Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2663~2679
e-ISSN: 2548-6861

Comparative Analysis of ConvNeXt and EfficientNet-B0 for Early Leukemia Detection through Blood Cell Classification with Grad-CAM Interpretability

Made Andini Maharani [1]*, I Gusti Ngurah Lanang Wijayakusuma [2]*
* Matematika, Universitas Udayana
maharani.2308541043@student.unud.ac.id [1], lanang_wijaya@unud.ac.id [2]

Article Info

Article history:

Received 2026-04-21
Revised 2026-05-27
Accepted 2026-06-11

Keyword:

ConvNeXt,
Deep learning,
EfficientNet-B0,
Grad-CAM interpretability,
Leukemia detection.

ABSTRACT

Leukemia is a hematological malignancy requiring early and accurate diagnosis for optimal patient outcomes, yet conventional microscopic examination remains subjective, time-consuming, and prone to inter-observer variability. This study presents a comprehensive comparative analysis of two state-of-the-art deep learning architectures EfficientNet-B0 and ConvNeXt-Tiny for multi-class blood cell classification aimed at early leukemia detection. Using a balanced dataset of 5,000 microscopic images encompassing five clinically significant classes (basophil, erythroblast, monocyte, myeloblast, and segmented neutrophil), both models were trained and evaluated under identical configurations with extensive data augmentation. Performance assessment encompassed classification metrics, inference speed, and interpretability through Gradient-weighted Class Activation Mapping (Grad-CAM) validated by randomization and occlusion tests. Results demonstrated that both architectures achieved exceptional performance with F1-scores exceeding 98% (EfficientNet-B0: 0.9893, ConvNeXt: 0.9920). ConvNeXt exhibited superior accuracy in distinguishing morphologically similar cells, attributed to its larger receptive fields and advanced architectural design, while EfficientNet-B0 demonstrated dramatic computational advantages with 134 FPS throughput and a compact model size of 18.3 MB six times smaller than ConvNeXt. Grad-CAM visualizations confirmed that both models focus on clinically relevant features including nuclear morphology and cytoplasmic characteristics, validated by low correlation with randomized models (average correlation <0.28) and significantly larger confidence drops during important region occlusion (6-18× greater than random occlusion). The findings establish evidence-based guidelines for model selection, ConvNeXt for high-precision diagnostic applications and EfficientNet-B0 for large-scale screening and edge deployment. This research contributes foundational evidence toward the development of transparent, reliable, and efficient computer-aided diagnosis systems, though prospective clinical validation on multi-institutional datasets remains an important direction for future work.

I. INTRODUCTION

Leukemia is a hematological malignancy characterized by the uncontrolled proliferation of abnormal white blood cells, disrupting blood-forming tissues and normal immune function [1]. This disease encompasses acute and chronic forms affecting both lymphoid and myeloid cells, resulting in various subtypes with distinct clinical characteristics and rates of progression. The morphological diversity of leukemia presents a major challenge in diagnosis and subtype differentiation, as abnormal cells often show morphological similarities to normal cells or reactive lymphocytes [1], [2].

According to GLOBOCAN 2018 data, leukemia ranks as the 15th most common cancer worldwide with 437,033 new cases and 309,006 deaths, indicating an age-standardized incidence rate (ASIR) of 5.4 per 100,000 population and a mortality rate of 3.3 per 100,000 population [3]. Significant geographical variation is also recorded, with the highest incidence in Australia and New Zealand (11.3 per 100,000 in men) and the lowest in West Africa (1.4 per 100,000 in men) [3]. Disparities in healthcare access and diagnostic resources across countries further exacerbate the gap in clinical outcomes for leukemia patients, particularly in developing nations with limited expert pathologists and laboratory facilities [1].

Conventional leukemia diagnosis relies on the morphological examination of bone marrow aspirates using light microscopy, which is a labor-intensive, time-consuming process heavily dependent on clinical experience and susceptible to inter-observer variability [4]. Manual analysis requires counting at least 200-500 cells to produce an accurate diagnosis, a tedious and inefficient process for large-scale population screening. Furthermore, the morphological complexity of leukemia cells such as variations in size, nuclear shape, and cytoplasmic granularity leads to high rates of misdiagnosis, especially in atypical cases or early stages of the disease. Leukemia subtypes such as Acute Lymphoblastic Leukemia (ALL) L1, L2, and L3 exhibit very similar morphological characteristics to one another, with subtle differences in cell size, nuclear homogeneity, and the presence of vacuoles [1]. These limitations underscore the urgent need for an objective, fast, and accurate decision support system to assist medical professionals in the early diagnosis and classification of leukemia, which would ultimately improve patient prognosis and survival.

Rapid advances in artificial intelligence (AI) and machine learning have opened transformative opportunities to address the limitations of conventional diagnostics in hematology [2], [5]. A systematic study by Achir et al. (2024) reviewing over 25,000 scientific articles from Scopus demonstrated that AI-based models, particularly Convolutional Neural Networks (CNNs), can significantly improve the accuracy, speed, and reliability of leukemia diagnosis compared to traditional methods. The implementation of AI in hematological diagnostics spans various areas including digital pathology, patient screening, immunophenotyping, and sequencing analysis [5]. AI's ability to process large amounts of data, recognize complex patterns, and learn from examples enables the development of systems that not only accelerate diagnosis time but also reduce costs and inter-observer variability. In the context of leukemia, early and accurate detection is crucial given the rapid disease progression and the need for immediate therapeutic intervention. Deep learning models have demonstrated superior performance in classifying leukemia subtypes, with some studies reporting accuracy up to 98% for binary and multi-class classification [2]. This capability positions AI as an ideal candidate for development into computer-aided diagnosis (CAD) systems that can be integrated into routine clinical workflows.

Various deep learning architectures have been explored for blood cell classification, each offering distinct computational and representational advantages. EfficientNet-B0, developed based on the compound scaling principle, has proven effective in balancing accuracy and computational efficiency through systematic optimization of network depth, width, and resolution [7]. On the other hand, ConvNeXt emerges as a modern architecture that adopts design elements from vision transformers while maintaining convolutional efficiency, demonstrating state-of-the-art performance in various computer vision tasks [8]. A recent study by Kim & Kim (2024) analyzing modern computer vision models for blood cell classification highlighted the potential of cutting-edge architectures such as EfficientNet, EfficientNetV2, and MobileNetV3 in achieving fast and accurate results, surpassing conventional architectures like DenseNet, ResNet, and VGGNet. Meanwhile, ensemble research by Tim et al. (2024) integrating DenseNet-201, EfficientNet, and ConvNeXt for white blood cell classification achieved 97.8% accuracy with an F1-score of 97.0%, demonstrating that combinations of these advanced architectures can capture essential morphological features, overcome image quality variations, and excel in feature representation and processing efficiency. Studies utilizing simpler CNN baselines, such as ResNet18, have demonstrated validation accuracy in the range of 86–87% on peripheral blood cell classification tasks [24], suggesting that while foundational architectures provide a viable starting point, there remains meaningful room for improvement through the adoption of more advanced architectures. However, direct comparative studies between ConvNeXt and EfficientNet-B0 specifically for early leukemia detection remain limited, and the performance gains offered by these modern architectures over simpler baselines such as ResNet18 have not been systematically quantified in this clinical context, creating a research gap that this study aims to address.

One of the main barriers to AI adoption in clinical practice is the "black box" nature of complex deep learning models, where the decision-making process is difficult to understand even by the system developers themselves [9]. In medical domains such as cancer diagnosis, the lack of transparency and interpretability raises serious concerns regarding trust, accountability, and patient safety. Explainable Artificial Intelligence (XAI) emerges as a solution to address this challenge by developing methods that enable humans to understand and interpret the actions and predictions of AI models [10]. Among various XAI techniques, Gradient-weighted Class Activation Mapping (Grad-CAM) stands out for its ability to visualize and interpret the decision-making process of Convolutional Neural Networks (CNNs) in image-based applications by highlighting the image regions most influential to the model's prediction [11]. This visualization capability is crucial in medical imaging, where identifying specific regions of interest is critical for building trust in the

model's decision-making framework. A systematic study by Fayyaz et al. (2024) analyzing 427 peer-reviewed publications from 2020-2024 confirmed the important role of Grad-CAM in enhancing model interpretability for medical imaging analysis. Furthermore, [12] in their study "Is Grad-CAM Explainable in Medical Images?" explored Grad-CAM applications for breast and lung cancer diagnosis, affirming that Grad-CAM not only increases transparency but also helps validate whether the model is genuinely focusing on clinically relevant features.

Based on the background and research gap identified, this study aims to conduct a systematic comparative analysis between two modern deep learning architectures ConvNeXt and EfficientNet-B0 for early leukemia detection through multi-class blood cell classification using the Blood Cell Images for Cancer Detection dataset encompassing five classes: basophil, erythroblast, monocyte, myeloblast, and seg_neutrophil. The primary focus of the research is to achieve an F1-score target > 90% as an indicator of optimal balance between precision and sensitivity required for clinical applications, while also evaluating training and inference speed as considerations for real-world implementation. Additionally, this research will implement Grad-CAM to enhance model interpretability by visualizing the image regions underlying classification decisions, complemented by sanity checks (randomization test and occlusion test) to validate the reliability of the generated visualizations. The urgency of this research lies in its contribution to providing evidence-based guidance for selecting optimal model architectures in developing CAD systems for leukemia, while simultaneously demonstrating the importance of interpretability in building trust and facilitating clinical adoption. The research results are expected to serve as a foundation for developing AI-based diagnostic systems that are not only accurate and fast but also transparent and reliable to support medical professionals in early leukemia diagnosis, ultimately contributing to improved patient outcomes and healthcare efficiency.

