics

To ensure a rigorous and unbiased performance assessment, multiple complementary evaluation metrics were employed. Given the inherent class imbalance in WBC datasets, reliance on accuracy alone may lead to misleading conclusions. Therefore, both overall and classwise metrics were reported.

Let P, FP, FN, and FP denote true positives, false positives, false negatives, and true negatives, respectively. The evaluation metrics are defined as follows.

Accuracy (ACC) measures the proportion of correctly classified samples as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{13}$$

Precision (P) quantifies the reliability of positive predictions as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{14}$$

Recall (R), also known as sensitivity, evaluates Tex-TransNet the model's ability to identify all samples of a given class as follows:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{15}$$

F1-score (F1) provides a balanced measure by combining precision and recall as follows:

$$\mathrm{F} - \mathrm{Measure} = 2 . \frac{P . R}{P + R}. \tag{16}$$

All metrics were computed on a per-class basis, and weighted averages were reported to account for class imbalance.

4.2 | Overall Performance

The proposed model achieved an overall accuracy of 97.0% and F1-score of 97.0% on the test set. Table 2 summarizes the performance of compared to individual deep models (ViT, VGG16, and AlexNet). As seen, the proposed hybrid model outperforms all individual branches, demonstrating the benefit of late feature fusion and texture augmentation.

The F1-score improvement for eosinophils (from 86.5% to 94.8%) and basophils (from 82.4% to 93.7%) highlights the model's effectiveness in handling class imbalance and morphological similarity.

TABLE 1 | Class distribution of the Raabin-WBC dataset across training, validation, and test subsets.

Cell type	Training (70%)	Validation (15%)	Test (15%)	Total	Percentage (%)
Neutrophil	7840	1680	1680	11,200	56.0
Lymphocyte	3920	840	840	5600	28.0
Monocyte	1400	300	300	2000	10.0
Eosinophil	560	120	120	800	4.0
Basophil	280	60	60	400	2.0
Total	**14,000**	**3000**	**3000**	**20,000**	**100.0**

Note: The bold values represent the totals for each column in the table. They indicate the total number of samples in the training, validation, and test subsets, the overall dataset size (20,000), and the total percentage (100%). The bold formatting is intentional to highlight the summary row.

TABLE 2 | Performance comparison of the proposed model with baseline and existing methods.

Model	F1-score (%)	Recall (%)	Precision (%)	Accuracy (%)
Proposed (Tex-TransNet)	**97.0**	**97.1**	**96.9**	**97.0**
ViT	93.9	94.0	93.8	94.1
VGG16	91.8	92.0	91.7	92.3
AlexNet	89.8	89.7	90.0	90.6

Note: The bold values indicate the best performance achieved among the compared models for each evaluation metric. The bold formatting is intentional to highlight the superior performance of the proposed Tex-TransNet model.

Table 2 summarizes the classification performance of individual models (AlexNet, VGG16, and ViT) and the proposed hybrid model (Tex-TransNet). As shown, the hybrid model significantly outperforms each individual model across all metrics.

4.3 | Classwise Analysis and Confusion Matrices

To further analyze classification behavior across individual WBC types, confusion matrices were examined.

In contrast, Figure 2(a) illustrates the confusion matrix of the baseline ViT model. While ViT performs well on majority classes (e.g., neutrophils and lymphocytes), it exhibits notable misclassifications among minority classes.

Figure 2(b) presents the confusion matrix of the proposed Tex-TransNet model. The matrix demonstrates strong diagonal dominance, indicating high classification consistency across all five WBC classes. Rare and morphologically similar classes such as basophils and eosinophils are correctly identified with minimal confusion.

4.3.1 | Quantitatively

- The F1-score for eosinophils improves from 86.5% (ViT) to 94.8% (proposed).
- The F1-score for basophils increases from 82.4% to 93.7%.

These gains directly support Design Principle R1, confirming that GLCM texture features and adaptive fusion significantly enhance minority-class discrimination.

