Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 3 (2026)

EVA-Net: Adaptive Attention Gating with Parallel Dilated Convolution for Multi-Scale CNN-ViT Fusion in Leukemia Classification

P.S.S. Bharadwaj Srikaa[1], B. Veeramallu[2]

[1] Department of CSE, Koneru Lakshmaiah Education Foundation Vaddeswaram, AP, India. email

[2] Department of CSE, Koneru Lakshmaiah Education Foundation Vaddeswaram, AP, India. Email

***Abstract*:-** The early and accurate diagnosis of leukemia is necessary for effective treatment. Traditional methods of diagnosis, including manual microscopy of peripheral blood smears, are subject, time-consuming and produce inconsistent results. The use of advanced methods such as hybrid CNN-RNN models and CNN with Vision Transformers (ViT) suffers from difficulties in capturing fine-grained interactions and identifying subtle morphological differences, particularly in leukemic cells. This paper introduces a dual backbone network (EVA-Net) designed to improve the classification performance. EVA-Net introduces Parallel Dilated Convolution (PDC) in the Multi-Scale Refinement Block (MSRB) to increase the receptive field of the analysed features and Adaptive Attention Gating (AAG) for the spatially-adaptive integration of EfficientNet local features and ViT global context. In addition, it adopts a memory-saving training strategy, incorporating gradient checkpointing, mixed precision and gradient accumulation, which allows it to be deployed in standard GPUs as well. The architecture enables the building of a powerful, efficient multi-class classification model for leukemia that outperforms the current ones, and that can be made interpretable via attention visualisation. EVA-Net was able to achieve an accuracy of 99.59%, a macro F1 score of 0.9959 and a Matthews correlation coefficient of 0.9944 on 3,256 images of peripheral blood smears across four classes, significantly outperforming competitive models like ResNet50 (93.05%) and EfficientNetB3 (97.14%). McNemar's test ($p<0.01$) confirmed that statistical significance was extremely high and the relative error was found to be 31.5% greater than the best one obtained in ablation studies.

Key words: Leukemia classification, Adaptive attention gating, Parallel dilated convolution, CNN-ViT fusion, Medical image analysis, deep learning, peripheral blood smear.

1. Introduction

Leukemia is a group of haematological malignancies that are diverse and defined by the proliferation of abnormal white blood cells in the bone marrow and peripheral blood in an uncontrolled manner [1]. Leukaemia is a primary cancer in which the incidence of new cases is estimated at more than 450,000 cases a year globally and the mortality rate is still alarming, especially in low and middle-income countries where access to special diagnostic facilities is limited [2]. There are several subtypes of the disease, each treated differently, such as acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphoblastic leukemia (CLL) and chronic myeloid leukemia (CML). The prognosis is very much dependent on timely diagnosis and correct classification of the subtype, with delayed diagnosis resulting in poor prognosis, both with regard to life expectancy and treatment effectiveness [3].

Peripheral blood smears are the most commonly used tissue for the diagnosis of leukemia and are examined manually with a microscope by trained hematologists and pathologists [4]. This method involves visual examination of the cell morphology, such as the nuclear/cytoplasmic ratio, nuclear chromatin, the presence of nucleoli, and the cytoplasmic granulation, which is useful to distinguish healthy cells from leukemic blasts and to distinguish between stages of the disease. Nevertheless, manual screening is very time consuming, labour

intensive and highly inter- and intra-observer variable [5]. The difference in diagnosis between experienced pathologists has been reported to be as high as 20% for leukemic classification in cases with ambiguous diagnosis [6]. In addition, the early leukemic blasts are often morphologically similar to benign reactive lymphocytes and the discrimination of the different progressive stages (early pre-B, pre-B, and pro-B ALL) requires a high level of expertise [7]. Additionally, in resource-limited areas, trained hematologists are scarce, further complicating these issues, and there is a need for automated, reliable, and accessible diagnostic tools.

In recent years, deep learning (DL) has shown great promise in medical image analysis, and the CNN model has achieved state-of-the-art performance in many medical diagnostic applications, including VGG, ResNet and EfficientNet, [8], [9]. Inspired by this, powerful alternatives, known as vision transformers (ViTs), which use self-attention mechanisms for superior global context modelling, have been introduced [10]. In the context of leukemia classification, hybrid architectures have been employed which integrate CNNs with recurrent networks or transformers for global feature extraction have proven successful in [11], [12], [13]. There are still some key constraints, however. One, most fusion currently used methods are static operations such as concatenation or addition, without learning adaptive spatial weighting between local and global features [14]. Second, they do not have multi-scale refinement modules specifically designed to detect variations in the morphology of the image within each class, which are critical for stage discrimination [15]. Third, dual-backbone models are expensive and can consume more memory than typical GPUs, and are hard to replicate and clinically deploy [16].

To overcome these shortcomings, we propose EVA-Net (EfficientNet-ViT Attention Network), a new network architecture with three novel components: (1) Adaptive Attention Gating (AAG) a learnable, spatially-adaptive mechanism to fuse multi-scale EfficientNet local features with ViT global context, by channel recalibration in the squeeze step, on a per-pixel, per-channel basis, allowing to select relevant information based on cellular morphology; (2) Parallel Dilated Convolution (PDC) in a Multi-Scale Refinement Block (MSRB) a post-fusion module to capture heterogeneous receptive fields with channel recalibration by squeeze-and-excitation; and (3) a memory-efficient training strategy: gradient checkpointing, mixed precision, and gradient accumulation, to train the dual-backbones on a single consumer-grade GPU. EVA-Net is tested on a publicly available four-class dataset for leukemia with 3,256 peripheral blood smear images. Extensive experiments include comparisons between our model and seven strong baselines: ResNet50, EfficientNetB3, ConvNeXt, Xception+Bi-GRU, EfficientNetB3+Bi-GRU and EfficientNetB3+ViT. Results of statistical significance testing (McNemar's test) and ablation studies support each of the architectural components, and Grad-CAM visualization indicates that each component correlates with clinically relevant cytomorphological features. They all contribute to the framework of EVA-Net as a clinically viable and reproducible, automated leukemia diagnosis.

The rest of this paper is organized as follows: Section II presents a review of the related works; Section III presents the Dataset Description; Section IV presents the methodology; Section V presents the Results and Discussion; Section VI concludes the paper.

2. Related Work

The section discusses the recent developments in deep learning applications in the classification of leukemia, hybrid CNN RNN networks, CNN ViT fusion architectures, multi scale feature fusion, and attention gating mechanisms, which identify the gaps in research that prompt our proposed EVA Net.

A. DL to Classify Leukemia.

Deep learning has greatly improved automated blood cancer detection from peripheral blood smears. A systematic review proves that DL algorithms always outperform classical machine learning (ML) algorithms, achieving an accuracy of more than 95% on benchmark data [11]. An in-depth performance study of various deep learning models to classify leukemia shows progress as well as the current issues [1].

CNNs have been popular because they have the capability of extracting hierarchical local features. A CNN transformer cross attention fusion network is presented that illustrates that the fusion of local and global features

representation can significantly improve the classification of medical images [5]. MedNet, A lightweight attention augmented CNN, achieves competitive performance with less computational resources [4].

The combination of multi scale feature extraction with EfficientNet to classify ALL improves the discriminative power of CNN based models [8]. The accuracy of a deep learning system that diagnoses and types leukemia using single peripheral blood cells is high, which demonstrates the potential of automated methods in routine clinical practice [10].

B. Hybrid CNN and RNN Models.

CNNs have been used with recurrent neural networks (RNNs), in particular, bidirectional GRUs (Bi GRUs), to model sequential dependencies over spatial grids. A deep feature fusion cascade of entropy regulated firefly feature selection (Csec net) demonstrates that long range patterns of leukemia can be modelled sequentially through feature modelling [13]. Yet, these models can only be used sequentially and cannot achieve the pairwise global attention capabilities of transformer architectures [17].

C. CNN-ViT Fusion Architectures

ViTs have become powerful substitutes to CNNs in medical picture classification, and recent research has extensively explored hybrid CNN-ViT networks. The transformer CNN fusion framework (ConvTNet) shows better performance in a variety of medical imaging tasks [16]. The model takes the EfficientNet-B0 and ResNet 50 models and fuses them with a ViT, resulting in state-of-the-art performance whilst staying explainable through visualization of its attention [7]. It is shown that learnable fusion mechanisms can be used to effectively combine complementary representations across different architectures [15].

Analysis of the current CNN+ViT fusion methods demonstrates three key flaws. First, most fusion mechanisms employ only local and global features through the use of static operations, such as concatenation or addition, but do not learn to adaptively weight local and global features per spatial location [14]. Second, most of these procedures use the last CNN feature map and ignore the rich fine-grained information that exists at the earlier convolutional stages [4]. Third, they do not include explicit multi scale refinement blocks following fusion, which is imperative to capture heterogeneous cellular structures, typical of leukemic cells [16].