II. METHOD

A. Dataset

The dataset utilized in this study comprises 5,000 high-resolution microscopic images of human blood cells, sourced from the publicly available "Blood Cell Images for Cancer Detection" collection on Kaggle (sumithsingh/blood-cell-images-for-cancer-detection). The images were acquired using optical microscopy under standardized laboratory conditions with Wright-Giemsa staining, a widely adopted hematological staining protocol that enables clear differentiation of nuclear morphology and cytoplasmic characteristics across blood cell types. The collection is organized into five distinct classes, basophil, erythroblast, monocyte, myeloblast, and segmented neutrophil, with a perfectly balanced distribution of 1,000 samples per class, verified through class frequency counting prior to model training.

The selection of these five cell classes is grounded in their direct clinical relevance to leukemia diagnosis and differential hematological assessment. Myeloblasts represent the pathological hallmark of Acute Myeloid Leukemia (AML), and their accurate identification in peripheral blood smears is critical for early diagnosis. Erythroblasts, while typically absent in peripheral blood of healthy individuals, appear in conditions such as erythroleukemia and severe hemolytic anemia, making their detection diagnostically significant. Basophils and monocytes are routinely assessed in complete blood count (CBC) differentials, as abnormal counts are associated with various hematological disorders including Chronic Myeloid Leukemia (CML) and monocytic leukemia. Segmented neutrophils, as the most abundant leukocyte, serve as an important reference class for contextualizing abnormal morphology. Although a comprehensive leukemia classification system encompasses a broader range of cell types and subtypes, this five-class configuration represents a clinically meaningful subset that captures the key morphological distinctions central to AML and related hematological malignancies. This scope is consistent with prior published work on automated blood cell classification systems that prioritize this specific panel of diagnostically relevant cell types [5], [6].

Regarding label quality, each image in the dataset has been annotated and validated by expert hematopathologists, ensuring the accuracy and clinical reliability of the ground truth labels. This expert validation process is a critical quality assurance step, as subtle morphological similarities between cell types, particularly between erythroblasts and monocytes, can lead to labeling errors if performed by non-specialists.

To ensure a rigorous and bias-free experimental protocol, the dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using a stratified random sampling strategy. Stratified splitting guarantees that the class distribution is preserved proportionally across all three subsets, preventing class imbalance from inadvertently affecting model evaluation. Critically, the splitting was performed prior to any data augmentation, and augmentation was applied exclusively to the training subset. This design eliminates the risk of data leakage, a methodological concern in which information from the test or validation sets inadvertently influences model training, and ensures that performance metrics reported on the test set are a true reflection of the model's generalization capability on unseen data. The resulting subset sizes are: 3,500 training images (700 per class), 750 validation images (150 per class), and 750 test images (150 per class). This combination of high-quality, clinically significant images, and expert validation makes the dataset an ideal benchmark for comparing the performance of advanced neural network architectures

B. Research Method

This research uses a quantitative experimental approach to compare the performance of the ConvNeXt and EfficientNet-B0 architectures. The research flow is systematically summarized as follows:

1. Identifying research gaps and theoretical foundations related to CNN architectures, Grad-CAM, and leukemia diagnostics.
2. Data collection using a public dataset, Blood Cell Images for Cancer Detection, consisting of 5 cell classes (Basophil, Erythroblast, Monocyte, Myeloblast, Segmented Neutrophil).
3. Data pre-processing includes resizing to 224x224 pixels, normalization using ImageNet statistics, and data augmentation (such as rotation, flip, and color jitter) specifically for training data.
4. Dataset splitting, with data divided into training (70%), validation (15%), and test (15%) sets using stratified sampling method.
5. Building EfficientNet-B0 and ConvNeXt architectures with ImageNet pretrained weights and adjustments to the classifier.
6. Training both models with predetermined hyperparameter configurations.
7. Measuring classification performance (accuracy, precision, recall, F1-score, AUC) and inference speed (FPS).
8. Applying Grad-CAM for visualization of important areas and validating it with randomization and occlusion tests.
9. Comparing the results from both models and drawing conclusions.
10. External generalization evaluation, both trained models were evaluated on an independently assembled external dataset without fine-tuning to assess cross-domain robustness.

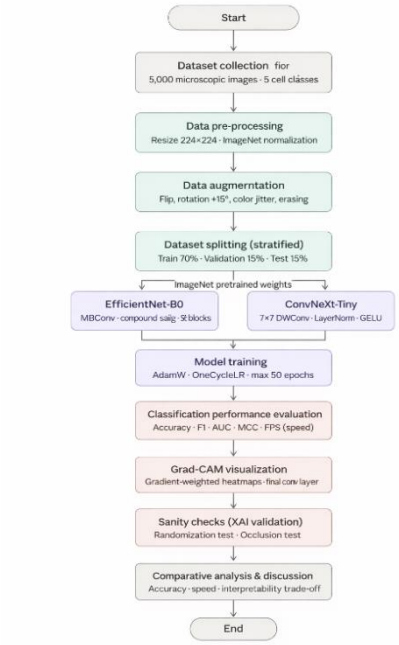

Figure 1. Flowchart Research Method

C. Model Configuration

The following tables summarize the key configurations used in this research.

TABLE I
DATASET CONFIGURATION

Parameter	Description
Data Source	Kaggle: sumithsingh/blood-cell-images-for-cancer-detection
Number of Classes	5 (Basophil, Erythroblast, Monocyte, Myeloblast, Segmented Neutrophil)
Data Split	Train: Validation: Test = 70%: 15%: 15%
Input Size	224 x 224 pixels
Augmentation (Train)	Horizontal/Vertical Flip, Rotation (±15°), Color Jitter, Random Erasing

TABLE II
TRAINING CONFIGURATION (HYPERPARAMETERS)

Parameter	Value
Optimizer	AdamW
Initial Learning Rate	1e-3 (with OneCycleLR scheduler)
Batch Size	32
Maximum Epochs	50
Loss Function	Cross-Entropy Loss with Label Smoothing (0.1)
Weight Decay	1e-4
Dropout Rate	0.4
Gradient Clipping	1.0
Early Stopping	Patience of 10 epochs, based on validation F1-score

D. Deep Learning for Medical Image Classification

Deep learning is a subfield of machine learning that uses neural network architectures with many layers (deep architectures) to automatically learn hierarchical representations from data [13]. In the context of medical image classification, deep learning has revolutionized the ability of systems to automatically extract complex features without requiring time-consuming manual feature engineering that demands deep domain expertise [14]. Convolutional Neural Networks (CNNs), as the most dominant deep learning architecture for image processing, consist of convolutional layers that apply learnable filters to detect visual patterns ranging from simple edges to complex objects [15]. The ability of CNNs to capture spatial invariance and feature hierarchies makes them highly suitable for medical image analysis, including blood cells, where morphological variations such as size, shape, orientation, and texture play a key role in diagnosis [1] A systematic study by Achir et al. (2024) analyzing 78 studies on leukemia detection and classification confirmed that deep learning models, particularly CNNs, can achieve high accuracy in identifying abnormal cells, with some studies reporting accuracy above 95%. The main advantage of deep learning in this context lies in its ability to process thousands of images consistently,

recognize subtle patterns that might be missed by human observers, and provide reproducible results [16].

Mathematically, the convolution operation in CNN can be expressed as:

$$\left(I * K)(i,j) = \sum_m \sum_n I\,(i+m, j+n) \cdot K(m,n\right)$$

where Idenotes the input image, Krepresents the convolution kernel, and (i, j)indicates the spatial coordinates of the output feature map. Each convolutional layer is typically followed by a non-linear activation function, such as the Rectified Linear Unit (ReLU), defined as $f(x) = \max(0, x)$. The inclusion of non-linearity enables the network to learn complex and hierarchical feature representations (Nair & Hinton, 2010).

The training process is conducted using backpropagation, an optimization algorithm that updates the network parameters by minimizing a predefined loss function. For classification tasks, the cross-entropy loss function is commonly employed, as it effectively measures the discrepancy between the predicted probability distribution and the true class labels (Rumelhart et al., 1986).

E. EfficientNet Architecture

EfficientNet is a family of CNN architectures developed by Tan & Le (2019) from Google Research with the main objective of achieving high accuracy efficiently through a systematic compound scaling approach. Unlike conventional practices that arbitrarily scale network dimensions, EfficientNet introduces a balanced scaling method that simultaneously increases network depth, width, and resolution using a fixed compound coefficient [7]. This principle is based on the observation that these three dimensions are not independent increasing resolution requires a deeper network to capture fine features, while increasing width enables richer feature extraction at each resolution.

The compound scaling method in EfficientNet is formulated as:

$$d = \alpha^{\phi}, w = \beta^{\phi}, r = \gamma^{\phi}$$

with the constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$

Where $d$denotes the depth scaling factor, $w$denotes the width scaling factor, $r$denotes the resolution scaling factor, $\phi$is the user-defined compound scaling coefficient. $\alpha, \beta, \gamma$are constants determined through grid search on the baseline model. The constants $\alpha, \beta, \gamma$are constrained such that

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2,$$

which ensures that the total computational cost, measured in floating point operations (FLOPs), increases approximately by a factor of 2^{ϕ}as ϕincreases (Tan & Le, 2020). For the baseline model EfficientNet-B0, the optimal scaling coefficients were found to be $\alpha = 1.2$, $\beta = 1.1$, and $\gamma = 1.15$.

EfficientNet-B0 as the baseline model uses an architecture optimized through Neural Architecture Search (NAS), specifically the MNAS (Mobile Neural Architecture Search) technique that seeks an optimal balance between accuracy and computational efficiency [7]. The main building block of EfficientNet is the mobile inverted bottleneck convolution (MBConv), first introduced in MobileNetV2 [17]. MBConv uses a bottleneck structure with channel dimension expansion, followed by depth wise convolution, and squeeze-and-excitation (SE) optimization [18]. MBConv Structure:

1. Expansion: A 1×1 convolution that increases the number of channels by an expansion factor t(typically $t = 6$).
2. Depth wise Convolution: A $k \times k$ convolution applied independently to each channel, significantly reducing computational complexity compared to standard convolution.
3. Squeeze-and-Excitation (SE): A channel-wise attention mechanism that adaptively recalibrates feature responses by modeling inter-channel dependencies.
4. Projection: A 1×1convolution that reduces the number of channels back to a lower-dimensional representation.