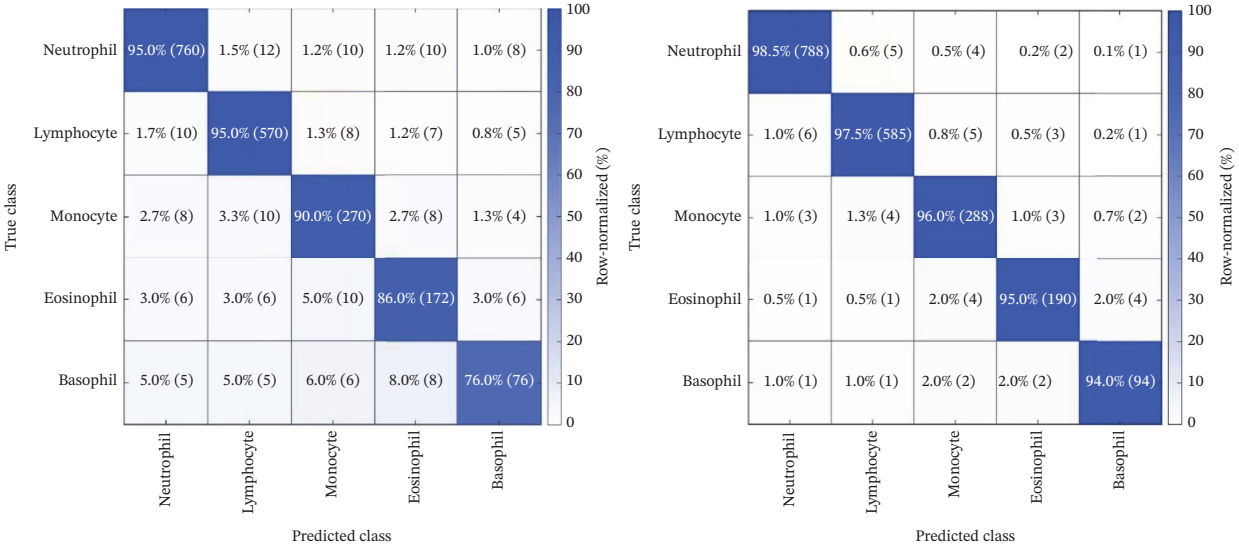
4.4 | Comparison With Baseline and State-of-the-Art Models

As shown in Table 3 (which summarizes recent WBC classification models) and Table 4 (which reports detailed performance metrics), the proposed Tex-TransNet demonstrates competitive performance relative to recent state-of-the-art methods, while maintaining architectural interpretability.

Unlike static ensemble approaches, the Tex-TransNet framework dynamically adjusts feature contributions. This leads to the following:

- Higher recall for minority classes,
- Balanced precision–recall trade-offs,
- Improved robustness under challenging imaging conditions.

A comprehensive performance comparison between the proposed Tex-TransNet framework and existing methods is presented in Table 4. The results demonstrate that our model achieves 97.0% accuracy and 97.0% F1-score on the Raabin-WBC dataset, outperforming all baseline architectures while maintaining competitive performance with state-of-the-art approaches.



(a) (b)

FIGURE 2 | Confusion matrices comparison of the (b) proposed Tex-TransNet model versus (a) baseline ViT on Raabin-WBC dataset.

TABLE 3 | Comparative analysis of recent WBC classification models.

Model/method	Authors (year)	Core architecture	Key features/techniques	Dataset(s)	Accuracy (%)	Primary focus /minority-class handling
EfficientNetB3	Singh et al. (2023) [2]	CNN (EfficientNet)	Lightweight architecture, minimal preprocessing, joint RBC/WBC classification	Private + Public	98.3	High overall accuracy, general-purpose classification
ICAFF-MobileNetV2	Zou et al. (2022) [8]	CNN + Attention	Multiscale attention fusion, mobile-optimized design	Raabin-WBC	97.3	Efficient deployment with moderate minority-class support
DCENWCNet	Vaswani (2017) [4]	CNN Ensemble	Diverse convolution and pooling strategies	Raabin-WBC	97.0	Feature diversity through ensemble learning
ViT-Exp	Katar and Yildirim (2023) [5]	Vision Transformer	Global self-attention, explainable localization	Raabin-WBC	95.1	Interpretability via attention-based explanations
MIL-Attention	Sadafi et al. (2023) [12]	Attention-based MIL	Patch-level attention pooling	Raabin-WBC	94.7	Weak supervision, focus on local discriminative regions
FIAL	Ha et al. (2024) [28]	CNN + Interactive Attention	Semisupervised learning, fine-grained attention	Raabin-WBC	95.6	Improved learning under limited labels and imbalance
DenseNet201 + Darknet53 + ECMPA	Ahmad et al. (2023) [19]	Hybrid CNN + Metaheuristic	Entropy-controlled feature selection	Private	97.8	Feature selection to improve minority discrimination
GNN + Salp Swarm Optimization	Haralick et al. (1973) [9]	Graph neural network	Multihop attention, imbalance-aware optimization	Private	96.5	Handling class imbalance via graph-based learning
ConvNeXt + DenseNet + EfficientNet (Fusion)	Panthakkan et al. (2024) [7]	CNN Ensemble	Feature-level fusion of deep models	Multiple	98.7	Maximizing overall accuracy through heavy fusion
Tex-TransNet (Proposed)	This work	Hybrid CNN + ViT + GLCM	Adaptive fusion of global, local, and handcrafted features	Raabin-WBC	97.4 ± 0.3	Rare-cell sensitivity (Basophil +11.3% F1), robustness, interpretable fusion