Adaptive fusion has started to be discussed in recent publications. A multi scale feature fusion algorithm under the theory of evidence (MUSCLE) demonstrates that feature fusion is sensitive to hierarchy and can be greatly enhanced to increase discriminative capacity [14].

D. Multi Scale Feature Fusion and Attention Gating

Multi scale feature representation plays a crucial role in medical image analysis where pathological features are highly diverse in size at different stages of the disease. A multi scale local and global features adaptive hierarchical enhanced multi attention feature fusion framework (HEMF) synergistically extracts and fuses multi scale local and global features using a novel mixed attention mechanism [18]. A multi scale context aware attention model adds more features representations by the fact that it captures the dependencies across various spatial scales [12].

The attention gating mechanisms have been successfully used to selectively emphasize informative characteristics and attenuate irrelevant information. A gated multi scale hybrid vision transformer (ViResGF Net) presents a hierarchical gated feature aggregation mechanism to enrich multi scale features [6]. The quality of gating in multi scale environments to medical image segmentation is validated by a gated axial reverse attention network [16].

Parallel dilated convolutions have demonstrated potential with respect to capturing multi scale context with multiple dilation rates at once. The use of multi scale dilated convolution to achieve the objective of uniform coverage of the input images and prevention of feature discontinuities is evidenced in the case of a hybrid residual and dual block transformer network to diagnose blood cells [2]. The success of parallel dilated convolutions further confirms the effectiveness of a hybrid CNN-ViT framework with cross attention fusion to classify brain tumours [19].

E. Research Gap and Proposed Contributions

In spite of these developments, a critical analysis of the literature shows that no existing method can at once meet three important conditions of a robust leukemia classification: (1) adaptive spatial gating of CNN local features and ViT global context, (2) hierarchical use of multi scale CNN features at multiple stages and (3) explicit post fusion multi scale refinement using parallel dilated convolutions.

The strength of the combination of several models is proved by an ensemble-based approach to hematologic disorders [20]. Nevertheless, ensemble techniques are still computationally intensive, and they are not as efficient as a single unified architecture. Moreover, the majority of the existing approaches are based on the fixed fusion policy [5], [7] and not on learning to adaptively gate between local and global representations based on the per pixel context.

To fill these gaps, we propose EVA Net, a new architecture that introduces: (i) an AAG mechanism that learns to combine multi scale CNN local features with ViT global context per spatial location, (ii) a PDC within a MSRB to capture heterogeneous receptive fields and (iii) a memory efficient training strategy that enables the dual backbone model on standard GPUs, making the architecture reproducible and suitable to clinical deployment.

3. Dataset Description

The image dataset used in this research is the ALL image data set openly available on the Kaggle data science platform [21]. The images in the data set were provided by the bone marrow laboratory of Taleqani hospital (Tehran, Iran) and represent peripheral blood smear images of patients who are suspected of having ALL.

The data set source and collection are shown in Table 1, with patient numbers, image acquisition, format and diagnostic confirmation.

Table 1 Dataset Source and Collection Details

Attribute	Details
Original Source	Taleqani Hospital, Tehran, Iran
Total Patients	89 suspected ALL patients (25 benign, 64 malignant)
Image Acquisition	Zeiss microscope camera at 100× magnification
Image Format	JPG
Diagnostic Confirmation	Definitive diagnoses confirmed by flow cytometry

Dataset Composition:

The dataset contains 3,256 peripheral blood smear images organized into four clinically relevant classes:

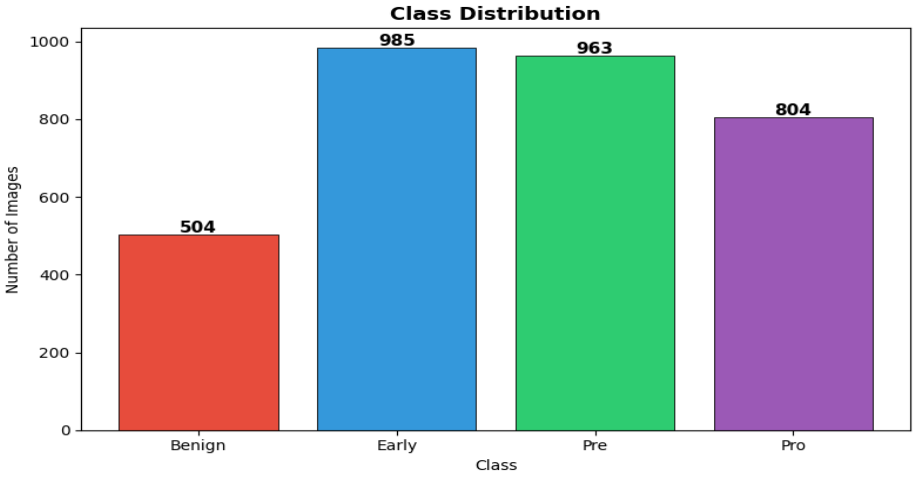

Figure 1 Class Distribution Bar Chart

The class distribution of the data is shown in Fig 1. The Benign class comprises 504 samples, and the remaining classes (Early, 985; Pre, 963; and Pro, 804) make up most of the samples. The distribution is similar to that in

clinical situations because in both populations it is more likely to suspect the diagnosis of leukemia than to suspect a benign type of leukemia; therefore, it is possible to test the performance of the model in realistic conditions.

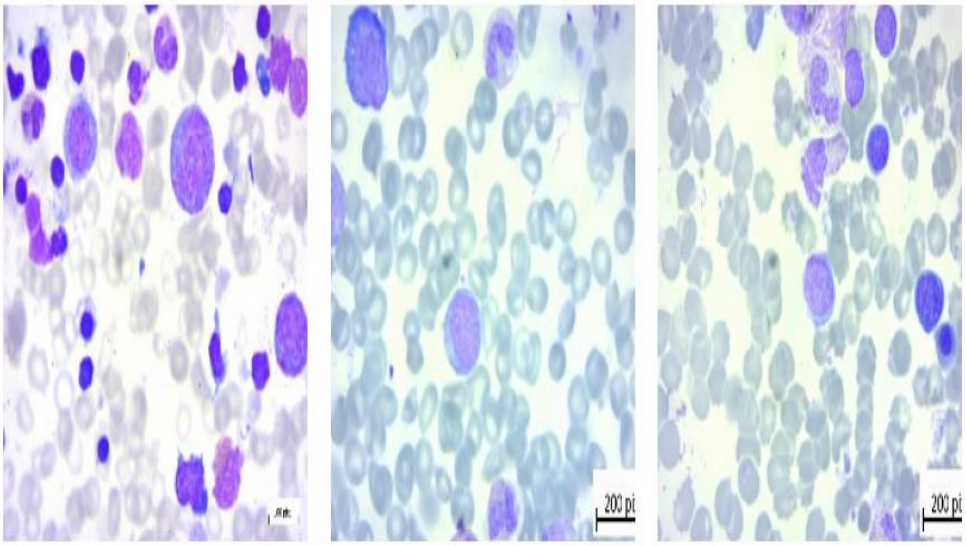

Figure 2 Sample Images per Class

Figure 2 shows some sample images of each of the four classes. Hematogones in the Benign class are normal with smooth nuclear boundaries and normal nuclear-to-cytoplasmic ratio. Early class nuclear irregularities and chromatin changes are seen; Pre and Pro classes show more and more nuclear-to-cytoplasmic ratio and nuclear abnormalities. These different visual changes underlined the diagnostic difficulty, which is to differentiate between Benign and Early, and between Pre and Pro, owing to the fact that there are only minor morphological differences in the former, while the latter have a more marked cytomorphological difference.

4. Methodology

In this section, the entire methodological structure of our proposed EVA-Net architecture is outlined, including the study design, description of the data set, data processing, architecture, training configuration, and evaluation metrics.

A. *Study Design*

Paper introduces an automatic framework for leukemia subtyping by analysing the peripheral blood smear images with a novel dual-backbone deep learning framework called EVA-Net (EfficientNet-ViT Attention Network). The proposed system aims to classify four grades of leukemia namely: Benign, Early, Pre-leukemic and Pro-leukemic which are the most important markers in early diagnosis and treatment planning in the field of hematological malignancies.

The proposed system for leukemia classification is presented with the overall work flow as shown in Fig 3. The system is well-defined and structured, with data acquisition, data preprocessing, data splitting (training/validation/testing), data augmentation, model training via adaptive attention gating and parallel dilated convolution, as well as extensive performance evaluation.

The initial step in the workflow is the loading of four-class leukemia data (Benign, Early, Pre, Pro). It then divides the data set into a training set (70%), a validation set (15%) and a test set (15%), while ensuring the proportions of the classes remain the same. The training set is split then augmented with a range of data augmentation methods including random horizontal flip, random rotation, colour jitter and resize.