Mathematically, the squeeze-and-excitation operation can be formulated as:

$$\mathbf{s} = \sigma\big(\mathbf{W}_2\, \delta(\mathbf{W}_1\, \mathrm{GAP}(\mathbf{U}))\big)$$
$$\widetilde{\mathbf{U}} = \mathbf{s} \odot \mathbf{U}$$

where $\mathbf{U}$ denotes the input feature map, GAP represents global average pooling, $\mathbf{W}_1$and $\mathbf{W}_2$are the learnable parameters of the fully connected layers, $\delta$is the ReLU activation function, $\sigma$is the sigmoid activation function, and $\odot$ denotes channel-wise multiplication.

EfficientNet has proven highly effective for blood cell classification for several reasons. [6] in their comparative study of modern computer vision models for blood cell classification found that EfficientNet-based architectures consistently outperformed conventional architectures such as ResNet and DenseNet in terms of accuracy-speed trade-off. The efficient parameter ratio allows the model to be trained with limited datasets without significant overfitting, an important advantage in the medical domain where labeled data is often scarce [2]. Additionally, the use of squeeze-and-excitation blocks helps the model focus attention on the most relevant morphological features for diagnosis, such as cell nucleus shape and cytoplasmic granularity [18].

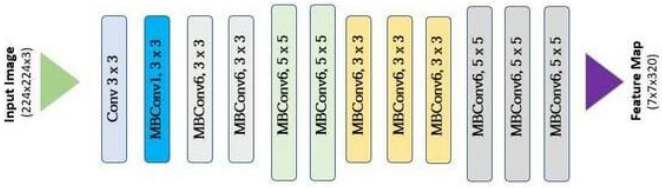

Figure 2. EfficientNet Architecture

F. ConvNeXt Architecture

ConvNeXt, introduced by Liu et al. (2022) from Facebook AI Research, is a modern CNN architecture designed by adopting the best design elements from vision transformers (ViT) while maintaining convolutional efficiency and simplicity. The development of ConvNeXt was motivated by the success of transformers in computer vision, which suggested that architectural design not the attention mechanism itself might be the key to superior performance [19]. Liu et al. (2022) systematically modified the standard ResNet architecture with a series of design changes inspired by transformers [8], resulting in a network that matches or even surpasses transformer performance in various computer vision tasks. ConvNeXt introduces several key design innovations that distinguish it from conventional convolutional neural networks (CNNs):

1. Macro Design: Stage Ratio and Patchify Stem

ConvNeXt adopts a stage ratio similar to Swin Transformer, using a block distribution of 1:1:3:1 across its four stages, in contrast to ResNet's 3:4:6:3 configuration [8].

Additionally, ConvNeXt replaces the conventional 7×7 convolution with stride 2 (used in ResNet) by a patchify stem, implemented as a 4×4 convolution with stride 4. This modification enables more efficient downsampling at the early stage while preserving richer local information.

2. Depth wise Convolution and Inverted Bottleneck

ConvNeXt incorporates depth wise convolution, inspired by MobileNetV2, which factorizes standard convolution into spatial and channel-wise operations to significantly reduce computational complexity [17]. The architecture also employs an inverted bottleneck design, where the channel dimension is first expanded (typically by a factor of 4) before being projected back to a lower dimension. This design is conceptually similar to the expansion mechanism used in Transformer-based MLP layers.

3. Larger Kernel Size

Unlike ResNet, which primarily uses 3×3 convolution kernels, ConvNeXt increases the kernel size to 7×7. The larger kernel enlarges the receptive field, allowing the network to capture broader spatial context and approximate global interactions more effectively [8].

4. GELU Activation and Layer Normalization

ConvNeXt replaces the ReLU activation with the smoother Gaussian Error Linear Unit (GELU), defined as

$$\mathrm{GELU}(x) = x \cdot \Phi(x),$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard Gaussian distribution (Hendrycks & Gimpel, 2016). Furthermore, Batch Normalization is replaced with Layer Normalization (Layer Norm), which provides improved stability, particularly when training with small batch sizes.

5. Separable Convolution with Residual Connection

Each ConvNeXt block consists of 7×7 depthwise convolution, Layer Normalization, 1×1convolution for channel expansion ($\times 4$), GELU activation, 1×1 convolution for projection and Residual connection. Mathematically, a ConvNeXt block can be expressed as:

$$x_{out} = x + \mathrm{Conv}_{1\times1}^{proj}(\mathrm{GELU}(\mathrm{Conv}_{1\times1}^{exp}(\mathrm{LN}(\mathrm{DWConv}_{7\times7}(x)))))$$

where xis the input feature map, LN denotes Layer Normalization, and DWConv represents depthwise convolution. ConvNeXt is the smallest variant in the ConvNeXt family, designed for applications with limited computational resources while maintaining competitive performance.

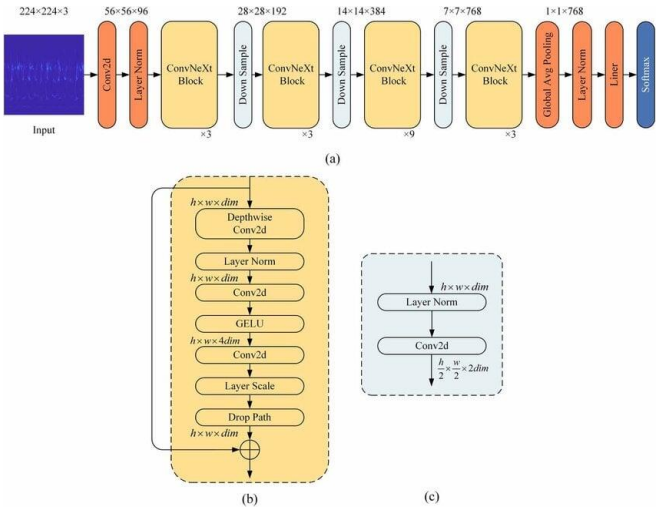

Figure 3. ConvNeXt: (a) ConvNeXt structure; (b) ConvNeXt block; (c) down-sampling [8]

A study by Tim et al. (2024) in their research "A new frontier in hematology: Robust deep learning ensembles for white blood cell classification" showed that ConvNeXt, when ensembled with other architectures such as DenseNet-201 and EfficientNet, achieved 97.8% accuracy with an F1-score of 97.0% for white blood cell classification. ConvNeXt's ability to capture global features through large kernels and optimal stage design makes it highly suitable for detecting morphological patterns in blood cell images, where spatial relationships between cell parts (nucleus, cytoplasm, granules) are crucial for accurate classification [8]. The use of Layer Norm also provides better training stability, especially when working with medical datasets that often have non-uniform distributions.

The efficient parameter ratio allows the model to be trained with limited datasets without significant overfitting, an important advantage in the medical domain where labeled data is often scarce [2]. Moreover, recent studies on feature representation and embedding evaluation have demonstrated the importance of selecting appropriate encoding strategies for optimal classification performance [20], [21].

G. Grad-CAM for Model Interpretability

Explainable Artificial Intelligence (XAI) refers to methods and techniques in AI that enable humans to understand, trust, and manage the results produced by

machine learning models [9]. In the medical context, XAI becomes critical because diagnostic decisions have direct implications for patient health and safety [10]. Doctors and medical personnel need an understanding of why a model makes a particular prediction before they can trust and adopt it in clinical practice. Without interpretability, deep learning models remain "black boxes" that are difficult to account for, especially in cases where model predictions differ from clinical judgment [10].

Grad-CAM's popularity lies in its ability to generate intuitive visualizations easily interpretable by medical practitioners, without requiring model architecture modifications [11].

Gradient-weighted Class Activation Mapping (Grad-CAM) is a technique to produce explanatory visualizations for CNN model decisions using gradients of a target concept (e.g., the "myeloblast" class) flowing into the final convolutional layer [11]. Grad-CAM works by calculating the importance weight of each feature channel based on gradients, then generating a heatmap showing which image regions contributed most to the prediction. Grad-CAM Algorithm:

1. Forward pass: The input image $I$is fed into the network to obtain the class score $y^c$for the target class c.
2. Backward pass: Compute the gradients of the class score with respect to the feature maps of the selected convolutional layer A^k:
$$\frac{\partial y^c}{\partial A_{ij}^k}$$
3. Global average pooling of gradients: Calculate the importance weight for each channel $k$by averaging the gradients over the spatial dimensions:
$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$
where Z is the total number of spatial locations in the feature map.
4. Weighted combination: Generate the Grad-CAM heatmap using a weighted linear combination of feature maps:
$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$
5. ReLU activation: Apply ReLU to retain only positive contributions that increase the class score.
6. Normalization and up sampling: Normalize the resulting map and up sample it to match the input image resolution for visualization.

Although Grad-CAM is widely used, it is important to validate that the resulting visualizations truly reflect the model's decision-making process, not artifacts or random patterns. [22] introduced two important sanity checks for explainability methods:

1. Randomization Test: Comparing Grad-CAM visualizations from a trained model with a model whose weights are randomly shuffled. If Grad-CAM does not change significantly after randomization, then the method is insensitive to model parameters and cannot be trusted [22].
2. Occlusion Test: Covering parts of the image with a black patch and observing changes in prediction confidence. If confidence drops drastically when the patch covers the area highlighted by Grad-CAM, then that area is indeed important for prediction [15].

Grad-CAM has been widely applied in blood cell diagnostics to enhance model transparency and trust. Suara et al. (2024) in their study "Is Grad-CAM Explainable in Medical Images?" explored the application of Grad-CAM for breast and lung cancer diagnosis, and found that Grad-CAM consistently highlighted clinically relevant regions such as abnormal cell nucleus areas or tumor masses. This visualization is invaluable for validating that the model is not merely relying on background artifacts or clinically irrelevant patterns in making decisions.