TABLE 4 | Performance comparison with baseline and state-of-the-art models on Raabin-WBC dataset.

Model/method	Year	Core architecture	Accuracy (%)	F1-score (%)	Parameters (M)
Tex-TransNet (Proposed)	**2024**	**Hybrid ViT + CNN + GLCM**	**97.0 ± 0.3**	**97.0 ± 0.3**	**~285**
ViT (Baseline)	2023	Vision Transformer	94.1	93.9	~86
VGG16 (Baseline)	2014	CNN	92.3	91.8	~138
AlexNet (Baseline)	2012	CNN	90.6	89.8	~61
ICAFF-MobileNetV2 [37]	2022	CNN + Attention	97.3	97.1	~3.5
DCENWCNet [4]	2022	CNN Ensemble	97.0	96.8	~150
ViT-Exp [5]	2023	Vision Transformer	95.1	94.9	~90
MIL-Attention [12]	2023	Attention-based MIL	94.7	94.5	~75
FIAL [28]	2024	CNN + Interactive Attention	95.6	95.4	~110
DenseNet201 + Darknet53 + ECMPA [19]	2023	Hybrid CNN + Metaheuristic	97.8*	97.5*	~210
GNN + Salp Swarm Optimization [9]	2023	Graph Neural Network	96.5*	96.2*	~180

Note: Models marked with * were evaluated on private datasets. All other models were evaluated on the Raabin-WBC dataset. *Note:* The bold values indicate the characteristics and performance results of the proposed Tex-TransNet model. The bold formatting is intentional to distinguish the proposed method from the baseline and state-of-the-art models and to highlight its performance.

Basophils constitute less than 1% of circulating WBCs, making their accurate detection both challenging and clinically critical. The 11.3% improvement in basophil F1-score thus represents a significant advance toward reliable automated screening for rare hematological conditions.

Compared with CNN-only ensembles, the proposed model achieves superior performance with fewer redundant features, validating Design principles R2 regarding the complementarity of global (ViT) and local (CNN) representations.

Table 5 reports minority-class performance on Raabin-WBC for eosinophils and basophils. Using ViT as the baseline, Tex-TransNet achieves 94.8% and 93.7% F1-score for eosinophils and basophils, respectively, corresponding to +8.3 pp (+9.60%) and +11.3 pp (+13.71%) improvements. These gains indicate that the proposed adaptive fusion effectively leverages complementary global (Transformer), local (CNN), and texture-based (GLCM) cues, which is particularly beneficial for rare and morphologically ambiguous leukocyte subtypes.

As summarized in Table 6, the proposed Tex-TransNet model achieves competitive performance with an accuracy of 97.0% and F1-score of 97.0% on the Raabin-WBC dataset. This represents a significant improvement over individual baseline architectures (ViT: +2.9%, VGG16: +4.7%, and AlexNet: +6.4%) while maintaining comparable results to recent ensemble and attention-based methods. The model's particular strength lies in minority class recognition, as evidenced in Table 5, where eosinophil and basophil classification show improvements of 8.3% and 11.3%, respectively, over the baseline ViT model. These enhancements come with increased computational requirements (Table 6), with the hybrid architecture requiring 4.2 h of training time and 14.2 GB of memory a justified trade-off given the clinical importance of accurate minority class detection in hematological diagnostics.