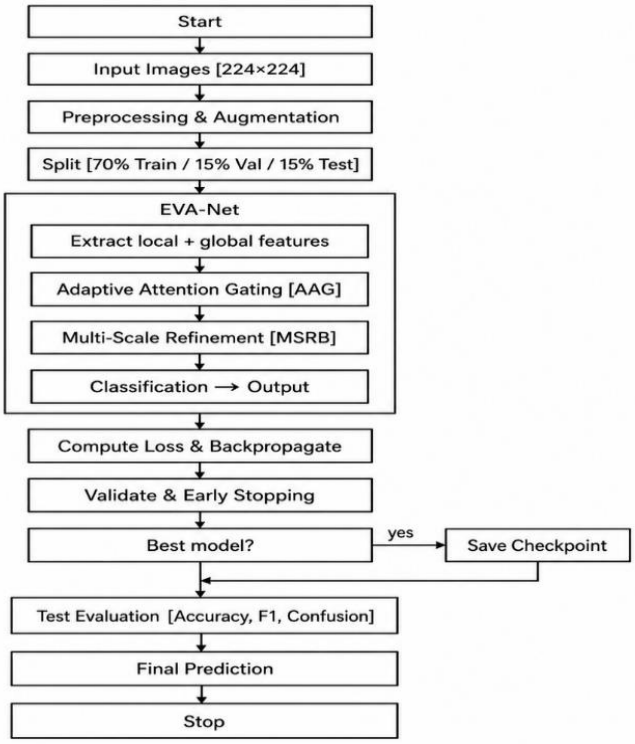

Figure 3 Overall workflow of the proposed system for Leukemia classification

This enhanced training data is then fed into the EVA-Net backbone network, which includes (i) an EfficientNet-B3 backbone network to capture multi-scale local features, (ii) a ViT backbone network for capturing global contextual features, (iii) AAG modules to combine local and global features at every spatial location, and (iv) MSRB with PDC for post-fusion refinement. AdamW optimizer and Cosine annealing are used to train the model. During training the validation set is employed to track the performance of the training process and to select the "best" model checkpoint. Finally, it provides an objective assessment of the trained model, by analysing the test data set which has not been used for the training of the model. The results are compared with base line architectures to show the superiority of EVA-Net.

B. *Data Processing*

Pre-processing: all of the images of the peripheral blood smears to the same standard, making them compatible with the model architecture. A stratified random sampling approach was used to create training, validation and testing sets with 70%, 15% and 15% respectively, to ensure class separation in the training, validation and testing sets.

Data Augmentation: Only used data augmentation on the training set, in order to improve generalization and not overfitting. The resizing and normalization of the validation and test sets were only done to measure true generalization performance. The data-processing steps to the data set are shown in Table 2.

Table 2 Image pre-processing parameters and stratified 70–15–15 dataset partitioning

Processing Step	Description
Resizing	All images resized to 224×224 pixels
Normalization	Pixel intensities normalized to [-1, 1] using mean and standard deviation of 0.5
Data Augmentation	Applied only to training set
Stratified Split	70% training, 15% validation, 15% test

Table 3 shows the augmentation techniques applied only to the training set. These enhancements mimic real-world phenomena of staining, lighting, cell orientation and image quality to make models more robust.

Table 3 Training-set augmentation strategies with probabilities

Augmentation	Details	Probability

Random Horizontal Flip	Flip image horizontally	50%
Random Vertical Flip	Flip image vertically	20%
Random Rotation	Rotate image by ±15 degrees	100%
Color Jitter	Brightness ±20%, Contrast ±20%, Saturation ±10%	100%
Gaussian Blur	Kernel size 3, sigma (0.1, 2.0)	30%
Gaussian Noise	Mean 0.0, Std 0.05	30%

These enhancements are added one at a time when training. Gaussian blur and Gaussian noise are added specifically to increase the robustness of the model to fluctuations in the quality of an image, which are typical of clinical images, including the quality of the staining, the focus and the imaging conditions. Only resizing and normalization were applied to the images using Validation and Test Sets for unbiased assessment. This way, reported performance metrics are actually generalisation on unmodified clinical data.

Data Loaders: The data was loaded with PyTorch's ImageFolder data loader where the splitting was done in a stratified manner to ensure class proportions. The training batches were shuffled, the validation and test batches were shuffled to ensure reproducible results.

C. *Proposed Architecture*

The proposed EVA-Net (EfficientNet ViT Attention Network) is a dual backbone deep learning model that is specifically designed to classify multi class leukemia on the basis of peripheral blood smear image analyses. It combines the feature extraction capability of the local regions of EfficientNet-B3 and the modelling capability of the global context of ViT Base (Vision Transformer). The architecture is shown in Fig. 4, consisting of five main modules: multi scale local feature extraction using EfficientNet-B3, global feature extraction using ViT Base, AAG modules that fuse local and global features at three scales, a MSRB that incorporates PDC and Squeeze and Excitation (SE), and a lightweight classification head.

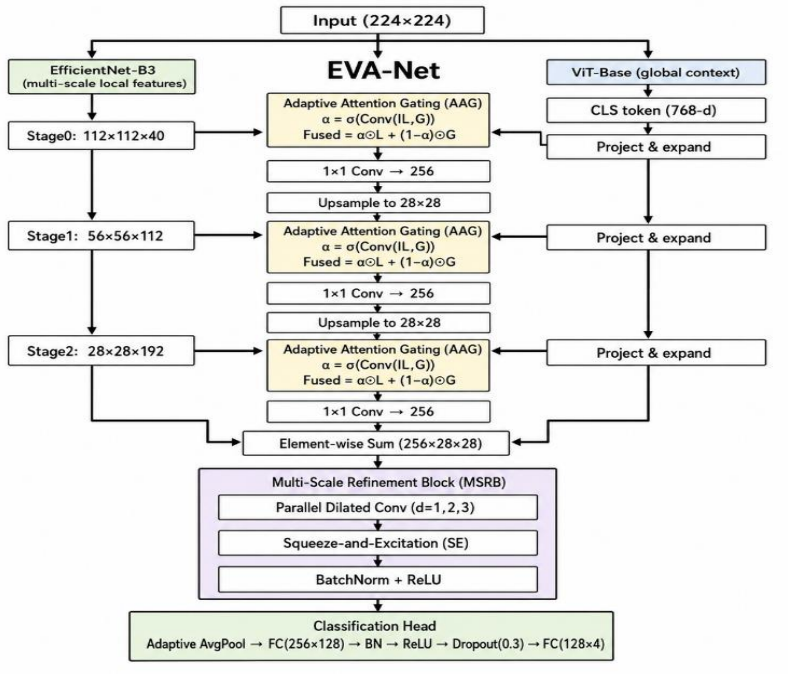

Figure 4 Architecture for Leukemia Classification (EVA-Net)

1) Multi-Scale Local Feature Extraction with EfficientNet-B3

EfficientNet-B3 is a scalable CNN that is pretrained on ImageNet. We set it with $features\ only = True$ to give out intermediate feature maps at various stages. The three distinct stages (indices 0, 1 and 2) are chosen due to the fact that they give more abstract representatisons and still maintain a high spatial resolution:

- Stage 0: spatial size 112×112, channels = 40. Resolves small details like edges of cells and texture.

- Stage 1: spatial size 56×56, channels = 112. Codes intermediate level features such as nuclei shape and chromatin pattern.
- Stage 2: spatial size 28×28, channels = 192. Represents high level semantics, e.g. cell type indicators.

These multi scale feature maps are represented as $L\_i \in \mathbb{R}\^\{B \times C\_i \times H\_i \times W\_i\}$, where B is batch size, and (H_i, W_i) are spatial dimensions.

2) Global Feature Extraction with Vision Transformer (ViT-Base)

The ViT-Base model transformer blocks, embedding dimension 768) takes the same input (224×224) and processes it. The image is broken down into 196 patches that are non-overlapping linearly projected into a token. Positional embeddings are inserted to preserve spatial data, a learnable $[CLS]$ token is prepended. The last representation of the $[CLS]$ token is the global context vector $g \in \mathbb{R}\^\{B \times 768\}$. This is a long-range dependency and holistic morphological information that cannot be acquired by any of the CNN receptive fields.

3) Adaptive Attention Gating (AAG) for Hierarchical Fusion

We linearly compose the local feature map L_i with the global feature g, and use a learnable, spatially adapting gate. It is a process that is broken down into three steps:

- Step 1 – Global projection and spatial expansion: The global characteristic is linearly mapped to correspond to the channel dimension of the local feature map:
  $g\_proj = W\_proj\^(i) \cdot g$ (1)

  Where,

  $$W_{proj}^{i} \in \mathbb{R}^{\{C_i \times 768\}} \quad (2)$$

  It is then scaled to the same spatial scale as L_i by replication to give $G\_i \in \mathbb{R}\^\{B \times C\_i \times H\_i \times W\_i\}$.
- Step 2 – Gate computation: The local and expanded global attributes are concatenated along the channel axis to make $[L\_i, G\_i] \in \mathbb{R}\^\{B \times 2C\_i \times H\_i \times W\_i\}$. A 1×1 convolutional layer is then used to reduce the channel dimension back to C_i followed by a sigmoid activation to generate a spatial gate map:
  $$\alpha\_i = \sigma(Conv\_\{1 \times 1\}([L\_i, G\_i])) \quad (3)$$
  In which σ is the sigmoid. L_i is the same as the dimension of the gate α_i, and is in$(0,1)$.
- Step 3 – Gated fusion: The fused feature map is calculated as a convex combination with weights of the gate:
  $$F\_i = \alpha\_i \odot L\_i + (1 - \alpha\_i) \odot G\_i \quad (4)$$
  Where $\odot$ is the element wise product. The mechanism enables the network to choose, by spatial location, and channel, whether to trust local information (α_i near 1) or global information (α_i near 0). An example is that when the boundaries between cells are clear, then the gate can prefer local information; in dense or blurry areas the gate can shift towards global information.