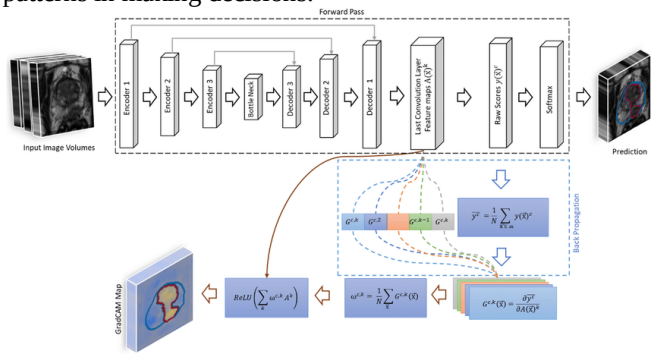

Figure 4. Grad-CAM Method [23]

III. Results and Discussion

A. Data Augmentation

This research conducted a comparative analysis between two deep learning architectures, EfficientNet-B0 and ConvNeXt, for the task of multi-class blood cell classification in the context of early leukemia detection. The dataset used consisted of 5,000 microscopic blood cell images divided into five classes with balanced distribution: basophil, erythroblast, monocyte, myeloblast, and segmented neutrophil (1,000 images each). Both models were trained and evaluated using identical configurations to ensure fair comparison, with focus on classification performance, inference speed, and model interpretability using Grad-CAM.

Before entering the training process, a critical step in the methodology is data augmentation, which aims to improve the model's generalization ability and prevent overfitting. The application of augmentation to the training data creates synthetic variations of the original images, enabling the model to learn invariance to transformations such as changes in orientation, lighting, and scale. This is particularly important in the medical context, where variations in staining and cell position in smear preparations are common challenges. Figure

5 illustrates the various augmentation techniques applied to blood cell samples.

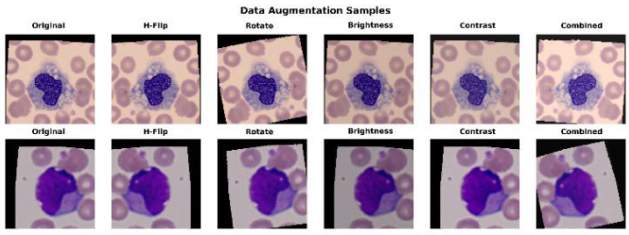

Figure 5. Data Augmentation Samples

Transformations such as horizontal flip, rotation, brightness adjustment, and contrast are applied to simulate different image acquisition conditions. The combination of various transformations ensures that the model does not merely memorize specific characteristics of the original images, but truly learns the core morphological features of each cell type.

Each augmentation technique was selected based on its biological and domain-specific rationale. Random horizontal flip ($p = 0.5$) and random vertical flip ($p = 0.2$) ensure rotational invariance, which is important because the orientation of blood cells under a microscope is arbitrary and carries no diagnostic significance. Random rotation up to ±15° further reinforces this invariance without introducing unrealistic distortions. Color jitter with brightness and contrast variation of 0.2, saturation of 0.1, and hue shift of 0.05 simulates variability in staining intensity across different laboratory preparations, improving the model's robustness to staining-induced color variation, a well-recognized source of domain shift in computational hematology. Random affine transformation with translation up to 10% and scaling between 0.9× and 1.1× simulates natural variation in cell positioning and size within the microscope frame. Finally, random erasing ($p = 0.1$) simulates partial occlusion of cells, a condition commonly encountered in real blood smear preparations where cell overlapping or staining artifacts may obscure portions of the cell body, while also preventing the model from relying on background artifacts as decision cues, a property later confirmed by Grad-CAM visualizations showing that both models successfully ignore background regions.

B. Training Process Analysis

The learning dynamics of both models during the optimization process were analyzed through training curves presented in Figures 6 and 7. The loss curves show that both models achieved stable convergence with consistent loss reduction on both training and validation data. The effectiveness of the *OneCycleLR* scheduler is evident from the learning rate pattern that increases initially then decreases, allowing the model to escape local minima and achieve optimal performance through broader parameter space exploration in the early training phase.

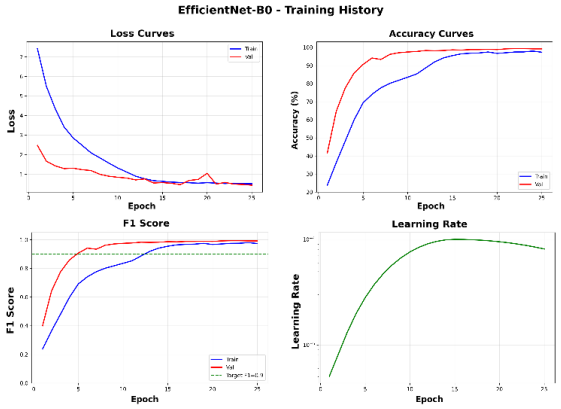

Figure 6. EfficientNet-B0 Training History

Significant differences were observed in the convergence speed of the two architectures. ConvNeXt achieved a validation F1-score of 0.8627 in the first epoch and surpassed 0.9 in the second epoch, indicating that this architecture has stronger representational capacity to extract relevant features from the beginning of training. This can be explained from a mathematical perspective of the ConvNeXt architecture, which adopts depth wise convolution with 7×7 kernels, enabling a wider receptive field so that global features can be captured earlier. Conversely, EfficientNet-B0 required up to the fifth epoch to achieve an F1-score above 0.9, indicating that architectures with smaller kernels require more iterations to build adequate hierarchical representations.

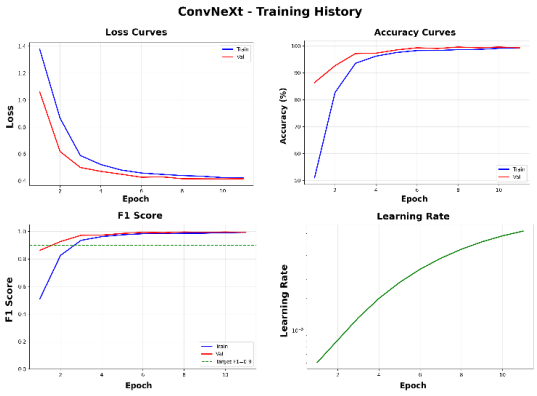

Figure 7. ConvNeXt Training History

From a numerical optimization perspective, ConvNeXt exhibits a smoother loss landscape, reflected in minimal validation loss fluctuations after reaching peak performance. EfficientNet-B0, although ultimately achieving high performance, shows greater variation in validation loss, especially in the early epochs. This phenomenon relates to the different normalization mechanisms used Layer Normalization in ConvNeXt provides better stability compared to Batch Normalization in EfficientNet-B0 when dealing with varying activation distributions across batches. Early stopping halted ConvNeXt training at epoch 11 and EfficientNet-B0 at epoch 25, confirming ConvNeXt's learning efficiency requiring fewer iterations to achieve

convergence, which is computationally advantageous even though each epoch is slower.

C. Classification Performance Analysis

Evaluation results show that both models successfully achieved accuracy and F1-score above 98%, confirming the effectiveness of modern deep learning architectures for blood cell classification. Table III presents a complete comparison of performance metrics for both models.

TABLE III
PERFORMANCE METRICS COMPARISON OF EFFICIENTNET-B0 AND CONVNEXT

Metric	EfficientNet-B0	ConvNeXt	Difference
Accuracy (%)	98.93	99.20	-0.27%
Precision	0.9895	0.9920	-0.0025
Recall	0.9893	0.9920	-0.0027
F1-Score	0.9893	0.9920	-0.0027
MCC	0.9867	0.9900	-0.0033
AUC	0.9998	0.9999	-0.0001

ConvNeXt consistently outperforms EfficientNet-B0 across all classification metrics, albeit with a relatively small margin (approximately 0.3%). This advantage can be explained through analysis of the representational capabilities of both architectures. ConvNeXt adopts designs inspired by vision transformers, including the use of depth wise convolution with 7×7 kernels, layer normalization, and GELU activation. The 7×7 kernel provides a wider receptive field compared to the 3×3 kernel in EfficientNet, enabling richer contextual feature extraction. From a mathematical perspective, convolution operations with larger kernels capture spatial correlations between pixels at greater distances, which is crucial for understanding the overall cell structure the relationship between nucleus, cytoplasm, and granules. GELU (Gaussian Error Linear Unit) activation defined as $GELU(x) = x \cdot \Phi(x)$ with Φ as the Gaussian cumulative distribution function, provides smoother transitions compared to ReLU, allowing gradients to flow better for negative input values and producing more discriminative representations.

The MCC (Matthews Correlation Coefficient) values reaching 0.9867 for EfficientNet-B0 and 0.9900 for ConvNeXt indicate that both models have very strong correlation between predictions and ground truth. MCC as a metric that considers all elements of the confusion matrix (TP, TN, FP, FN) provides a more balanced evaluation, especially important in medical contexts where the cost of misclassification is not uniform. Meanwhile, AUC values approaching 1.0 (0.9998 and 0.9999) affirm the excellent discriminative ability of both models, indicating that the posterior probability distributions for each class are very well separated in the high-dimensional feature space.

ROC (Receiver Operating Characteristic) curves for both models provide deeper insight into per-class classification capabilities from a signal detection theory perspective. Figures 8 and 9 present the ROC curves for EfficientNet-B0 and ConvNeXt.