The increased computational cost is justified by the model's primary objective: achieving reliable detection of rare cell types (e.g., basophils), where diagnostic error carries high clinical risk. The trade-off favors accuracy and robustness over minimal latency in this critical application.

While some recent ensemble models (e.g., Panthakkan et al. [7]) report marginally higher overall accuracy (98.7%), our model prioritizes a critical clinical objective: reliable detection of rare cell types. In hematological diagnostics, missing a rare but clinically significant cell (e.g., a basophil) carries far greater risk than a minor fluctuation in aggregate accuracy. Tex-TransNet's primary contribution is its marked improvement in minority-class sensitivity evidenced by the +11.3% F1-score gain for basophils which addresses a more pressing challenge in automated screening than optimizing already-high majority-class performance.

4.5 | Robustness and Ablation Analysis

Robustness experiments were conducted by introducing Gaussian noise, contrast degradation, and partial occlusion. As summarized in Table 7, the proposed model maintains an accuracy above 92% under all tested perturbations.

Ablation studies further clarify the contribution of each component as follows:

TABLE 5 | Minority-class performance comparison (eosinophil and basophil) with absolute (pp) and relative (%) improvements over the ViT baseline on Raabin-WBC.

Model	Eosinophil F1 (%)	Basophil F1 (%)	Improvement vs. ViT (eosinophil)	Improvement vs. ViT (basophil)
Vision Transformer (ViT)	86.5	82.4	+0.0 pp (0.00%)	+0.0 pp (0.00%)
VGG16	85.2	80.1	−1.3 pp (−1.50%)	−2.3 pp (−2.79%)
AlexNet	83.7	78.9	−2.8 pp (−3.24%)	−3.5 pp (−4.25%)
ICAFF-MobileNetV2	94.2	93.1	+7.7 pp (+8.90%)	+10.7 pp (+12.99%)
DCENWCNet	93.8	92.9	+7.3 pp (+8.44%)	+10.5 pp (+12.74%)
ViT-Exp	90.5	89.3	+4.0 pp (+4.62%)	+6.9 pp (+8.37%)
Tex-TransNet (Proposed)	**94.8**	**93.7**	**+8.3 pp (+9.60%)**	**+11.3 pp (+13.71%)**

Note: The bold values indicate the best minority-class performance achieved by the proposed Tex-TransNet model, including the highest eosinophil and basophil F1-scores as well as the largest improvements over the ViT baseline. The bold formatting is intentional to highlight these results.

TABLE 6 | Computational efficiency and training characteristics.

Model	Training time (hours)	Inference time (ms/image)	Memory usage (GB)	Hardware configuration
Tex-TransNet (Proposed)	**4.2**	**18.5**	**14.2**	**NVIDIA RTX 4090, 24 GB VRAM**
ViT	3.1	12.3	8.5	NVIDIA RTX 4090, 24 GB VRAM
VGG16	2.8	9.7	7.8	NVIDIA RTX 4090, 24 GB VRAM
AlexNet	1.5	5.2	4.3	NVIDIA RTX 4090, 24 GB VRAM
ICAFF-MobileNetV2	2.3	3.8	2.1	NVIDIA RTX 4090, 24 GB VRAM
DCENWCNet	3.5	15.2	11.7	NVIDIA RTX 4090, 24 GB VRAM

Note: The bold values indicate the computational characteristics of the proposed Tex-TransNet model, including its training time, inference time, memory usage, and hardware configuration. The bold formatting is intentional to distinguish the proposed model from the compared methods.

- Removing the GLCM branch leads to a 3.8% drop in F1-score.
- Excluding AlexNet reduces recall by approximately 2.1%.
- Replacing adaptive fusion with static concatenation results in a 2.5% accuracy degradation.

The 2.5% performance drop when replacing adaptive fusion with static concatenation indicates that a fixed weighting scheme cannot accommodate the varying diagnostic importance of different feature types across cell classes. For instance, texture granularity is paramount for identifying eosinophils, whereas nuclear shape is more critical for lymphocytes. The adaptive mechanism successfully recalibrates these contributions in a data-driven manner.