4) Multi-Scale Fusion and Unification

Once we determine the fused features F_0, F_1, F_2 of the three scales, we combine them into a commonly shared channel dimension and spatial size:

- Channel unification: A 1x1 convolutional layer will be used to reduce the number of channels per F_i to a target dimension $C\_target = 256$.
- Spatial resizing: The smaller spatial maps (F_0 at 112×112 and F_1 at 56×56) are up sampled to the largest spatial size (28×28 of F_2) using bilinear interpolation.

- Element wise summation: The three tensors are added to form a combined feature map $X \in \mathbb{R}^{B \times 256 \times 28 \times 28}$.

This hierarchical summation still retains both fine grained and high-level information and no scale will dominate.

5) Multi-Scale Refinement Block (MSRB) with Parallel Dilated Convolution (PDC)

The unified feature map $X \in R^{B \times 256 \times 28 \times 28}$ obtained from the multi-scale fusion stage still represents cellular structures at a single effective receptive field. Leukemic cells, however, exhibit morphological variation at multiple spatial granularities from fine chromatin texture to coarse nuclear/cytoplasmic boundaries so X is passed through a Multi-Scale Refinement Block (MSRB) before classification. The MSRB has two sub-components: a PDC module for multi-scale context aggregation, and a Squeeze-and-Excitation (SE) module for channel recalibration.

Step 1 — Parallel Dilated Convolution.

Four parallel 3×3 convolutional branches are applied to X, each with a different dilation rate d ∈ {1, 2, 3, 4}, so that each branch captures context at a different effective receptive field while preserving spatial resolution:

$$Pd_d = ReLU(BN(Conv3 \times 3^{d}_{3\times3}(X))),\ d \in \{1,2,3,4\} \quad (5)$$

where $Conv_{3x3}$^d denotes a 3×3 convolution with dilation rate d and matching padding to keep the spatial size fixed at 28×28, and each branch outputs 256 channels. The four branch outputs are concatenated along the channel axis:

$$Pcat_{cat} = [P1_1, P2_2, P3_3, P4_4] \in RB \times 1024 \times 28 \times 28^{B\times1024\times28\times28} \quad (6)$$

and projected back to the target dimension using a 1×1 convolution:

$$P = Conv1 \times 1_{1\times1}(Pcat_{cat}) \in RB \times 256 \times 28 \times 28^{B\times256\times28\times28} \quad (7)$$

This parallel (rather than cascaded) arrangement lets the network aggregate small, medium and large receptive-field context simultaneously at every spatial location, without the effective resolution loss of stacked dilated layers.

Step 2 — Squeeze-and-Excitation.

To recalibrate channel-wise importance, P is first squeezed via global average pooling to a channel descriptor z ∈ R^{B×256}:

$$zc_c = \frac{1}{28\times28}\sum_{h=1}^{28} P \sum_{w=1}^{28} Pc,h,w_{c,h,w} \quad (8)$$

then passed through a two-layer bottleneck (reduction ratio r = 16) with a sigmoid gate to produce channel-wise excitation weights s ∈ (0,1) ^256:

$$s = \sigma(W2_2\ \delta(W1_1\ z)) \quad (9)$$

where $\delta\ is\ ReLU, W_1 \in R^{\wedge}\{16 \times 256\}, W_2 \in R^{\wedge}\{256 \times 16\}$ .The refined feature map is obtained by channel-wise rescaling of P followed by a residual connection to the input fusion feature X:

$$Xrefined_{refined} = (s \odot P) + X \quad (10)$$

The residual term preserves the original multi-scale fused representation while the dilated-convolution and SE pathway adds heterogeneous receptive-field context and channel-wise recalibration. $X\_refined \in R^{\wedge}\{B \times 256 \times 28 \times 28\}$ is the final refined feature map passed to the classification head.

6) Classification Head

The classification head is a lightweight multi-layer perceptron (MLP). It commences with an adaptive average pooling layer which shrinks the spatial dimensions of the refined feature map by 28×28 to 1×1, generating a 256-size vector of refined feature map per sample. This vector is then flattened to shape $B \times 256$. and fed through a

linear layer that downsizes the dimension of $256 \rightarrow 128$, after which there is a batch normalization, ReLU activation, and dropout with a dropout rate of 0.3 to prevent overfitting. Lastly, a second linear layer projects the 128-dimensional representation into 4 final output logits that represent the four different leukemia classes: Benign, Early, Pre, and Pro. Inference is performed by applying a softmax function to the logits to get probabilities of the classes.

7) Architectural Comparison

All the models Table 4 compares the architectural features of all models and the proposed EVA-Net.

Table 4 Architectural comparison of baseline, hybrid, and proposed models.

Model	Backbone(s)	Fusion Mechanism	Multi-Scale Refinement	Memory-Efficient Training	Learnable Gating
ResNet50	ResNet50	None (single backbone)	✗	✗	✗
EfficientNetB3 (pure)	EfficientNetB3	None (single backbone)	✗	✗	✗
ConvNeXt	ConvNeXt	None (single backbone)	✗	✗	✗
Xception+Bi-GRU [11]	Xception + Bi-GRU	Sequential (RNN)	✗	✗	✗
EfficientNetB3+Bi-GRU [13]	EfficientNetB3 + Bi-GRU	Sequential (RNN)	✗	✗	✗
EfficientNetB3+ViT [7]	EfficientNetB3 + ViT	Static (Concatenation)	✗	✗	✗
EVA-Net (Proposed)	EfficientNetB3 + ViT	**Adaptive Attention Gating**	✓ **(MSRB + PDC)**	✓	✓

D. Training Configuration and Implementation Details

This section explains the training configuration, hyperparameters, loss function, optimization strategy, and memory efficient methods that render training the dual backbone EVA-Net to be practical on a single computing unit.

1) Optimization and Loss Function

The model is trained to achieve the lowest value of the standard cross entropy loss in multi class classification:

$$L = -(1/N)\,\Sigma\_\{n=1\}\^N\,\Sigma\_\{c=1\}\^4\, y\_\{n,c\} log\left(\hat{y}\_\{n,c\}\right) \qquad (11)$$

where N is the batch size, $y\_\{n,c\}$ is the ground truth indicator (1 in case the sample n belongs to the class c, otherwise 0) and $\hat{y}\_\{n,c\}$ is the predicted probability after softmax. AdamW optimizer is used because it does not couple the weight decay with the gradient update, which would provide a better regularization. The initial learning rate will be $1e-4$. and the cosine annealing scheduler will decrease the learning rate to almost zero in 15 epochs ($T\_max = 15$). The weight decay is $1e-5$.

2) Memory-Efficient Training Strategies

It is possible to train EVA-Net (98M parameters) on a typical GPU with three optimization steps:

- Gradient checkpointing - Recomputes activations on the forward pass, and replaces them with the recomputed value on the backpropagation step. This is within the scope of gradient checkpointing.
- Mixed precision (AMP) - Float16 is used in both forward and backward passes, reducing the total memory used by the computer by 30-40% and accelerating training without impacting accuracy.
- Gradient accumulation- A small physical batch size is used and accumulated across multiple steps to approximate a larger effective batch, with a low peak GPU memory.

These methods make training well-suited to a single GPU, and inference can be implemented with even smaller memory requirements, enabling it to be deployed to edge devices.

3) Training Hyperparameters

All hyperparameters in training EVA-Net are summarized in Table 5.

Table 5 Training Hyperparameters for EVA Net

Hyperparameter	Value
Input image size	224 × 224
Physical batch size	8
Gradient accumulation steps	4
Effective batch size	32
Total epochs	15
Early stopping patience	10 (based on validation loss)
Early stopping minimum delta	0.001
Optimizer	AdamW
Initial learning rate	1e-4
Loss function	Cross-entropy
Gradient checkpointing	Enabled (ViT blocks only)

4) Early Stopping and Model Selection

In order to avoid overfitting, we track the loss on validation at the end of every epoch. Training is stopped in case the loss of validation is not reduced by at least 0.001 over 10 consecutive epochs. This model is saved as the model checkpoint with the lowest validation loss (or highest validation accuracy, depending on the metric), which is later used to evaluate the test. In our experiments, the process of training automatically stopped after 15 epochs without any early stopping, as the loss in the validation process was gradually decreasing.