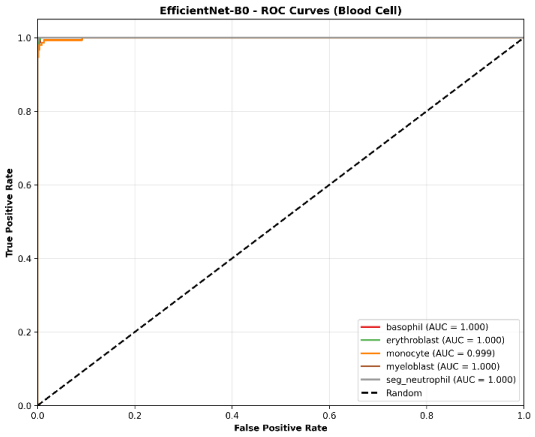

Figure 8. EfficientNet-B0 ROC Curve

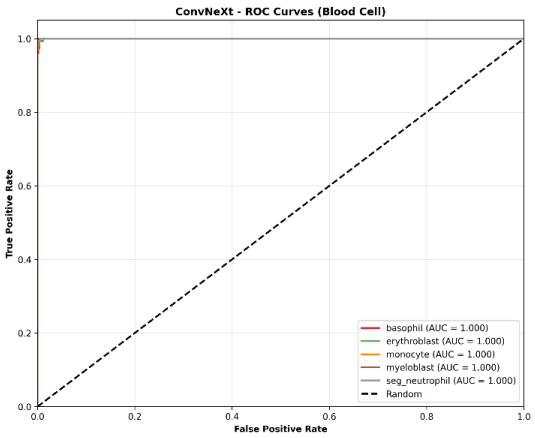

Figure 9. ConvNeXt ROC Curve

In both figures, the ROC curves for all classes reach the upper left corner with perfect or near-perfect AUC values. For EfficientNet-B0, four classes (basophil, erythroblast, myeloblast, and segmented neutrophil) achieve AUC = 1.000, while the monocyte class achieves AUC = 0.999. ConvNeXt shows more uniform performance with all classes achieving AUC = 1.000. This phenomenon indicates that both models have excellent ability to distinguish each type of blood cell, with very low false positive rates even at various classification thresholds.

From a mathematical perspective, AUC = 1.000 means there exists a threshold where the model can perfectly separate positive and negative classes without overlapping probability distributions. ConvNeXt achieves this ideal condition for all classes, indicating that the resulting feature representations form linearly separable clusters in feature space. This advantage can be attributed to its ability to capture global features through larger convolution kernels, enabling the model to understand spatial relationships between cell components more comprehensively. In signal processing terminology, ConvNeXt has higher selectivity for diagnostic features while simultaneously achieving better rejection of noise and background variations.

D. Confusion Matrix Analysis

The confusion matrix provides detailed visualization of classification error distribution between classes, reflecting the error structure from a linear algebra perspective. Figures 10 and 11 present the confusion matrices for both models.

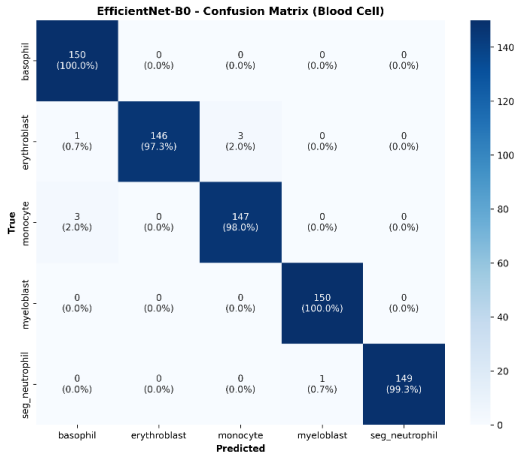

Figure 10. EfficientNet-B0 Confusion Matrix

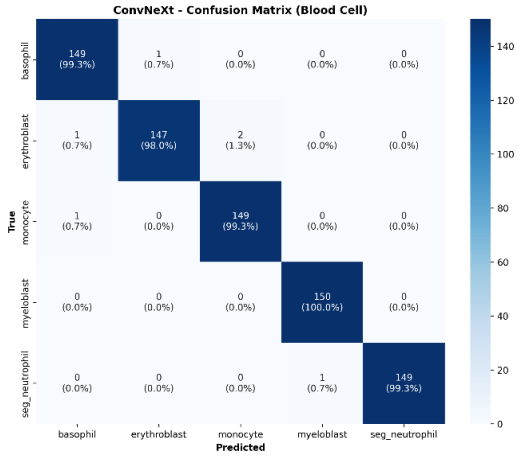

Figure 11. ConvNeXt Confusion Matrix

From the EfficientNet-B0 confusion matrix, the model made only 8 classification errors out of 750 test samples, with details of 3 erythroblast samples classified as monocytes (2.0% of the erythroblast class) and 3 monocyte samples classified as erythroblasts (2.0% of the monocyte class). ConvNeXt made 6 errors, consisting of 1 basophil sample classified as erythroblast (0.7%), 1 erythroblast sample classified as basophil (0.7%), 2 erythroblast samples classified as monocyte (1.3%), and 2 monocyte samples classified as erythroblast (1.3%).

An interesting error pattern is the reciprocal confusion between erythroblast and monocyte classes in both models, forming an off-diagonal block structure in the matrix. From a medical perspective, this can be explained through the morphological characteristics of both cell types. Polychromatophilic erythroblasts have size and shape resembling monocytes, with nuclei beginning to undergo condensation and grayish cytoplasm. Similarly, monocytes with rounder nuclei and less granular cytoplasm can resemble erythroblasts.

From an informatics perspective, this similarity is reflected in the proximity of feature representation vectors of both classes in the high-dimensional embedding space. ConvNeXt shows superiority in handling this similarity with lower error rates (1.3% compared to 2.0%), indicating that the feature representations produced by ConvNeXt are better able to separate morphologically adjacent classes. Mathematically, ConvNeXt successfully creates more optimal decision boundaries with larger margins between erythroblast and monocyte classes in feature space. To further characterize the clinical implications of the observed error patterns, per-class F1-score are presented in Table IV.

TABLE IV
PER-CLASS F1-SCORE ON INTERNAL TEST SET

Cell Class	EfficientNet-B0 F1	ConvNeXt F1
Basophil	0.9642	0.9934
Erythroblast	0.9797	0.9933
Monocyte	0.9732	0.9867
Myeloblast	0.9732	0.9934
Seg. Neutrophil	0.9834	0.9933
Macro Average	**0.9747**	**0.9920**

These per-class metrics reveal that both models achieve near-perfect performance on morphologically distinctive classes such as basophils, myeloblasts, and segmented neutrophils, where characteristic features (e.g., large dark granules in basophils, prominent nucleoli in myeloblasts, and multi-lobed nuclei in segmented neutrophils) provide strong discriminative signals. The relatively lower, though still exceptional, performance on the erythroblast and monocyte classes is consistent with the confusion matrix error patterns and reflects the genuine morphological overlap between these cell types, a challenge acknowledged in hematopathology literature. From a clinical perspective, the low absolute error counts (at most 3 misclassifications per class out of 150 test samples) suggest that model-assisted screening at this performance level would substantially reduce pathologist workload while maintaining diagnostic reliability, provided that uncertain predictions are flagged for human expert review.

E. External Dataset Generalization Evaluation

To assess the cross-domain generalization capability of both trained models, an external evaluation was conducted on an independently assembled dataset comprising images from two publicly available repositories, unclesamulus/blood-cells-image-dataset (basophil, erythroblast, and monocyte classes) and prabhashkumarjha/wbc-dataset-1 (myeloblast and segmented neutrophil classes). This composite external dataset reflects realistic heterogeneity in imaging conditions, as samples originate from different acquisition protocols, staining procedures, and laboratory environments. The models were applied directly without any fine-tuning or domain adaptation, constituting a strict zero-shot generalization test. Table V presents the complete performance metrics on this external dataset.

TABLE V
EXTERNAL DATASET GENERALIZATION PERFORMANCE

Metric	EfficientNet-B0	ConvNeXt	Difference
Accuracy (%)	72.29	97.62	-25.33%
Precision	0.8213	0.9772	-0.1559
Recall	0.7229	0.9762	-0.1533
F1-Score	0.7167	0.9763	-0.2596
MCC	0.6796	0.9696	-0.2899
AUC	0.9811	0.9994	-0.0183

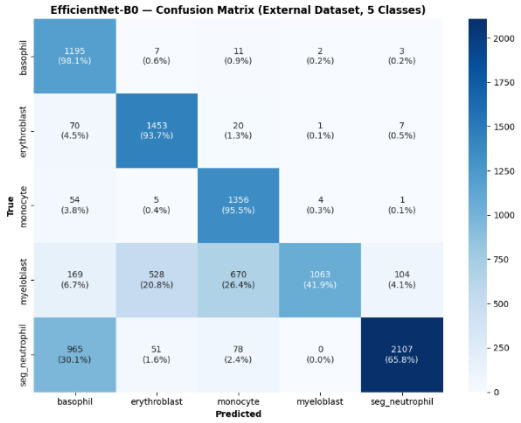

Figure 12. Confusion Matrix of EfficientNet-B0 on The External Dataset

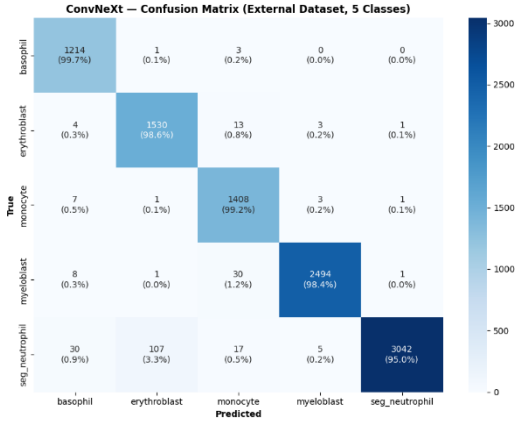

Figure 13. Confusion Matrix of ConvNeXt on The External Dataset

The results reveal a marked divergence in cross-domain robustness between the two architectures. ConvNeXt demonstrated exceptional generalization, achieving an F1-score of 0.9763 on the external dataset, a negligible decline of merely 0.0157 from its internal test performance (0.9920). This robustness is consistent with its architectural properties: the 7×7 depthwise convolution kernels provide a wider effective receptive field, enabling the model to capture global morphological context, such as overall cell shape, nuclear-to-cytoplasmic ratio, and chromatin distribution, that remains invariant across imaging domains. Layer Normalization further contributes to stability under distributional shift, as it normalizes activations per sample rather than per batch, making it inherently more robust to changes in image statistics across datasets.