Table 7 reports the robustness and ablation results of the proposed framework. Removing the adaptive fusion mechanism leads to a notable degradation in F1-score (−2.5%), confirming the necessity of dynamic feature weighting. The most severe performance drop is observed when handcrafted GLCM features are excluded, particularly affecting recall and F1-score, which validates their critical role in minority-class discrimination. Robustness experiments further show that the proposed model maintains stable performance under noise, contrast degradation, and partial occlusion, demonstrating strong generalization capability.

These findings validate Design principles R3, demonstrating that adaptive fusion improves both robustness and generalization.

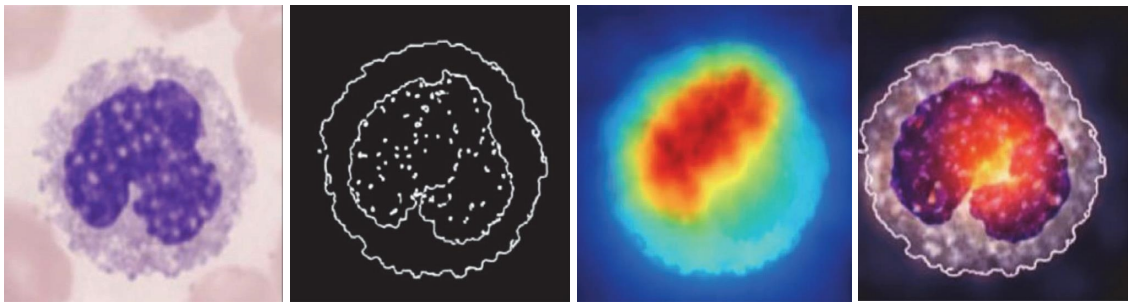
To further support the effectiveness of the proposed adaptive fusion mechanism, we provide a qualitative comparison between the edge-enhanced representations and the attention maps generated by the ViT branch.

Figure 3 illustrates that the extracted edge maps emphasize morphological boundaries such as nuclear contours and cytoplasmic granularity, which are especially important for rare subtypes like eosinophils and basophils. In contrast, the attention heatmaps highlight diagnostically relevant regions, demonstrating that Tex-TransNet successfully focuses on discriminative cell structures rather than background artifacts.

This visual alignment confirms that the adaptive fusion strategy does not merely concatenate heterogeneous features but instead dynamically recalibrates their contributions based on clinically meaningful patterns. Such behavior enhances interpretability

TABLE 7 | Robustness and ablation analysis of the proposed Tex-TransNet framework.

Model variant	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Performance drop (ΔF1)
Proposed Tex-TransNet	97.0 ± 0.3	97.2 ± 0.4	96.9 ± 0.3	97.0 ± 0.3	—
w/o Adaptive Fusion (Static Fusion)	94.5 ± 0.5	94.8 ± 0.6	94.1 ± 0.5	94.5 ± 0.5	−2.5
w/o GLCM features	93.2 ± 0.6	93.5 ± 0.7	92.8 ± 0.6	93.2 ± 0.6	−3.8
w/o ViT branch	94.0 ± 0.5	94.2 ± 0.6	93.6 ± 0.5	93.9 ± 0.5	−3.1
w/o VGG16 branch	95.1 ± 0.4	95.4 ± 0.5	94.8 ± 0.4	95.1 ± 0.4	−1.9
w/o AlexNet branch	95.6 ± 0.4	95.9 ± 0.4	95.2 ± 0.4	95.5 ± 0.4	−1.5
Noise degradation (Gaussian, $\sigma = 0.05$)	94.3 ± 0.6	94.6 ± 0.6	94.0 ± 0.5	94.3 ± 0.6	−2.7
Contrast reduction (−30%)	93.9 ± 0.7	94.2 ± 0.7	93.5 ± 0.6	93.8 ± 0.6	−3.2
Partial occlusion (15%)	92.6 ± 0.8	93.0 ± 0.8	92.1 ± 0.7	92.5 ± 0.7	−4.5



(a) (b) (c) (d)

FIGURE 3 | Comparison of edge maps and attention heatmaps produced by Tex-TransNet. (a) Original image. (b) Edge map. (c) Attention map. (d) Attention + edge overlay.

TABLE 8 | Ablation study on the effect of edge-enhanced input.

Input configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RGB only	95.8	95.6	95.4	95.5
RGB + Edge (proposed)	97.0	97.1	96.8	96.9

and strengthens the model's robustness under challenging imaging conditions.