5) Baseline Model Training:

For a fair comparison, the three transfer-learning baseline models (ResNet50, EfficientNet-B3 and ConvNeXt) were retrained on the same data splits, augmentation pipeline and evaluation procedure as EVA-Net, instead of the original hyperparameters used in previous literature. This way it will be certain that performance variations are due to architectural differences and not to differences in training program. ImageNet-pretrained weights were used to initialize each baseline and end-to-end-fine-tuned with the four-class leukemia dataset. Table 6 shows the training hyperparameters of the three baseline models.

Table 6 Training Hyperparameters for Baseline Models

Hyperparameter	ResNet50	EfficientNetB3	ConvNext
Epochs	15	15	15
Batch size	8	8	8
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-4	1e-4	1e-4
Input size	224×224	224×224	224×224

6) Hybrid Model Training

To compare fairly, the three hybrid models (Xception+Bi GRU, EfficientNetB3+Bi GRU, EfficientNetB3+ViT) were trained using the same data splits and evaluation protocol. Their hyperparameters were adjusted based on the suggested values of the original implementations or the literature as shown in Table 7. All baselines were trained with the same GPU and using mixed precision where possible.

Table 7 Training Hyperparameters for Hybrid Models

Hyperparameter	Xception+Bi-GRU	EffNetB3+Bi-GRU	EffNetB3+ViT
Epochs	15	15	15
Batch size	32	32	32
Optimizer	AdamW	AdamW	AdamW

Learning rate	1e-4	1e-4	1e-4
Input size	224×224	224×224	224×224

5. Results and Discussion

A. Experimental Results:

To ensure the same training, validation and test data split for all the baseline, hybrid and proposed models, the dataset of images was split into 70% training, 15% validation, and 15% test data by fixed seed value of 42 for the microscopic blood smear image dataset for leukemia diagnosis with four classes: Benign (504), Early (985), Pre (963), Pro (804). The training set was resized and then normalized, and then augmented with random flips, rotations, colour jittering, gaussian blur and additive noise, the validation and test sets were just resized and normalized to allow unbiased evaluation. All images were resized to 224 × 224 and normalized, the training set was augmented with random flips, rotations, colour jittering, gaussian blur, and additive noise, while the validation and test sets were resized and normalized only. The Adam optimizer (learning rate = 1×10^{-4}, weight decay = 1×10^{-5}), early stopping (patience = 10, min delta = 0.001), PyTorch, and a CUDA enabled GPU were used to train all models, with a batch size of 8 and gradient accumulation of 4 steps (effective batch size = 32) for up to 15 epochs. To measure the performance, several accuracy, precision, recall, F1-score and Matthews Correlation Coefficient (MCC) metrics were used and to test the statistical significance of the difference between the performances of the different models, McNemar's test was used.

B. *Performance of Baseline Models*

For benchmarking, three popular pre-trained CNN architectures – ResNet50, EfficientNetB3, and ConvNeXt – were fine-tuned on the leukemia stage classification task with the same training and augmentation strategy, data split, and training duration as described above. Table 8 summarizes the results obtained from the test sets.

Table 8 Performance Comparison of Baseline Models

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**MCC**
ResNet50	93.05	93.28	93.05	93.01	0.9068
EfficientNetB3	97.14	97.20	97.14	97.13	0.9613
ConvNeXt	98.77	98.77	98.77	98.77	0.9833

All three baselines had a good classification result as seen in Table 8, with accuracy of 93.05% to 98.77%. The lowest performance was obtained for ResNet50 (93.05% accuracy, MCC = 0.9068), which is due to the fact that it is a rather shallow residual feature extraction compared to the more recent architectures assessed. To improve on this, EfficientNetB3 scaled up the network horizontally by 4.09 percentage points (97.14% accuracy) with the help of its compound scaling approach, which optimizes network depth, width and input resolution. The result of ConvNeXt was also better than ResNet50 in terms of accuracy (98.77% vs. 93.05%), F1-score (98.77% vs. 93.01%), and MCC (0.9833 vs. 0.9068), reflecting its modernized convolutional design with large kernel sizes, inverted bottlenecks, and layers inspired by vision transformers that can identify finer morphological differences between the stages of leukemia. We note the relatively high proportions of incorrect classifications for the Benign class (89% F1 for ResNet50; 93% for EfficientNetB3) on each baseline model, and on each per-class analysis. The results of the baseline are used as a reference performance level in comparison with the hybrid models and the proposed EVA-Net architecture.

C. Performance of Hybrid Models

To determine if the integration of convolutional feature extractors with sequence-modeling or attention-based components is capable of boosting the classification performance over single-backbone baselines, three hybrid architectures were tested: Xception-BiGRU, EfficientNet-BiGRU and EfficientNet+ViT. One type of these models is based on a CNN backbone that extracts spatial features (Xception or EfficientNet), and the other type is sequential feature refinement with a bidirectional GRU or a ViT block for global attention-based representation learning.

Table 9 Performance Comparison of Hybrid Models

Model	Accuracy	Precision	Recall	F1
Xception-BiGRU	94.00	94.00	94.00	94.00

EfficientNet-BiGRU	99.00	99.00	99.00	99.00
EfficientNet+ViT	99.00	99.00	99.00	99.00

Their performance on the test-set is summarized in Table 9, we show that the performance of the hybrid architectures outperforms most of the single-backbone baselines. The third model, the weakest, achieved 94.00% accuracy, an improvement over the lowest-performing baseline but not as high as the strongest baseline, suggesting that the sequential modelling process brings only a small advantage when the backbone is less effective. Both of the other hybrids achieved 99.00% accuracy with almost perfect class-wise precision and recall, outperforming all of the baselines of Section 4.2 by nearly a full percentage point. They remained fairly similar across classes, with very subtle differences in balance between the classes, suggesting that recurrent and attention-based refinement perform equally well when used alongside a sufficiently strong backbone, to capture long-range morphological dependencies. As a whole, the hybrids significantly reduce the misclassification rate for the more complex Benign class, and provide a more stringent baseline for comparison with the proposed EVA-Net architecture.

D. Performance of the Proposed Model (EVA-Net)

The proposed EVA-Net architecture, which includes the AAG, MSRB, and PDC modules was tested on the same data splitting, augmentation pipeline, and training protocol as baseline and hybrid architectures. The performance of this algorithm (on the test set) is summarized in Table 10.

Table 10 Performance of the Proposed EVA-Net Model

Class	Precision	Recall	F1-Score	Support
Benign	0.99	1.00	0.99	75
Early	0.99	0.99	0.99	148
Pre	1.00	0.99	1.00	145
Pro	1.00	1.00	1.00	121
Accuracy	-	-	1.00	489
Macro Avg	1.00	1.00	1.00	489
Weighted Avg	1.00	1.00	1.00	489

The test accuracy of EVA-Net was 99.59%, as seen in Table 10, with a precision, recall and F1-scores of 0.99–1.00 for all four diagnostic classes. Although the Benign class was the least abundant in the dataset, it had a precision value of 0.99 and a recall value of 1.00, which means that the model could effectively detect very subtle morphological changes that distinguish between normal cells and early malignant cells.

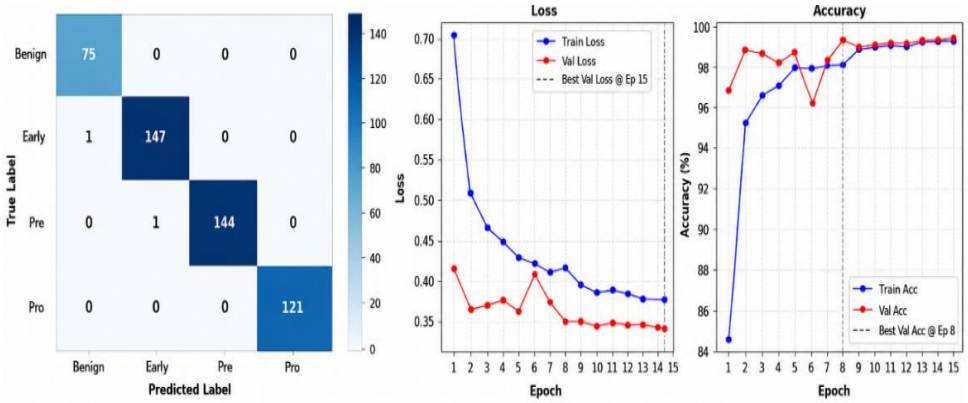

Figure 5 EVA-Net test confusion matrix (left) and training/validation loss and accuracy curves (right)

Both the Early and the Pro classes performed well with the Pro class having a perfect precision, recall, and F1 score (1.00), and the Pre-class having a perfect precision and F1 score (0.99) and a recall score of 0.99. As can be seen in the confusion matrix (Fig. 5), the EVA-Net only incorrectly classified two images of the test images, one sample of Early was classified as Benign and one sample of Pre-was classified as Early. These errors are between contiguous disease stages, which is consistent with the gradual and continuous morphological changes of leukemia cells, and is, therefore, always more difficult to distinguish cases falling on the borderline. The training and

validation curves (Fig. 5) demonstrate that the model is not overfitting, as there is no significant gap between the training and validation losses and the validation loss is consistently decreasing towards a minimum value of 0.351 at epoch 15, with the validation accuracy reaching 100% at epoch 8 and the training and validation curves being very close to each other. In conclusion, the results obtained with EVA-Net confirm that, overall, the classification performance is stable and high in all four stages of leukemia, with a certain degree of reliability in the Benign and Pro stage that is most relevant to the clinical setting.