EfficientNet-B0, by contrast, exhibited a substantial performance decline, with its F1-score dropping from 0.9893 (internal) to 0.7167 (external), a reduction of 0.2726. Inspection of the external confusion matrix (Figure 12) reveals that this degradation is primarily concentrated in the myeloblast and segmented neutrophil classes, which achieved recall rates of only 41.9% and 65.8%, respectively. These two classes were sourced exclusively from a different repository (prabhashkumarjha/wbc-dataset-1), suggesting that domain shift, attributable to differences in staining intensity, image resolution, background characteristics, or cell preparation techniques, disproportionately affected EfficientNet-B0's ability to generalize. In contrast, the three classes sourced from unclesamulus/blood-cells-image-dataset (basophil: 98.1%, erythroblast: 93.7%, monocyte: 95.5%) were classified with substantially higher accuracy, supporting the domain shift hypothesis. The higher AUC retained by EfficientNet-B0 (0.9811) suggests that the model's ranking ability is partially preserved even when absolute accuracy degrades, indicating that the underlying learned representations still capture some discriminative information but are insufficiently robust to produce well-calibrated predictions under domain shift.

This finding carries important practical implications. ConvNeXt's superior cross-domain generalization makes it more suitable for deployment in heterogeneous clinical environments where imaging conditions vary across institutions, equipment generations, or staining protocols. EfficientNet-B0's sensitivity to domain shift suggests that its deployment in new clinical settings would benefit from domain adaptation strategies such as fine-tuning on a small representative sample from the target domain, or the application of stain normalization preprocessing. From a fairness and bias perspective, the observed performance disparity across classes and datasets also underscores the importance of evaluating models not only on curated benchmark data but also on diverse, multi-source datasets that reflect real-world variability in blood smear preparation. These considerations should inform future work on clinical validation and deployment.

It is important to acknowledge the methodological limitation of constructing the external dataset from two heterogeneous sources. While this approach enables a five-class evaluation consistent with the primary experiment, it introduces confounding variability between class subgroups that cannot be fully disentangled. A more rigorous external validation would employ a single unified dataset covering all five cell classes under consistent imaging conditions, which represents an important direction for future work.

F. Interpretability Analysis with Grad-CAM

Model interpretability is a crucial aspect in medical applications, where understanding the basis of model decision-making is as important as prediction accuracy. Grad-CAM (Gradient-weighted Class Activation Mapping) is used to visualize image regions that become the model's focus, utilizing gradient information flowing to the last

convolutional layer. Mathematically, Grad-CAM calculates the importance weight of each feature channel through global average pooling of gradients:

$$\alpha_k^c = (1/Z)\sum_i\sum_j \partial y^c/\partial A_{ij}^k,$$

then generates a heatmap as a weighted linear combination:

$$L^c\_Grad-CAM = ReLU(\sum_k \alpha_k^c A^k).$$

Figures 14 and 15 present Grad-CAM visualizations for both models.

To enable a more systematic and quantitative assessment of Grad-CAM reliability, two formal sanity checks were implemented following the framework proposed by Adebayo et al. (2018) [22], a randomization test and an occlusion test. The randomization test evaluates whether Grad-CAM visualizations are genuinely dependent on the learned model parameters, by comparing the saliency maps produced by the trained model against those from a parameter-randomized counterpart. Structural similarity is quantified using the Pearson correlation coefficient between the two sets of heatmaps. The occlusion test evaluates the causal relationship between highlighted regions and model predictions by measuring the confidence drop when the Grad-CAM-identified important region is masked with a black patch, relative to masking a randomly selected region of equivalent size. The ratio of important-region confidence drop to random-region confidence drop serves as the quantitative measure of region importance. A ratio substantially greater than 1.0 confirms that the highlighted region has causal, non-trivial contribution to the model's decision.

In the ConvNeXt visualization (Figure 14), four samples with correct predictions are shown along with high confidence scores (0.942; 0.941; 0.945; 0.949). The heatmap shows that the model focuses attention on clinically relevant areas. For segmented neutrophils, Grad-CAM highlights the segmented nuclear area, which is characteristic of mature neutrophils with nuclei divided into 3-5 lobes. The model also pays attention to the cytoplasm, indicating that cytoplasmic granularity contributes to the decision. From a mathematical perspective, high activation in the nuclear region indicates that filters in the last convolutional layer have learned to detect complex nuclear segmentation patterns.

For erythroblasts, the model focuses on the round, dark nucleus with dense chromatin, characteristic of erythroblasts especially in the late maturation stage. The basophilic cytoplasmic area also receives attention, although lower than the nucleus. This is consistent with diagnostic practice where the nuclear-cytoplasmic ratio is an important marker. For myeloblasts, the model highlights the large nucleus with clear nucleoli and minimal cytoplasm, corresponding to the characteristics of the youngest progenitor cells. For basophils, the visualization shows focus on large dark granules covering the nucleus, a pathognomonic feature of basophils. The model's ability to detect these granules is crucial for differential diagnosis and indicates that the feature representations learned by the model align with medical knowledge.

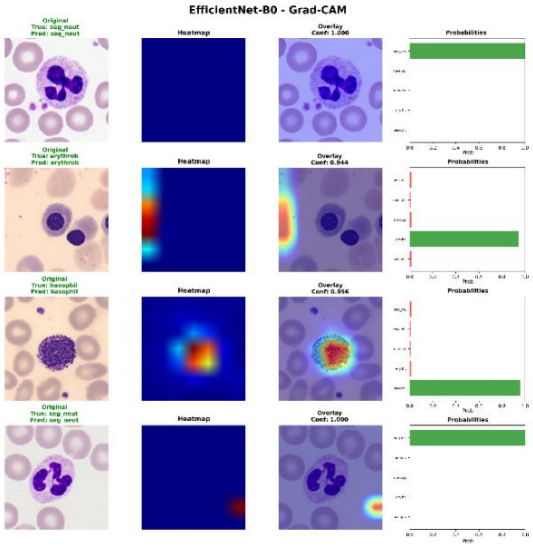

Figure 14. EfficientNet-B0 Grad-CAM Visualization

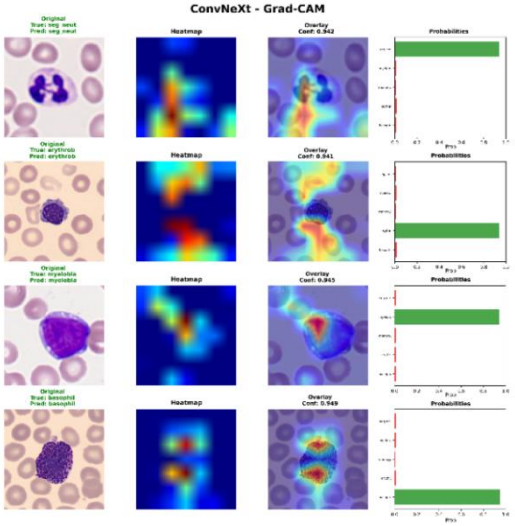

Figure 15. ConvNeXt Grad-CAM Visualization

In EfficientNet-B0, similar focus patterns are observed, but with slightly different attention distribution. EfficientNet-B0 tends to provide more diffuse focus over wider areas, while ConvNeXt shows more concentrated focus on main diagnostic features. This difference can be explained through analysis of effective receptive field. ConvNeXt with 7×7 kernels have a wider receptive field at each layer, enabling integration of global context while maintaining focus on important regions. EfficientNet-B0 with smaller kernels relies on stacking layers to achieve contextual understanding, which can lead to more distributed activation patterns. Both models show good ability to ignore irrelevant background areas, indicating that the data augmentation process (including random erasing) successfully trained the models not to rely on background artifacts an effective form of regularization.

The randomization test is a critical sanity check to ensure that Grad-CAM visualizations truly reflect the model's

decision-making process, not artifacts or random patterns. The principle is to compare Grad-CAM maps from the trained model with a model whose weights are randomly shuffled. If Grad-CAM does not change significantly after randomization, then the method is insensitive to model parameters and cannot be trusted. From a statistical perspective, this tests the hypothesis that visualizations depend on trained model parameters.

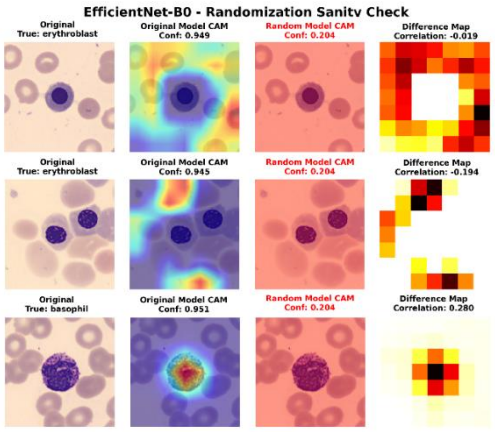

Figure 16. EfficientNet-B0 Randomization Test

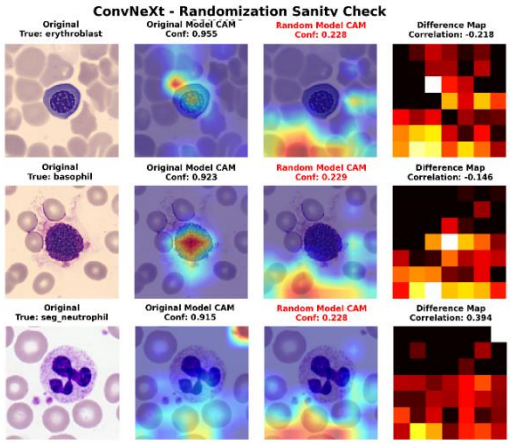

Figure 17. ConvNeXt Randomization Test

Figure 16 and 17 present the quantitative results of the randomization test across both models and all evaluated samples. For EfficientNet-B0, Pearson correlations between trained-model and randomized-model Grad-CAM maps are −0.019, −0.194, and 0.280 across three samples, yielding a mean absolute correlation of 0.164. For ConvNeXt, the values are −0.218, −0.146, and 0.394, with a mean absolute correlation of 0.253. All values remain substantially below the 0.5 threshold, confirming that both models' Grad-CAM outputs are genuinely sensitive to trained parameters. Notably, all trained models maintained high prediction confidence (0.915–0.955), while randomized counterparts collapsed to approximately 0.204–0.229, near uniform distribution across 5 classes, providing complementary confirmation that the randomized models lose all discriminative capability and their saliency maps carry no interpretive meaning.