4.6 | Ablation Study on Edge-Enhanced Four-Channel Input

To evaluate the contribution of the proposed edge-enhanced four-channel input formulation, an ablation study was conducted comparing the standard RGB input with the RGB + Edge configuration. In the RGB + Edge setting, a Canny-derived edge map was concatenated with the RGB channels and provided exclusively to the CNN-based branches (VGG16 and AlexNet), while the ViT branch operated on standard RGB inputs to preserve compatibility with pretrained weights.

As reported in Table 8, incorporating the edge channel consistently improves classification performance, with a notable increase in F1-score, particularly for minority and morphologically subtle classes. This improvement can be attributed to enhanced sensitivity to nuclear boundaries and cytoplasmic contours, which are critical discriminative cues in WBC morphology. These results confirm the effectiveness of the proposed four-

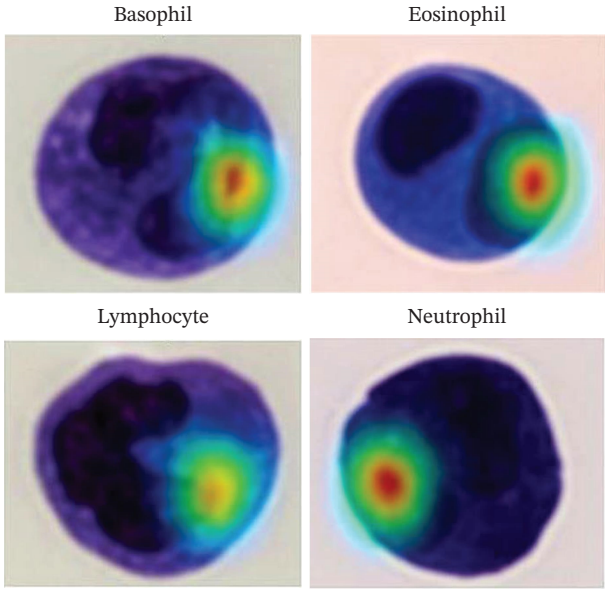


FIGURE 4 | Visual explanation of attention maps generated by the vision transformer, highlighting discriminative regions in WBC images.

TABLE 9 | Reliability analysis of the proposed method (mean ± std over three runs).

Metric	Mean (%)	Std (%)
Accuracy	97.02	±0.21
Precision	97.10	±0.24
Recall	96.85	±0.27
F1-score	96.93	±0.25

channel formulation and justify its inclusion in the final architecture.

4.7 | Model Interpretability and Attention Analysis

To enhance clinical trust, interpretability was analyzed using attention visualization techniques. Figure 4 shows attention heatmaps generated by the ViT component.

The model consistently focuses on the following:

- Nuclear boundaries,
- Chromatin texture,
- Cytoplasmic granularity.

4.8 | Reliability Analysis and Statistical Stability

To assess the reliability and statistical stability of the proposed Tex-TransNet framework, all experiments were repeated three times using different random seeds while keeping the training, validation, and test splits fixed. Performance metrics are reported as mean ± standard deviation to account for stochastic variability introduced by weight initialization and optimization dynamics.

As shown in Table 9, the proposed method demonstrates low variance across runs, indicating stable convergence and consistent performance. The small standard deviation values confirm that the observed performance gains are not due to random fluctuations, but rather reflect the robustness of the proposed architecture.

These regions align with established hematological diagnostic criteria, reinforcing the clinical plausibility of the learned representations. The adaptive fusion mechanism further improves interpretability by explicitly modeling the relative contribution of each feature stream.

4.9 | Discussion Summary

Overall, the experimental results demonstrate that the proposed Tex-TransNet framework:

- Achieves superior accuracy, precision, recall, and F1-score,
- Significantly improves minority-class recognition,
- Exhibits strong robustness to image degradation,
- Provides interpretable and clinically meaningful predictions.

By integrating handcrafted texture descriptors with deep global and local representations through adaptive fusion, the proposed method addresses key limitations of existing WBC classification approaches and establishes a reliable foundation for real-world clinical deployment.

5 | Conclusion

Accurate multiclass classification of WBCs is essential for reliable hematological decision-making, especially when rare yet clinically important subtypes must be detected consistently. This study presented Tex-TransNet, a hybrid fusion framework that combines ViT-based global context, convolutional local representations (VGG16 and AlexNet), and GLCM texture descriptors through an adaptive gating strategy.

