Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00380-9

Original Article

Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM

Purna Siswantomo [1], Ahmad Ilham [1], Wendy Sarasjati [1]

1. *Informatics, Universitas Muhammadiyah Semarang, Semarang, IDN*

Received: June 17, 2026 | Review began: September 25, 2026 | Review ended: October 08, 2026 | Published: November 19, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Malaria remains a major public health challenge, particularly in endemic regions with limited access to trained microscopists and laboratory infrastructure. Conventional microscopy is labor-intensive, expertise-dependent, and affected by observer fatigue and inter-observer variability. In these settings, automated support is more likely to be trusted and used safely when microscopists can check the basis of each prediction. This study presents an explainable deep-learning benchmark for automated malaria cell classification using the NIH Malaria Cell Images Dataset, comprising 27,558 thin blood smear images. Five convolutional neural network backbones (EfficientNetV2-S, VGG16, VGG19, ResNet34, and ResNet50) were compared, and MixUp and CutMix augmentation were examined for their effects on generalization. Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to show which cell regions drove each prediction, so that a reviewer can judge whether a positive call is supported by parasite-like structures or by staining artifacts.

Models were evaluated with 5-fold stratified cross-validation using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC), together with learning curve analysis. ROC-AUC was the primary metric for the Friedman test with Nemenyi post-hoc analysis. EfficientNetV2-S achieved its best performance under the CutMix scenario, with a mean accuracy of 97.56 ± 0.30%, ROC-AUC of 99.72 ± 0.06%, and F1-score of 97.53 ± 0.31%. The Friedman test showed significant differences among backbones in all augmentation scenarios, also after Holm correction, although the absolute ROC-AUC differences were below 0.002. Nemenyi post-hoc analysis showed the significant superiority of EfficientNetV2-S over ResNet34 in all scenarios and over VGG19 under CutMix, but not over every baseline.

Grad-CAM maps for cases selected by a fixed rule suggested that EfficientNetV2-S focused on parasite-like stained inclusions in correctly classified infected cells, whereas misclassified cells had probabilities close to the decision threshold and activations on faint staining irregularities or cell boundaries. A deletion-insertion check on 200 validation images supported the faithfulness of these maps for parasitized predictions but not for uninfected predictions. These observations were not validated by experts, but such heatmaps could help a microscopist confirm or reject a flagged cell. EfficientNetV2-S had the fewest parameters (20.84 million), the lowest number of multiply-accumulate operations (3.54 GMACs), and the smallest model size (79.49 MB), but was slower on GPU than all four baselines (23.87 vs. 6.12-10.80 ms per image), so its compactness refers to size rather than speed. Overall, EfficientNetV2-S is a promising backbone for automated malaria cell classification, offering strong discriminative performance, a compact model size, and qualitative Grad-CAM observations that may support human-in-the-loop screening. External validation on clinically diverse datasets and expert assessment of explainability outputs are required before clinical deployment.

Categories: Advances in Biomedical data analysis, Explainable AI, Deep Learning

Keywords: malaria cell classification, efficientnetv2, deep learning, grad-cam, explainable ai, transfer learning, thin blood smear

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

Introduction

Malaria is a parasitic disease caused by protozoa of the genus Plasmodium and transmitted through the bites of infected female Anopheles mosquitoes. According to the World Health Organization World Malaria Report 2025, there were an estimated 282 million malaria cases globally in 2024, with an incidence of 64 cases per 1,000 population at risk. In the same year, malaria caused an estimated 610,000 deaths worldwide, with a mortality rate of 13.8 deaths per 100,000 population at risk [1]. These figures show that malaria remains a major public health problem despite substantial progress in global intervention programs, particularly in endemic regions with limited diagnostic infrastructure [1]. In Indonesia, malaria transmission is still concentrated in several eastern regions, including Papua, Maluku, East Nusa Tenggara, and Southeast Sulawesi, where access to trained microscopists and laboratory facilities is often limited [2]. Accurate and timely diagnosis is therefore essential to support early treatment and reduce the risk of severe complications.

Peripheral blood smear examination under a light microscope remains widely used as a reference method for malaria diagnosis because it enables parasite visualization and supports species identification and parasitemia estimation [3]. However, this method has several practical limitations. Its accuracy depends heavily on the expertise of trained microscopists, while the examination process is labor-intensive and prone to observer fatigue and inter-observer variability. In addition, sensitivity may decrease in cases of low parasitemia, which are commonly found during the early stages of infection. These limitations highlight the need for computer-aided diagnostic systems that can support microscopic image interpretation, especially in resource-limited settings.