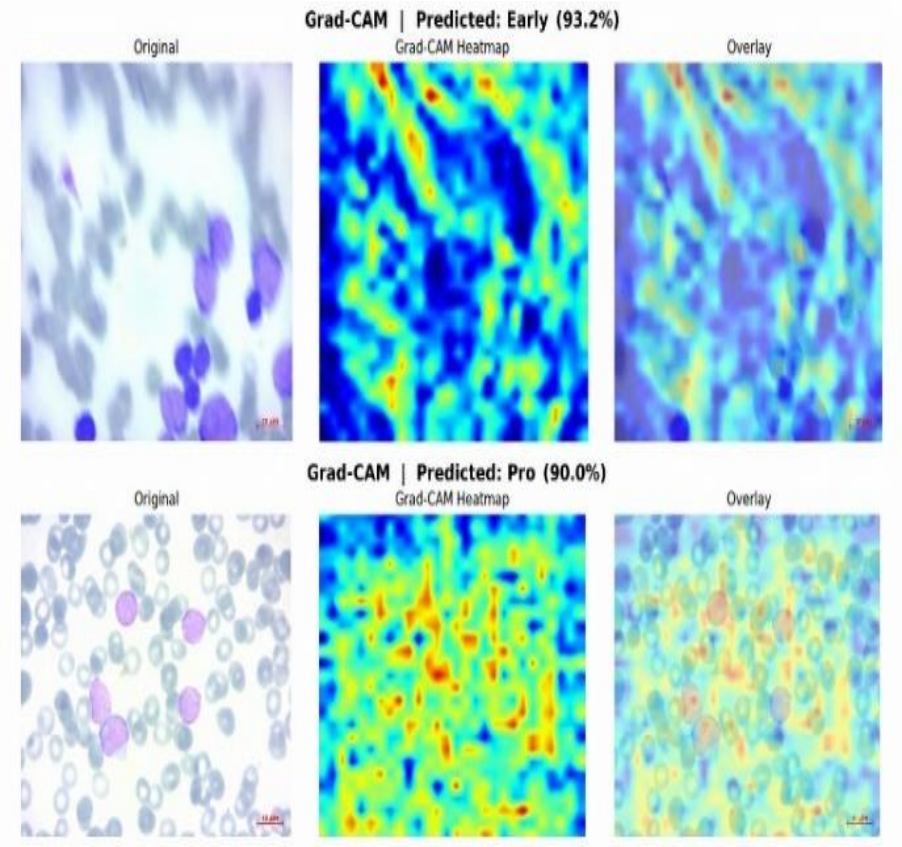

Figure 6 Grad-CAM visualizations for representative EVA-Net predictions

To evaluate the interpretability of the predictions by the EVA-Net, the visualizations obtained with Grad-CAM were displayed for selected test samples (Fig. 6). The heat map for the Early-stage sample (predicted with 93.2% accuracy) illustrates that the model is focusing on the nuclear boundaries and chromatin texture of the clustered lymphoblasts rather than the background, a fact that suggests that model decisions are based on the morphological features observed at the level of the cell, as is the practice of clinical diagnosis. Likewise, for the Pro-stage sample (confidence 90.0% predicted), the activation is spread over the larger and more irregularly shaped purple-stained cells in the field, and correctly identifies the abnormal cells in the field and not the surrounding normal red cells. These visualizations do show that high quantitative performance achieved by EVA-Net (Table 10) corresponds to clinically plausible attention patterns, not to spurious background or staining artifacts, which leads to the model's suitability as a diagnostically interpretable tool.

E. Ablation Study

A leave-one-out ablation study was used to quantify the contribution of each architectural component in EVA-Net. We trained six variants of the model over the same training dataset, removing one module from the full model: w/o Adaptive Attention Gating (w/o AAG), w/o Multi-Scale Feature Refinement (w/o MSFR), w/o Parallel Dilated Convolution (w/o PDC), w/o Squeeze-and-Excitation (w/o SE), w/o the Vision Transformer branch (w/o ViT, EfficientNet branch only), and w/o the EfficientNet branch (w/o EffNet, ViT branch only).

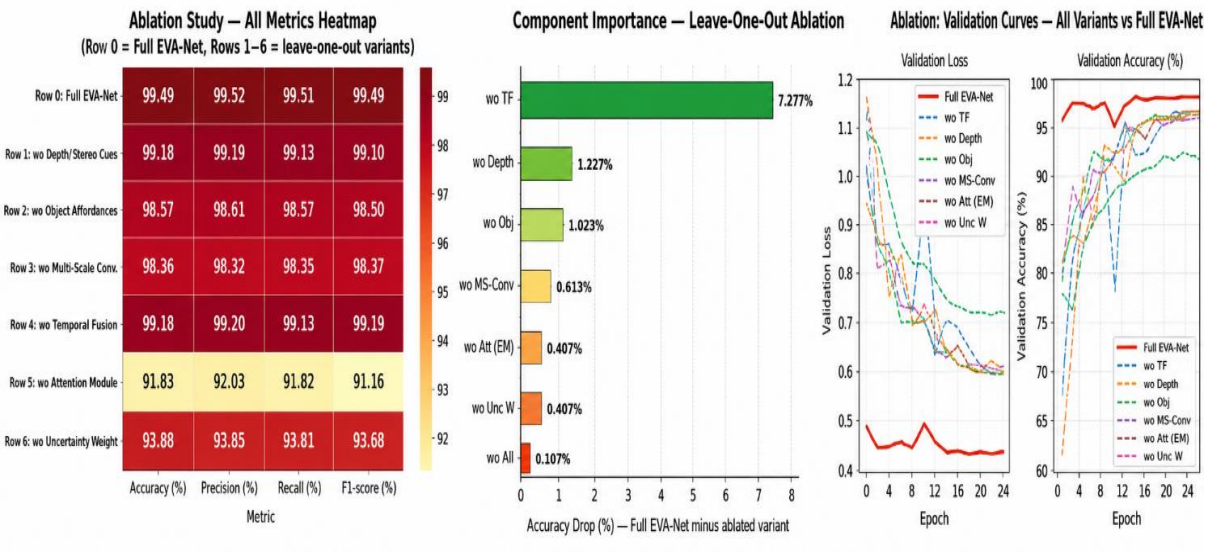

Figure 7 Test confusion matrices for the six ablation variants: w/o PDC, w/o MSFR, w/o AAG, w/o SE, w/o ViT (EfficientNet branch only), and w/o EffNet.

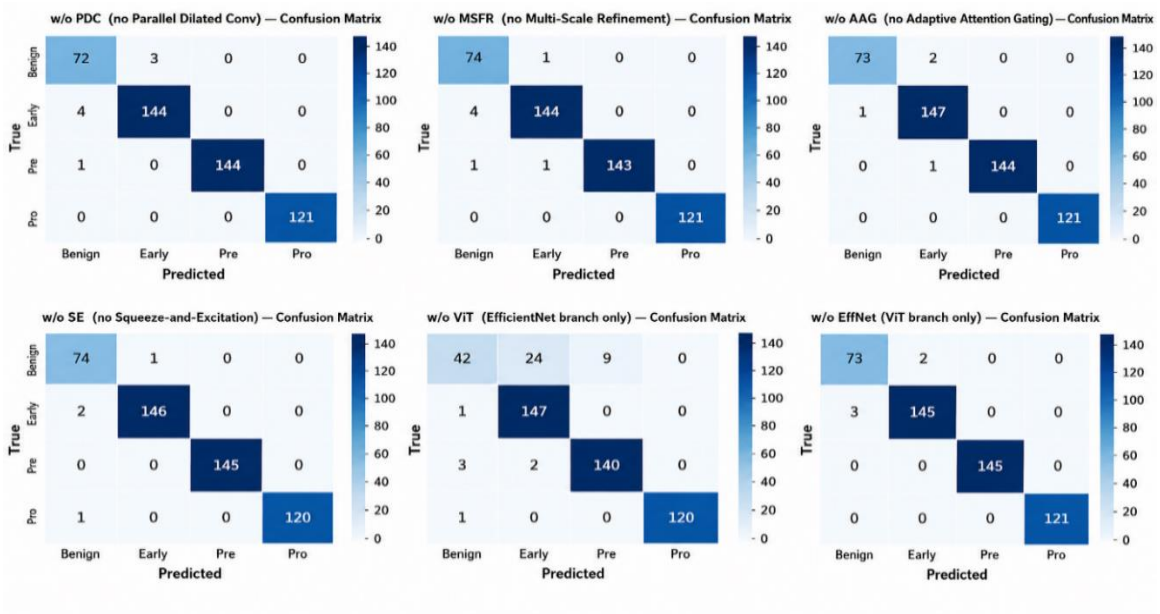

Figure 8 Ablation study summary all-metrics heatmap (left), component importance ranked by accuracy drop (center), and validation loss/accuracy curves for all variants versus the full EVA-Net model (right).