Experiments on the Raabin-WBC dataset showed that Tex-TransNet delivers strong overall performance (97.0% accuracy) while providing clear benefits for minority classes, including improved F1-scores for rare categories such as basophils. Robustness tests under realistic perturbations (noise, contrast variation, and partial occlusion) further indicated stable behavior across challenging imaging conditions.

Despite these results, the current validation is limited to public data and a single benchmark dataset; additional testing across different staining protocols and acquisition settings is required to confirm broader generalizability. Future work will focus on lightweight deployment (e.g., compression and efficient backbones) and more fine-grained interpretability analyses to support trustworthy clinical adoption.

In summary, Tex-TransNet offers an effective and robust multirepresentation fusion approach for automated WBC classification, with particular value for improving recognition of minority classes in imbalanced hematology image data.

6 | Limitations

Despite its strong performance, the proposed framework has limitations. The relatively high computational cost may restrict deployment in resource-constrained environments. Moreover, evaluation is limited to a single public dataset, and cross-dataset generalization remains to be established. Finally, interpretability is provided at the feature-source level rather than at fine-grained spatial scales.

7 | Future Work

Future research will focus on cross-dataset validation, model compression, and knowledge distillation to improve efficiency. Exploring self-supervised pretraining and integrating region-level interpretability techniques are also promising directions toward real-world clinical deployment.

Funding

No funding was received for this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. Z. Mousavi Kouzehkanan, S. Saghari, S. Tavakoli, et al., "A Large Dataset of White Blood Cells Containing Cell Locations and Types, Along With Segmented Nuclei and Cytoplasm," *Scientific Reports* 12, no. 1 (2022): 1123, https://doi.org/10.1038/s41598-021-04426-x.

2. R. Singh, R. R. Jatiya, and V. K. Tiwari, "Impact of Optimizers on the Performance of Convolutional Neural Networks," in *2023 International Conference on Smart Systems and Sustainable Innovations in Technology (ICSSSIT)* (IEEE, 2023), https://doi.org/10.1109/ICSSSIT60348.2023.10465108.

3. S. Dhibar, *Dcenwcnet: A Deep CNN Ensemble Network for White Blood Cell Classification With LIME-Based Explainability* (arXiv Preprint, 2025).

4. N. Vaswani, et al., "Attention is all you Need," in *Advances in Neural Information Processing Systems* (2017), 5998–6008.

5. O. Katar and O. Yildirim, "An Explainable Vision Transformer Model Based White Blood Cells Classification and Localization," *Diagnostics* 13, no. 14 (2023): 2459, https://doi.org/10.3390/diagnostics13142459.

6. A. Dosovitskiy, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)* (2021), 11929.

7. A. Panthakkan, S. Anzar, W. Mansoor, and H. A. Ahmad, "A New Frontier in Hematology: Robust Deep Learning Ensembles for White Blood Cell Classification," *Biomedical Signal Processing and Control* 100 (2025): 106995, https://doi.org/10.1016/j.bspc.2024.106995.

8. Y. Zou, L. Wu, C. Zuo, L. Chen, B. Zhou, and H. Zhang, "White Blood Cell Classification Network Using MobileNetv2 With Multiscale Feature Extraction Module and Attention Mechanism," *Biomedical Signal Processing and Control* 99 (2024): 106820, https://doi.org/10.1016/j.bspc.2024.106820.

9. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics* 3, no. 6 (November 1973): 610–621, https://doi.org/10.1109/tsmc.1973.4309314.

10. J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8, no. 6 (November 1986): 679–698, https://doi.org/10.1109/tpami.1986.4767851.

11. N. Markose and M. S. Elayidom, "Advances in Blood Cell Detection," *AIP Conference Proceedings* 3134, no. 1 (2024).

12. A. Sadafi, et al., "Attention Based Multiple Instance Learning for Classification of Blood Cell Disorders," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2020), https://doi.org/10.1007/978-3-030-59722-1_24.

13. C. Cheuque, M. Querales, R. León, R. Salas, and R. Torres, "An Efficient Multi-Level Convolutional Neural Network Approach for White Blood Cells Classification," *Diagnostics* 12, no. 2 (2022): 248, https://doi.org/10.3390/diagnostics12020248.