Deep learning has shown considerable potential in automated medical image analysis. Convolutional Neural Networks and their variants have achieved competitive performance in various medical imaging tasks, including radiology, dermatology, histopathology, and microscopy [4]. Image analysis and machine learning methods for malaria microscopy have been widely investigated [3]. More recent deep learning studies on thin blood smear images have explored transfer learning, EfficientNet-based architectures, ensemble learning, and custom CNN models. Transfer learning studies have evaluated architectures such as ResNet, VGG, MobileNet, and Inception on the NIH Malaria Cell Images Dataset, with several studies reporting high classification performance [5-10], and EfficientNet-B0 and EfficientNet-B7 have been directly compared for malaria blood cell classification [11]. Ensemble-based approaches have also been introduced to improve robustness by combining multiple CNN feature extractors or decision-level voting strategies [12-14]. Nevertheless, ensemble methods may increase computational requirements, model size, and deployment complexity, which are important constraints for malaria screening in remote or resource-limited healthcare settings. At the same time, studies on custom architectures and cross-dataset validation have shown that strong performance on the NIH dataset does not always guarantee reliable generalization under real clinical image acquisition conditions [15-18].

Although these studies demonstrate the potential of deep learning for malaria diagnosis, several methodological gaps remain. First, to our knowledge, few studies have integrated modern efficient CNN backbones, classical CNN baselines, augmentation ablation, statistical testing, computational efficiency, and explainability within one unified protocol. Second, among the 15 previously published malaria studies summarized in Table *1*, six described conventional geometric augmentation such as rotation, zooming, shifting, and flipping, three mentioned augmentation without specifying the techniques, and six did not report image augmentation; none used mixing-based strategies such as MixUp or CutMix. To our knowledge, these strategies have not been evaluated in a unified ablation framework for malaria cell classification. Third, visual explainability is not always treated as a core part of model evaluation, even though interpretability is important in medical diagnostics; only one of the studies in Table *1* reported saliency-based explanations. Explainability can help verify whether a model focuses on biologically meaningful regions rather than image artifacts or irrelevant background patterns [19-25].

Study	Dataset	Model(s)	Image augmentation	MixUp/CutMix	XAI

Muhandisin and Azhar [5]	NIH	Fine-tuned ResNet50	Rotation (30°), zoom (0.3), horizontal flip	No	No
Jameela et al. [6]	NIH	ResNet34, ResNet50, VGG16, VGG19	Resizing and normalization only	No	No
Laghari et al. [7]	NIH	Eight transfer-learning CNNs (e.g., ResNet50, ResNet101, Xception)	Resizing and normalization; augmentation mentioned without details	No	No
Hemachandran et al. [8]	NIH	CNN, MobileNetV2, ResNet50	Not reported	No	No
Mujahid et al. [9]	NIH	Modified EfficientNet-B2 vs. 11 CNNs	Resizing only	No	No
Boit and Patil [10]	NIH	EDRI (EfficientNetB2-Dense-Residual-Inception hybrid)	Rotation, shift, zoom, shear, flip	No	No
Maulana et al. [11]	NIH	EfficientNet-B0, EfficientNet-B7	Rotation (30°), zoom, width/height shift, horizontal flip	No	No
Shahin et al. [12]	NIH	ResNet50, VGG16, DenseNet201 feature fusion with SVM/LSTM	Resizing only	No	No
Bhuiyan and Islam [13]	NIH	VGG16, VGG19, DenseNet201 ensemble	Rotation, zoom, horizontal/vertical flip	No	No
Reddy et al. [14]	NIH	CNN with VGG16, VGG19, ResNet50V2, DenseNet201	Rotation (±30°), zoom (up to 20%), horizontal/vertical flip	No	No
Siłka et al. [15]	BBBC041	Custom encoder-decoder CNN	Horizontal/vertical flip, rotation, zoom	No	No

Acherar et al. [16]	Private clinical dataset + NIH	Custom CNN, VGG19, ResNet50, EfficientNet-B7	Not reported (dataset enriched with artifact images)	No	No
Swastika et al. [18]	NIH	ResNet50V2, EfficientNetB0, InceptionV3	Used in selected experiments; techniques not specified	No	No
Parveen et al. [21]	Kaggle malaria collection + Harvard Dataverse	MobileNetV2, VGG19, InceptionV3, ResNet18, Xception, Inception-ResNetV2	Mild augmentation; techniques not specified	No	Grad-CAM, LIME, SHAP
Rajaraman et al. [26]	NIH	AlexNet, VGG-16, Xception, ResNet50, DenseNet-121, custom CNN	Not reported	No	No
This study	NIH	EfficientNetV2-S, VGG16, VGG19, ResNet34, ResNet50	Geometric + MixUp, CutMix, MixUp + CutMix (ablation)	Yes	Grad-CAM (TP/FP/FN/TN)

TABLE 1: Data augmentation and explainability in previously published deep learning studies on malaria cell images cited in this study.