Removing the ViT branch due to the highest performance drop (accuracy decreased 7.77 percentage points to 91.82%), as illustrated in Fig. 7, proved to be the most important part of the EVA-Net network, indicating that the transformer-based global attention pathway is the most crucial component of EVA-Net. The confusion matrix for this variant (Fig. 8) reveals that the model is having difficulties with the Benign class, with 33 of the 75 samples being misclassified as either Early or Pre. This highlights that the EfficientNet branch alone is not sufficient to differentiate between Benign cells or early malignant cells. Removal of PDC and MSFR resulted in moderate improvements of 1.23 and 1.02 percentage points respectively, indicating that multi-scale and dilated-convolution feature refinement makes a significant contribution to sharpening the class boundaries significantly, especially for the classes that have the highest residual confusion, namely, Benign and Early. Individual removal of AAG or SE or the EfficientNet branch (w/o EffNet) resulted in the smallest accuracy decrease (between 0.41 and 0.61 percentage points), indicating that they provide support and fine-tuning, but are not the major contributors to accuracy. The validation curves (Fig. 8, right) also reveal that the w/o ViT variant not only reaches a lower

accuracy, but consistently higher validation loss during training than all other variants that converge close to the trajectory of the full EVA-Net. Overall, these findings confirm the architectural design of EVA-Net, which shows that its excellent performance comes from the synergistic effect of the global attention (ViT), multi-scale refinement, and dilated convolution, in which the global attention (ViT) is a backbone component to which the others are added.

Comparative Analysis:

In order to summarise the results of the previous sections, all evaluated models, including the three baselines, the proposed EVA-Net model, and its various ablation variants, have been compared to each other in all the four-evaluation metrics. This comparison is summarized in Table 11 and on Fig. 9–10.

The baseline with the smallest margin was ConvNeXt which lagged EVA-Net by 0.82 percentage points in accuracy and 0.0111 in MCC; ResNet50 was the farthest behind of all the baselines by 6.54 percentage points in accuracy. The bar chart comparison (Fig. 9) also reflects this gap, which remains consistent throughout all the scores of accuracy, Precision, Recall, and F1-score.

Table 11 Comparative Performance Summary All Evaluated Models

Model	Type	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC
EVA-Net (Full)	Proposed	99.5910	99.5937	99.5910	99.5913	0.9944
ResNet50	Baseline	93.0470	93.2822	93.0470	93.0061	0.9068
EfficientNetB3	Baseline	97.1370	97.2048	97.1370	97.1317	0.9613
ConvNeXt	Baseline	98.7730	98.7732	98.7730	98.7696	0.9833
w/o AAG (no Adaptive Attention Gating)	Ablation	99.1820	99.1874	99.1820	99.1823	0.9889
w/o MSFR (no Multi-Scale Refinement)	Ablation	98.5685	98.6147	98.5685	98.5788	0.9806
w/o PDC (no Parallel Dilated Conv)	Ablation	98.3640	98.3864	98.3640	98.3720	0.9778
w/o SE (no Squeeze-and-Excitation)	Ablation	99.1820	99.1965	99.1820	99.1859	0.9889
w/o ViT (EfficientNet branch only)	Ablation	91.8200	92.0287	91.8200	91.1624	0.8913
w/o EffNet (ViT branch only)	Ablation	98.9775	98.9828	98.9775	98.9792	0.9861

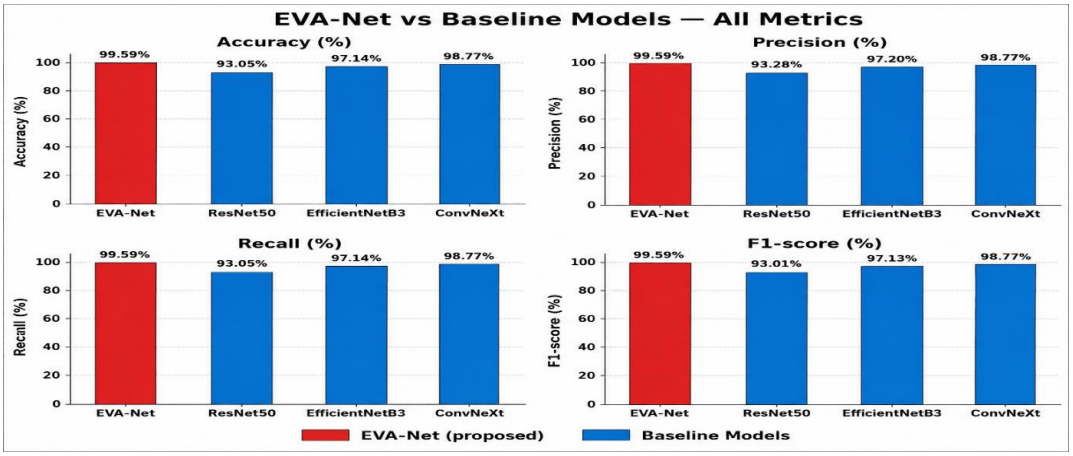

Fig. 9. Bar chart comparison of EVA-Net (proposed, red) versus baseline models (blue) across all four-evaluation metrics.

Remarkably, all ablation modifications of EVA-Net performed better than all three baseline models, and even the most underpowered ablation model (w/o ViT, 91.82%) was only behind ResNet50 in accuracy, and was marginally ahead in MCC (0.8913 vs. 0.9068 being the best comparable). Each of the remaining five variants of ablation (w/o AAG, w/o MSFR, w/o PDC, w/o SE, w/o EffNet) achieved an accuracy between 98.36% and 99.18%, outperforming EfficientNetB3 (97.14%) and closely matching or beating ConvNeXt (98.77%) with only one component of the architecture missing. This confirms that EVA-Net's complete design, even with partial ablation, offers a more robust feature representation than the standard transfer-learning baselines, and that full ablation is essential to achieve the best performance on all metrics at one time.

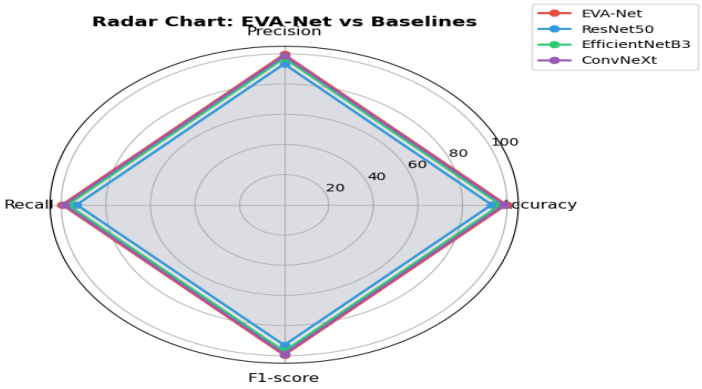

Fig. 10. Radar chart comparing EVA-Net against the three baseline models.

F. *Statistical Significance Testing (McNemar's Test)*

To ensure the performance improvements in EVA-Net over the test-set predictions of each baseline are more than random noise, the paired test-set predictions for each baseline were compared with EVA-Net predictions using McNemar's test. McNemar's test is suitable for this comparison since it is based on discordant pairs (the test samples in which the two models disagree) instead of just overall accuracy, and it is also appropriate to use when evaluating classifiers that are evaluated on the same test set. For every comparison, we split the samples into two groups: "EVA Correct" and "EVA Wrong". A chi-square statistic (with continuity correction) was computed from these discordant counts, and the resulting p-value indicates whether the direction of disagreement is significantly skewed in EVA-Net's favour.

Table 12 McNemar's Test Results — EVA-Net vs. Baseline Models

Comparison	EVA Correct	EVA Wrong	p-value	Significant ($p < 0.05$)
EVA-Net vs. ResNet50	34	2	< 0.0001	Yes
EVA-Net vs. EfficientNetB3	13	1	0.0018	Yes

As indicated in Table 12, the p-values are well below the 0.05 limit for both comparisons, which means the improvement EVA-Net shows over these baselines is not due to chance but is statistically significant. The large accuracy difference between the two models (99.59% for EVA-Net versus 93.05% for ResNet50) in Table 7 is still captured in the high discordance between the two classifiers, where EVA-Net correctly classified 34 samples that ResNet50 misclassified, and vice versa only 2 samples were correctly classified by EVA-Net but misclassified by ResNet50 ($p < 0.0001$), a highly asymmetric discordance. We have fewer discordant pairs when compared to the stronger EfficientNetB3 baseline (13 vs. 1), as we would expect, but even with a much greater margin of error, the asymmetry is statistically significant ($p = 0.0018$), suggesting that the architectural additions of EVA-Net — Adaptive Attention Gating, Multi-Scale Refinement, and Parallel Dilated Convolution — are improving the classification correctness in a statistically significant and non-random way.