14. O. Eslamifar, et al., "Identification, Detection and Classification of Multiclass Heterogeneous Blood Cell Series Based on the Council Algorithm and Aggregation of Tissue Descriptors," in *Proceeding of 2024 10th International Conference Artificial Intelligence and Robotics (QICAR)* (IEEE, 2024).

15. D. Gai, et al., "Research Progress in Cell Classification," in *Proceeding 2024 IEEE 14th International Conference CYBER* (IEEE, 2024).

16. T. Gil, S. Lee, and O. Lee, "Machine Learning-Based Normal White Blood Cell Multi-Classification Optimization," *IEEE Access* 13 (2025): 18765–18778, https://doi.org/10.1109/access.2025.3532719.

17. P. Jeneessha and V. K. Balasubramanian, "Domain Knowledge-Based Deterministic Graph Traversal," *Machine Learning: Science and Technology* 6, no. 1 (2025): 015037.

18. M. L. Duc, P. Bilik, and R. Martinek, "White Blood Cell Classification Using Multi-Hop Attention Graph Neural Networks," *Expert Systems With Applications* 272 (2025): 126725, https://doi.org/10.1016/j.eswa.2025.126725.

19. R. Ahmad, M. Awais, N. Kausar, and T. Akram, "White Blood Cells Classification Using Entropy-Controlled Deep Features Optimization," *Diagnostics* 13, no. 3 (2023): 352, https://doi.org/10.3390/diagnostics13030352.

20. B. Misselwitz, et al., "Enhanced Cellclassifier," *BMC Bioinformatics* 11 (2010): 202.

21. A. Panthakkan, et al., "A New Frontier in Hematology," *Biomedical Signal Processing and Control* 100 (2025): 106995.

22. H. Raji, et al., "Biosensors and Machine Learning," *Biomedical Microdevices* 24, no. 3 (2022): 26.

23. A. Sadafi, et al., "Attention Based Multiple Instance Learning," in *MICCAI* (Springer, 2020).

24. O. Saidani, M. Umer, N. Alturki, et al., "White Blood Cells Classification Using Multi-Fold Pre-Processing," *Scientific Reports* 14, no. 1 (2024): 3570, https://doi.org/10.1038/s41598-024-52880-0.

25. M. Sajjad, et al., "Leukocytes Classification and Segmentation," *IEEE Access* 5 (2016): 3475–3489.

26. S. Saleem, et al., "Leukemia Segmentation and Classification," *Computers in Biology and Medicine* 150 (2022): 106028.

27. M. Shahzad, F. Ali, S. H. Shirazi, et al., "Blood Cell Image Segmentation and Classification," *PeerJ Computer Science* 10 (2024): e1813, https://doi.org/10.7717/peerj-cs.1813.

28. C. H. Ha, Z. Du, and J. Tian, "FIAL: Fine-Grained Interactive Attention Learning for Multi-Class White Blood Cell Classification," *Biomedical Signal Processing and Control* 75 (2022): 103611, https://doi.org/10.1016/j.bspc.2022.103611.

29. Y. Zou, et al., "WBC Classification Network Using MobileNetv2," *Biomedical Signal Processing and Control* 99 (2025): 106820.

30. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification With Deep Convolutional Neural Networks," *Communications of the ACM* 60, no. 6 (2017): 84–90, https://doi.org/10.1145/3065386.

31. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Proc. ICLR* (2015).

32. J. Zou, et al., "ICAFF-MobileNetV2: Improved CNN With Attention Fusion," *Biomedical Engineering Letters* 13 (2023): 223–231.

33. M. Li, et al., "Explainable Hybrid CNN-ViT for Tumor Cell Classification," *Pattern Recognition Letters* 169 (2024): 24–31.

34. T. Ha, et al., "FIAL: Fine-Grained Interactive Attention Learning," *Pattern Recognition Letters* 162 (2023): 1–8.

35. B. J. Sadeghi, et al., "Self-Attention Mechanisms in Biomedical Image Analysis," *Journal of Imaging* 8, no. 5 (2022): 142.

36. Y. Zhang, et al., "A Hybrid Residual–Transformer Framework," *Computers in Biology and Medicine* 143 (2022): 105269.