To address these gaps, this study presents an explainable deep learning benchmark for malaria cell classification using the NIH Malaria Cell Images Dataset. EfficientNetV2-S is compared with four classical CNN backbones, namely VGG16, VGG19, ResNet34, and ResNet50, using the same preprocessing, training, and evaluation pipeline within each augmentation scenario. Four augmentation scenarios are evaluated: no augmentation, MixUp, CutMix, and MixUp + CutMix. Model performance is assessed using 5-fold stratified cross-validation and multiple classification metrics, with receiver operating characteristic curve (ROC-AUC) used as the primary metric for non-parametric statistical analysis through the Friedman test, Nemenyi post-hoc test, and Kendall's W effect size. Learning curve analysis is also conducted to examine training stability, convergence behavior, and possible signs of overfitting or underfitting across epochs. Computational efficiency is evaluated using parameter count, multiply-accumulate operations (MACs), model size, latency, and throughput, while Gradient-weighted Class Activation Mapping (Grad-CAM) is used to provide visual interpretation for representative correct and incorrect predictions.

The main contributions of this study are as follows:

1. A unified benchmark on the NIH Malaria Cell Images Dataset comparing EfficientNetV2-S with VGG16, VGG19, ResNet34, and ResNet50 under the same 256 × 256 preprocessing pipeline, ImageNet initialization, binary classification head, 5-fold stratified cross-validation, ROC-AUC-based non-parametric statistical testing, and learning curve analysis.

2. An ablation study of no augmentation, MixUp, CutMix, and MixUp + CutMix, including their probability settings and Beta distribution parameters, to evaluate the individual and combined effects of augmentation strategies on malaria cell classification.

3. A Grad-CAM-based explainability analysis on representative True Positive, False Positive, False Negative, and True Negative cases to examine whether EfficientNetV2-S focuses on morphologically relevant regions in Parasitized and Uninfected cell images.

4. A computational efficiency analysis covering parameter count, MACs, model size, CPU/GPU inference latency, and throughput to support practical model selection for resource-constrained healthcare environments.

Materials And Methods

Dataset

This study used the NIH Malaria Cell Images Dataset, curated and published by Rajaraman et al. [26]. The dataset contains 27,558 single-cell images segmented from Giemsa-stained thin blood smear slides of 150 *Plasmodium falciparum*-infected and 50 healthy patients. The slides were collected and photographed with a smartphone camera at Chittagong Medical College Hospital, Bangladesh, and the segmented cells were manually annotated by an expert slide reader at the Mahidol-Oxford Tropical Medicine Research Unit [26]. It consists of two balanced classes: 13,779 images of *P. falciparum*-infected cells labeled as Parasitized and 13,779 images of uninfected cells labeled as Uninfected. Because the two classes contain the same number of samples, explicit class reweighting was not applied during training. Instead, robustness was addressed through Focal Loss and data augmentation; these methods were used to reduce the influence of visually easy samples and increase training variability rather than to correct numerical class imbalance.

The images vary in spatial resolution, ranging from 40 × 40 pixels to 385 × 394 pixels, with an average size of approximately 130 × 130 pixels. This variation reflects differences in cell morphology, staining quality, and microscopic image acquisition conditions.

This dataset also has several limitations. All images were obtained from a single hospital with one acquisition setup, so variation in staining protocols, microscopes, and cameras across healthcare facilities is not represented. The Parasitized class contains only *P. falciparum*, so other *Plasmodium* species are not covered. The cells were pre-segmented by an automated algorithm and labeled individually, which differs from routine diagnosis, in which microscopists examine whole fields of view containing overlapping cells, artifacts, and other blood components. As the 27,558 images were derived from only 200 patients, many images share patient-specific characteristics such as staining intensity and background appearance, which may make sample-level validation optimistic. Finally, each cell was annotated by a single expert reader, so labeling errors cannot be fully excluded. These limitations are further considered in the Clinical Implications and Limitations section.