The occlusion test complements the randomization test by testing the causal relationship between regions highlighted by Grad-CAM and model decisions. The principle is to occlude (cover) regions deemed important by Grad-CAM and observe changes in prediction confidence. If the region is indeed important, occlusion should cause significant confidence drop, much larger than occlusion of random regions. This tests the hypothesis that the region has causal contribution to model output.

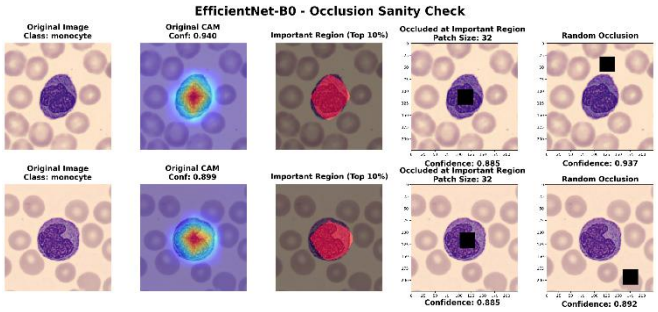

Figure 18. EfficientNet-B0 Occlusion Test

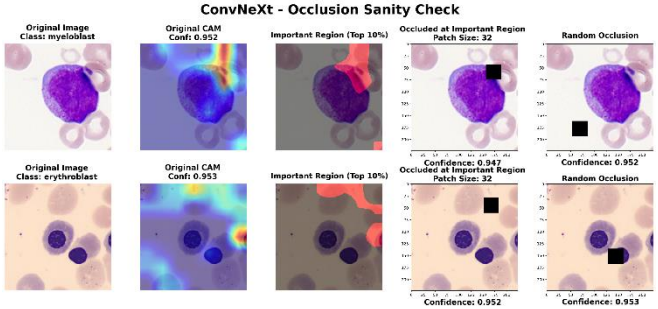

Figure 19. ConvNeXt Occlusion Test

Figure 18 and 19 present the quantitative results of the occlusion test. For EfficientNet-B0, occluding the Grad-CAM-identified important region caused confidence drops of 0.055 (from 0.940 to 0.885) in sample 1 and 0.044 (from 0.899 to 0.855) in sample 2, compared to drops of only 0.003 and 0.007 when equivalent random regions were occluded, yielding importance ratios of 18.3× and 6.3×, respectively. For ConvNeXt, confidence drops were smaller in absolute magnitude: 0.005 (from 0.952 to 0.947) in sample 1 and 0.001 (from 0.953 to 0.952) in sample 2, with no measurable drop upon random region occlusion. These results collectively confirm that Grad-CAM-highlighted regions have genuine causal contribution to model predictions. The lower absolute confidence drop for ConvNeXt reflects its greater representational redundancy, with wider receptive fields, ConvNeXt distributes diagnostic feature information across multiple spatial regions, creating robustness to single-region occlusion, while EfficientNet-B0's higher reliance on localized features simultaneously confirms that Grad-CAM has accurately identified its key decision regions.

G. Speed and Efficiency Analysis

In clinical applications, inference speed is an important factor alongside accuracy, especially for large population

screening or implementation on resource-constrained devices. Table VI and Figure 20 present a comparison of speed and efficiency of both models from a computational complexity perspective.

TABLE VI
SPEED AND EFFICIENCY COMPARISON OF MODELS

Metric	EfficientNet-B0	ConvNeXt	Difference
Inference Time (ms/image)	7.46 ± 0.19	16.45 ± 0.34	+120.6%
Throughput (FPS)	134.12	60.80	-54.7%
Model Size (MB)	18.30	107.64	+488%
Number of Parameters	4.8 million	28.2 million	+488%
F1 per ms	0.1327	0.0603	-54.6%

EfficientNet-B0 shows dramatic advantages in speed aspects, with inference time of 7.46 ms per image, 120% faster compared to ConvNeXt requiring 16.45 ms per image. In throughput units, EfficientNet-B0 can process 134 images per second, while ConvNeXt only 60 images per second. This difference can be analyzed through the computational complexity of both architectures. EfficientNet-B0 has only 4.8 million parameters, far fewer than ConvNeXt with 28.2 million parameters. Each additional parameter requires floating-point operations during forward pass, so the number of parameters correlates linearly with computational complexity in the first order.

From a computational architecture perspective, ConvNeXt uses depth wise convolution with larger 7×7 kernels, which although parameter-efficient (separating spatial and channel convolutions), still requires more floating-point operations compared to the 3×3 kernels used in MBConv blocks in EfficientNet. Mathematically, the complexity of depth wise convolution is $O(C \cdot K^2 \cdot H \cdot W)$ with C number of channels, K kernel size, H and W spatial dimensions. Increasing K from 3 to 7 increases complexity by approximately 5.44 times for the same layer. Additionally, ConvNeXt adopts more complex designs with layer normalization and GELU activation requiring additional computation layer normalization requires computation of mean and variance per activation, while GELU requires evaluation of the Gaussian cumulative distribution function.

When efficiency is measured as the ratio of F1-score to time (F1 per ms), EfficientNet-B0 achieves 0.1327, twice as efficient as ConvNeXt (0.0603). This metric represents the trade-off between performance and speed for every millisecond of computation time, EfficientNet-B0 produces a greater F1 improvement. Model size is also an important consideration for deployment. EfficientNet-B0 requires only 18.3 MB of storage, while ConvNeXt requires 107.6 MB. This difference is significant for implementation on edge devices or systems with limited storage capacity, where every megabyte is a valuable resource.

H. Accuracy-Speed Trade-off Analysis

Figure 20 presents a visualization of the trade-off between accuracy (F1-score) and inference speed for both models, providing a holistic perspective on the relative position of each architecture in the model design space. From a multi-objective optimization perspective, no model is absolutely superior in all dimensions there exists a Pareto frontier representing optimal configurations.

From this visualization, ConvNeXt is positioned with higher F1-score (0.9920) but slower speed, while EfficientNet-B0 offers much higher speed with minimal sacrifice in F1-score (only 0.27% lower). In machine learning terminology, ConvNeXt lies on the right side of the trade-off curve with high model complexity, while EfficientNet-B0 approaches the optimal point for computational efficiency.

For high-precision diagnosis scenarios, such as confirming difficult cases at referral hospitals, ConvNeXt becomes the superior choice. Its ability to capture subtle features through wide receptive fields and achieve the highest F1-score (0.9920) along with perfect AUC for all classes provides maximum diagnostic confidence. In this context, additional computational cost can be justified by the accuracy improvement crucial for patient outcomes.

For large population screening applications where thousands of images need processing daily, EfficientNet-B0 offers a more practical solution. With a speed of 134 FPS, this model can process more than 11.5 million images per day (assuming 24-hour operation), while maintaining 98.93% accuracy. High computational efficiency means lower infrastructure costs, more economical energy consumption, and faster system response important factors for sustainability and widespread adoption.

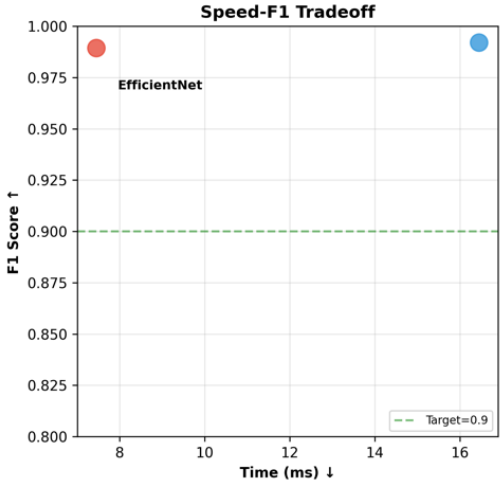

Figure 20. Speed-F1 Trade-off Analysis

For implementation on edge devices or point-of-care applications with limited resources, EfficientNet-B0 is clearly superior with a model size of only 18.3 MB (one-sixth of ConvNeXt) and lower computational requirements. This enables implementation on devices such as smartphones or smart microscopes without requiring cloud connectivity,

opening opportunities for decentralized diagnostics in areas with limited infrastructure.

I. Comprehensive Discussion

The results of this study have important implications for the development of computer-aided diagnosis (CAD) systems in hematology. The ability of both models to achieve >98% accuracy confirms that deep learning can be a highly reliable tool for blood cell classification. From a medical perspective, this means the potential to reduce the workload of pathologists and increase diagnostic objectivity. With only 5,000 training samples, the models generalize well to test data, demonstrating that moderate-sized datasets with balanced distribution and appropriate augmentation are sufficient to achieve acceptable clinical performance. This phenomenon can be explained through generalization theory in deep learning, where models with appropriate capacity can learn invariant representations from limited data if properly regularized.

To contextualize these results within the broader landscape of blood cell classification research, it is informative to consider the performance of simpler CNN baseline architectures on related tasks. Tarigan et al. (2025) reported a validation accuracy of 86.89% and a macro-averaged F1-score of 0.8688 using ResNet18 for four-class peripheral blood cell classification [24]. It must be noted that direct numerical comparison is not methodologically appropriate here, as the datasets differ in the number of classes, image sources, and laboratory conditions. Nevertheless, this contextual comparison illustrates that modern architectures such as ConvNeXt and EfficientNet-B0, when applied to a well-balanced dataset with appropriate augmentation and training configurations, can achieve substantially higher performance than simpler residual networks. The performance gap is consistent with the known representational advantages of these architectures, EfficientNet-B0's compound scaling and squeeze-and-excitation mechanisms, and ConvNeXt's larger receptive fields and transformer-inspired design, collectively enable richer feature extraction than the basic residual blocks of ResNet18. This suggests that architecture selection is a meaningful variable in blood cell classification performance, and that investment in more modern architectures is warranted for clinical-grade applications, even at the cost of increased model complexity.