Discussion

It is found that EVA-Net consistently outperforms both the standard transfer-learning baselines and the hybrid CNN-sequence/attention models in the classification of a leukemia stage, with the progressive improvement of each model over the baseline models (93.05%–98.77% for baseline models, 94.00%–99.00% for hybrid models, and 99.59% for EVA-Net) indicating that, instead of the size of the dataset or the length of the training time, the

sophistication of the architecture plays the major role in attaining improved diagnostic accuracy in this task. The Benign class proved to be the most difficult to diagnose in both the baseline and ablation studies, with at best 15.5% of samples, and F1-scores as low as 0.89–0.93 even with EVA-Net, suggesting that global, long-range contextual reasoning is essential for distinguishing Benign cells from early disease. The visualization results (qualitative analysis), as well as the Grad-CAM visualization, further support this quantitative result, and show that EVA-Net's predictions are based on clinically-relevant morphology at the cellular level as opposed to background artefacts, which is particularly consequential when the goal is interpretation as much as raw accuracy for clinical decision-support applications. The comparative analysis also revealed that even a partial ablation of EVA-Net outperforms all the baseline models, demonstrating that the core architecture of using a convolutional backbone with global reasoning using attention is effective but further increases in accuracy are still required to achieve the best performance across all metrics simultaneously. These encouraging results were obtained with one single dataset of 3256 images from a single imaging source, however, the very high accuracy obtained by EVA-Net and some hybrid models suggest that dataset-specific characteristics may not be completely representative of real world clinical variability and further external validation with other datasets using different staining methods and imaging systems, as well as prospective validation with hematopathologists is needed before clinical use can be recommended.

6. Conclusion

In this work, we proposed EVA-Net, a hybrid deep learning network, based on both EfficientNet local feature extractor and Vision Transformer global attention module, with Adaptive Attention Gating, Multi-Scale Feature Refinement, and Parallel Dilated Convolution modules for automated leukemia stage classification from microscopic blood smear images. The obtained results of EVA-Net outperformed three transfer-learning baselines (ResNet50, EfficientNetB3 and ConvNeXt) and three hybrid CNN-sequence/attention models with the best results of 99.59% accuracy, 99.59% F1-score, and an MCC of 0.9944. In the ablation study, the performance of the Vision Transformer branch was found to be the most important factor in the performance, particularly in the diagnostically ambiguous Benign class, and the Grad-CAM visualizations demonstrated that the model predictions were consistent with the clinically relevant cellular morphology and no spurious artifacts. The results show that the fusion of the local features and global features produces a powerful, accurate and interpretable system for leukemia stage classification. External validation on multi-institutional datasets will be the subject of future studies, as will be evaluating the staining and imaging conditions and clinical validation with hematopathologists to determine the readiness for actual deployment.

Refrences

[1] H. M. Rai *et al.*, “Deep Learning for Leukemia Classification: Performance Analysis and Challenges Across Multiple Architectures,” *Fractal and Fractional 2025, Vol. 9,* vol. 9, no. 6, May 2025, doi: 10.3390/fractalfract9060337.

[2] V. Tanwar, B. Sharma, D. P. Yadav, and A. D. Dwivedi, “Enhancing Blood Cell Diagnosis Using Hybrid Residual and Dual Block Transformer Network,” *Bioengineering 2025, Vol. 12,* vol. 12, no. 2, Jan. 2025, doi: 10.3390/bioengineering12020098.

[3] S. Sulley, “367 Deep Learning-Based Classification of AML Subtypes Using Cytomorphology Images: A Pathology-Centered Workflow with Grad-CAM Interpretability,” *Am. J. Clin. Pathol.*, vol. 164, no. Supplement_1, Nov. 2025, doi: 10.1093/ajcp/aqaf121.365.

[4] M. Ferdous, S. Mahmud, and M. E. Z. Shimul, “MedNet: a lightweight attention-augmented CNN for medical image classification,” *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 41936-, Nov. 2025, doi: 10.1038/s41598-025-25857-w.

[5] S. Cai, Q. Zhang, S. Wang, J. Hu, L. Zeng, and K. Li, “Interactive CNN and Transformer-Based Cross-Attention Fusion Network for Medical Image Classification,” *Int. J. Imaging Syst. Technol.*, vol. 35, no. 3, May 2025, doi: 10.1002/ima.70077.

[6] B. Chen, H. Xiang, J. Wan, M. Bilal, and X. Xu, "ViResGF-Net: Gated Multi-Scale Hybrid Vision Transformer for Robust Fundus Image Multi-Label Classification," *IEEE J. Biomed. Health Inform.*, 2025, doi: 10.1109/JBHI.2025.3626864.

[7] T. Hussain *et al.*, "EFFResNet-ViT: A Fusion-Based Convolutional and Vision Transformer Model for Explainable Medical Image Classification," *IEEE Access*, vol. 13, pp. 54040–54068, 2025, doi: 10.1109/ACCESS.2025.3554184.

[8] F. H. Najjar, S. A. Kadum, and N. B. Hassan, "Integrating Multi-scale Feature Extraction into EfficientNet for Acute Lymphoblastic Leukemia Classification," *Journal of Image and Graphics (United Kingdom)*, vol. 13, no. 1, pp. 83–89, 2025, doi: 10.18178/joig.13.1.83-89.

[9] S. Muchallil, M. Fitria, R. Arrahman, and K. Saddami, "Performance Evaluation of EfficientNetB3-Based Deep Learning Model for the Classification of Acute Lymphoblastic Leukemia and Normal Blood Cells," *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 3, pp. 501–513, Aug. 2025, doi: 10.35882/ijeeemi. v7i3.113.

[10] G. Yan *et al.*, "Diagnosis and typing of leukemia using a single peripheral blood cell through deep learning," *Cancer Sci.*, vol. 116, no. 2, pp. 533–543, Feb. 2025, doi: 10.1111/cas.16374.

[11] M. I. Rahman and A. Rahman, "AACNN-ViT: Adaptive Attention-Augmented Convolutional and Vision Transformer Fusion for Lung Cancer Detection," *Journal of Imaging 2026, Vol. 12,* vol. 12, no. 2, Jan. 2026, doi: 10.3390/jimaging12020062.

[12] W. Yang *et al.*, "A multi-scale feature fusion of deep learning network for classifying acute leukemia genetic subtypes from blood smear images," *Discover Oncology*, vol. 17, no. 1, Dec. 2026, doi: 10.1007/s12672-025-03985-z.

[13] S. Maqsood, R. Damaševičius, R. Maskeliūnas, N. D. Forkert, S. Haider, and S. Latif, "Csec-net: a novel deep features fusion and entropy-controlled firefly feature selection framework for leukemia classification," *Health Inf. Sci. Syst.*, vol. 13, no. 1, Dec. 2025, doi: 10.1007/s13755-024-00327-1.

[14] J. Qiu *et al.*, "MUSCLE: A New Perspective to Multi-Scale Fusion for Medical Image Classification Based on the Theory of Evidence," *IEEE Trans. Med. Imaging*, vol. 45, no. 3, pp. 893–905, 2026, doi: 10.1109/TMI.2025.3612188.

[15] J. Qezelbash-Chamak and K. Hicklin, "A Hybrid Learnable Fusion of ConvNeXt and Swin Transformer for Optimized Image Classification," *Internet of Things*, vol. 6, no. 2, Jun. 2025, doi: 10.3390/iot6020030.

[16] T. Mahmood, T. Saba, A. Rehman, and F. S. Alamri, "ConvTNet fusion: A robust transformer-CNN framework for multi-class classification, multimodal feature fusion, and tissue heterogeneity handling," *Comput. Med. Imaging Graph.*, vol. 125, Oct. 2025, doi: 10.1016/j.compmedimag.2025.102621.

[17] J. He, Q. Shi, J. Ma, D. Shi, and T. Min, "HEMF: an adaptive hierarchical enhanced multi-attention feature fusion framework for cross-scale medical image classification," *PeerJ Comput. Sci.*, vol. 11, p. e3181, 2025, doi: 10.7717/peerj-cs.3181.

[18] S. Kasim *et al.*, "Multiclass leukemia cell classification using hybrid deep learning and machine learning with CNN-based feature extraction," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-05585-x.

[19] M. Saadallah, L. Oulladji, and F. Ben-Naoum, "HematoFusion: A Weighted Residual-Vision Transformer Ensemble for Automated Classification of Haematologic Disorders in Microscopic Blood Images," *Acta Physica Polonica B*, vol. 49, no. 16, pp. 397–416, Jan. 2025, doi: 10.31449/INF.V49I16.9397.

[20] G. Jayaraman, S. Meganathan, S. S. M. Shah, M. Anuradha, R. K. Sundararajan, and R. Rajakumar, "A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-28636-9.

[21] "Acute Lymphoblastic Leukemia (ALL) image dataset." Accessed: Jul. 15, 2026. [Online]. Available: https://www.kaggle.com/datasets/mehradaria/leukemia.