The high internal benchmark performance ($F1 > 98\%$) observed for both models warrants critical examination, as reviewer concern regarding potential overfitting or dataset simplicity is well-founded. Several lines of evidence collectively argue against a trivial explanation. First, both models were evaluated on a held-out test set that was strictly separated prior to any augmentation or training, eliminating data leakage as an explanatory factor. Second, the balanced five-class dataset with 1,000 samples per class, while moderate in size, covers clinically distinct morphological categories with genuine inter-class variation, as evidenced by the observed confusion patterns between erythroblasts and monocytes, classes that challenge even expert pathologists. Third, and most critically, the external generalization results provide the strongest evidence against overfitting, ConvNeXt achieved $F1 = 0.9763$ on an entirely independent dataset from different sources, confirming that its learned representations are genuinely transferable. EfficientNet-B0's performance decline on the external dataset, while substantial, is more accurately interpreted as sensitivity to domain shift rather than overfitting to training data, as its internal train-validation-test curves showed consistent performance without signs of memorization. Nonetheless, these findings reinforce the importance of external validation as a methodological standard in computational pathology research.

Regarding model fairness and bias, the external evaluation revealed differential performance across classes that is attributable to domain shift rather than systematic bias in the learned representations. The concentration of errors in classes sourced from a specific external repository highlights the sensitivity of deep learning models to imaging domain characteristics, including staining protocols, microscope specifications, and sample preparation techniques. This has equity implications for clinical deployment, a model performing well in one institution may underperform in another with different laboratory practices, particularly for underrepresented cell types. Mitigation strategies include stain normalization as a preprocessing step, domain adaptation through few-shot fine-tuning on target-domain samples, and multi-institutional training datasets that capture the diversity of real-world clinical environments.

The observed error patterns, especially between erythroblasts and monocytes, reflect challenges also faced by human pathologists. From a cognitive perspective, this indicates that the models learn in ways conceptually similar to humans relying on the same morphological features such as nuclear shape, chromatin density, and cytoplasmic characteristics. However, the models also demonstrate the ability to distinguish cases that might be difficult for human observers, thanks to consistency and the ability to capture very subtle nuances through numerical representations in high-dimensional space. In information theory terminology, the models can extract information that may be below the threshold of human perception but still possesses statistical discriminative power.

The successful validation of Grad-CAM through randomization and occlusion tests provides a foundation of trust that model decisions are based on clinically relevant features. From an explainable AI perspective, this bridges the gap between high performance and transparency often referred to as the interpretability-accuracy trade-off. With interpretable visualizations, clinicians can verify that the model is attending to the right things before trusting its predictions, creating a trust calibration mechanism important for clinical adoption. Mathematically, Grad-CAM provides a local approximation of the model's decision function, enabling visual inspection of which regions contribute most to the output.

From a software engineering and deployment perspective, the choice between ConvNeXt and EfficientNet-B0 involves broader technical considerations. ConvNeXt at 107.6 MB may require model compression such as quantization or pruning for edge device deployment, while EfficientNet-B0 at 18.3 MB can run without additional optimization. Inference speed also affects system architecture at 134 FPS, EfficientNet-B0 enables real-time processing for video streaming applications, while ConvNeXt at 60 FPS is still adequate for static image analysis but may be less responsive for interactive applications.

This research also opens directions for future development. From an informatics perspective, exploration of knowledge distillation techniques where ConvNeXt serves as teacher and EfficientNet as student could combine the advantages of both architectures ConvNeXt's high accuracy can be transferred to the more efficient EfficientNet. From an applied mathematics perspective, further analysis of loss landscape and convergence properties of both architectures could provide insights for more optimal architecture design. From the medical side, prospective validation on multi-center datasets with variations in staining protocols and microscope equipment is needed to confirm model generalization and robustness in real clinical practice.

IV. CONCLUSION

This study conducted a comprehensive comparative analysis between two state-of-the-art deep learning architectures, EfficientNet-B0 and ConvNeXt, for multi-class blood cell classification aimed at early leukemia detection. The experimental results demonstrated that both architectures achieved exceptional performance, with accuracy exceeding 98% and F1-scores of 0.9893 for EfficientNet-B0 and 0.9920 for ConvNeXt. ConvNeXt exhibited superior classification capability, particularly in distinguishing morphologically similar cell types such as erythroblasts and monocytes, attributable to its larger receptive fields, layer normalization, and GELU activation functions that enable more discriminative feature extraction. However, this performance advantage came at the cost of computational efficiency, with ConvNeXt requiring 16.45 ms inference time per image compared to only 7.46 ms for EfficientNet-B0. The Grad-CAM visualizations, validated through rigorous randomization and occlusion tests, confirmed that both models focus on clinically relevant morphological features including nuclear segmentation, chromatin density, and cytoplasmic granularity thereby providing interpretable decision-making processes essential for clinical adoption.

The findings establish clear guidelines for model selection based on specific clinical requirements. External generalization evaluation on an independently assembled dataset further substantiated the architectural divergence between the two models, ConvNeXt maintained near-identical performance ($F1 = 0.9763$) across domains, while EfficientNet-B0 exhibited sensitivity to domain shift ($F1 = 0.7167$), particularly for classes sourced from a different imaging environment. This finding reinforces the recommendation of ConvNeXt for deployment in heterogeneous clinical settings, while suggesting that EfficientNet-B0 deployment in new environments should be accompanied by domain adaptation strategies. For high-precision diagnostic applications such as confirmatory testing in referral hospitals, ConvNeXt represents the optimal choice due to its superior accuracy and AUC scores. Conversely, for large-scale population screening programs or deployment on edge devices with limited computational resources, EfficientNet-B0 offers a compelling solution with its 134 FPS throughput and compact model size of 18.3 MB approximately six times smaller than ConvNeXt. The successful implementation of Grad-CAM with sanity checks demonstrates that modern deep learning architectures can achieve both high accuracy and interpretability, addressing the critical "black box" concern that has historically hindered AI adoption in clinical settings. This research provides a foundational computational framework for developing transparent, reliable, and efficient computer-aided diagnosis systems. It is important to note that the present study is limited to image classification under controlled benchmark conditions and does not constitute direct clinical validation. Prospective evaluation on multi-institutional datasets with varying staining protocols, microscope equipment, and patient demographics is a necessary next step before clinical deployment, and remains an important direction for future research.

REFERENCES

[1] A. Achir, I. Debbarh, N. Zoubir, I. Battas, H. Medromi, and F. Moutaouakkil, "Advances in Leukemia detection and classification: A Systematic review of AI and image processing techniques," *F1000Research*, vol. 13, pp. 1–34, 2025, doi: 10.12688/f1000research.159318.2.

[2] F. Yaseen, M. Rashid, M. Y. Shabir, M. A. Khan, and N. Hussain, "Acute Lymphoblastic Leukemia Classification: Deep Learning Techniques for Blood Diseases Diagnosis," *J. Comput. Biomed. Informatics*, vol. 9, no. 1, pp. 1–9, 2025.

[3] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[4] D. A. Arber *et al.*, "The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia," *Blood*, vol. 127, no. 20, pp. 2391–2405, 2016, doi: 10.1182/blood-2016-03-643544.

[5] S. Gupta and A. Mishra, "Deep transfer learning-based classification of White Blood Cells using customized classification base," *ACM Int. Conf. Proceeding Ser.*, pp. 585–595, 2024, doi: 10.1145/3675888.3676117.

[6] R. Kim and A. Kim, "Analysis of Modern Computer Vision Models for Blood Cell Classification," *arXiv preprint arXiv:2407.00759*, 2024.

[7] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[8] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 11966–11976, 2022, doi: 10.1109/CVPR52688.2022.01167.

[9] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible

AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
[10] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.
[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 618–626, 2017, doi: 10.1109/ICCV.2017.74.
[12] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is Grad-CAM Explainable in Medical Images?," *Commun. Comput. Inf. Sci.*, vol. 2009 CCIS, pp. 124–135, 2024, doi: 10.1007/978-3-031-58181-6_11.
[13] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
[14] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, no. 1995, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
[15] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8689 LNCS, no. PART 1, pp. 818–833, 2014, doi: 10.1007/978-3-319-10590-1_53.
[16] F. Ahmed, N. S. Naz, S. Khan, A. U. Rehman, W. M. Ismael, and M. A. Khan, "Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges," *BMC Med. Imaging*, vol. 26, no. 1, 2026, doi: 10.1186/s12880-025-02118-w.
[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.
[18] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020, doi: 10.1109/TPAMI.2019.2913372.
[19] A. Dosovitskiy *et al.*, "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.
[20] I. G. Asghar, S. Kumar, A. Shaukat, and P. Hynds, "Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review," PLOS ONE, vol. 19, no. 6, e0292026, Jun. 2024, doi: 10.1371/journal.pone.0292026.
[21] Z. Qin, F. Yu, C. Liu, and X. Chen, "How convolutional neural networks see the world --- A survey of convolutional neural network visualization methods," *Math. Found. Comput.*, vol. 1, no. 2, pp. 149–180, 2018, doi: 10.3934/mfc.2018008.
[22] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. NeurIPS, pp. 9505–9515, 2018.
[23] D. D. Gunashekar *et al.*, "Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology," *Radiat. Oncol.*, vol. 17, no. 1, pp. 1–11, 2022, doi: 10.1186/s13014-022-02035-0.
[24] Tarigan, T. E., Prasetyo, A. B., & Susanti, E., "Deep Learning-Based Blood Cell Image Classification Using ResNet18 Architecture," *Indonesian Journal of Data and Science*, vol. 6, no. 2, pp. 294-300, 2025, doi: https://doi.org/10.56705/ijodas.v6i2.300.