To obtain a more reliable estimate of model generalization, this study used 5-fold stratified cross-validation. The fold partitioning was performed at the sample level, where individual images were divided while maintaining the class proportion in each fold. In each iteration, approximately 80% of the data were used for training and 20% were used for validation. This process was repeated five times, ensuring that each sample appeared once in the validation set. All performance results were reported as the mean and standard deviation across the five folds. Patient-level splitting was not applied in this study, so cells from the same patient may appear in both the training and validation folds of the same iteration. Therefore, the results should be interpreted as sample-level validation rather than independent external or patient-level validation.

Preprocessing and augmentation strategies

The preprocessing pipeline ensured consistent input dimensions across all evaluated CNN backbones while aligning the input distribution with ImageNet pretraining statistics. All images were resized to 256 × 256 pixels and normalized using the ImageNet mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225]. In the no-augmentation scenario, resizing and normalization were the only operations applied to both training and validation images.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

In the MixUp, CutMix, and MixUp + CutMix scenarios, training images were additionally processed with a basic augmentation pipeline before batch-level mixing. This pipeline consisted of RandomResizedCrop with an output size of 256 × 256 pixels and a crop scale range of 0.85-1.0 (replacing direct resizing), horizontal flipping with $p = 0.5$, vertical flipping with $p = 0.2$, affine transformation with $p = 0.5$ (scaling of 0.9-1.1, translation of up to 5%, and rotation between −10° and +10°), and ColorJitter with $p = 0.3$ using the default Albumentations parameters. Validation images were always resized directly to 256 × 256 pixels without augmentation in all scenarios.

In addition to basic augmentation, this study evaluated MixUp and CutMix as advanced batch-level augmentation strategies. MixUp generates synthetic training samples by linearly interpolating pairs of images and labels. This encourages smoother decision boundaries and improves model regularization [27]. However, MixUp can also create visually blurred or biologically unnatural samples, which may be less suitable when diagnostic cues are very small and localized. CutMix replaces a rectangular region of one image with a region from another image and adjusts the target label according to the proportion of the mixed area. This encourages the model to learn more discriminative local image regions [28]. These characteristics are relevant to malaria cell classification because parasite-related visual cues often appear as small local structures inside red blood cells.

Four augmentation scenarios were evaluated: no augmentation, MixUp, CutMix, and MixUp + CutMix. The no-augmentation scenario served as a minimal reference in which training images were only resized and normalized, whereas the three mixing scenarios combined the basic augmentation pipeline described above with the respective batch-level mixing strategy. Consequently, differences between the no-augmentation scenario and each mixing scenario reflect the combined effect of basic and mixing-based augmentation, while differences among the three mixing scenarios, which share the same basic augmentation, isolate the effect of the mixing strategy. When MixUp was enabled, it was applied with a probability of 0.5. CutMix was evaluated only when MixUp was not applied, making both methods mutually exclusive within a single batch. When MixUp and CutMix were both enabled, the effective probabilities were 0.5 for MixUp, 0.25 for CutMix, and 0.25 for neither method. MixUp used a Beta distribution with alpha = 0.4, while CutMix used alpha = 1.0. These values were fixed across all backbones and folds to maintain a controlled comparison. The MixUp alpha value was selected as a moderate mixing strength based on the original MixUp formulation [27], while CutMix alpha = 1.0 was selected following the common setting used in the original CutMix strategy [28]. No separate hyperparameter search was conducted for alpha values, which should be considered a limitation of the augmentation analysis.

As MixUp and CutMix mix both images and labels, the training target was represented as a soft label rather than a single hard label. In this study, Focal Loss was used as the base binary loss, and the mixed-sample loss was computed as a weighted sum of the Focal Loss values from both original labels. For a mixed sample generated from two images x_i and x_j with labels y_i and y_j, the loss was computed as follows:

$$\mathcal{L}_{mixed} = \lambda\,\mathcal{L}(f(\tilde{x}), y_i) + (1-\lambda)\,\mathcal{L}(f(\tilde{x}), y_j) \quad (1)$$

where $\tilde{x}$ is the mixed input, $f(\tilde{x})$ is the model output for that input, y_i and y_j are the labels of the two mixed images, $L(\cdot)$ is the Focal Loss function, and λ is the mixing coefficient ($0 \le \lambda \le 1$). In MixUp, λ was sampled directly from the Beta distribution. In CutMix, λ was set according to the proportion of the image area replaced by the inserted patch. Therefore, the contribution of each label to the final loss reflected the degree of image mixing. The hybrid setting was intended to combine the global distribution-smoothing effect of MixUp with the local feature regularization effect of CutMix, while preserving a mutually exclusive batch-level application strategy to avoid excessive label mixing [28-30].

The overall preprocessing and augmentation pipeline, including the four evaluated scenarios, is summarized in Figure *1*.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

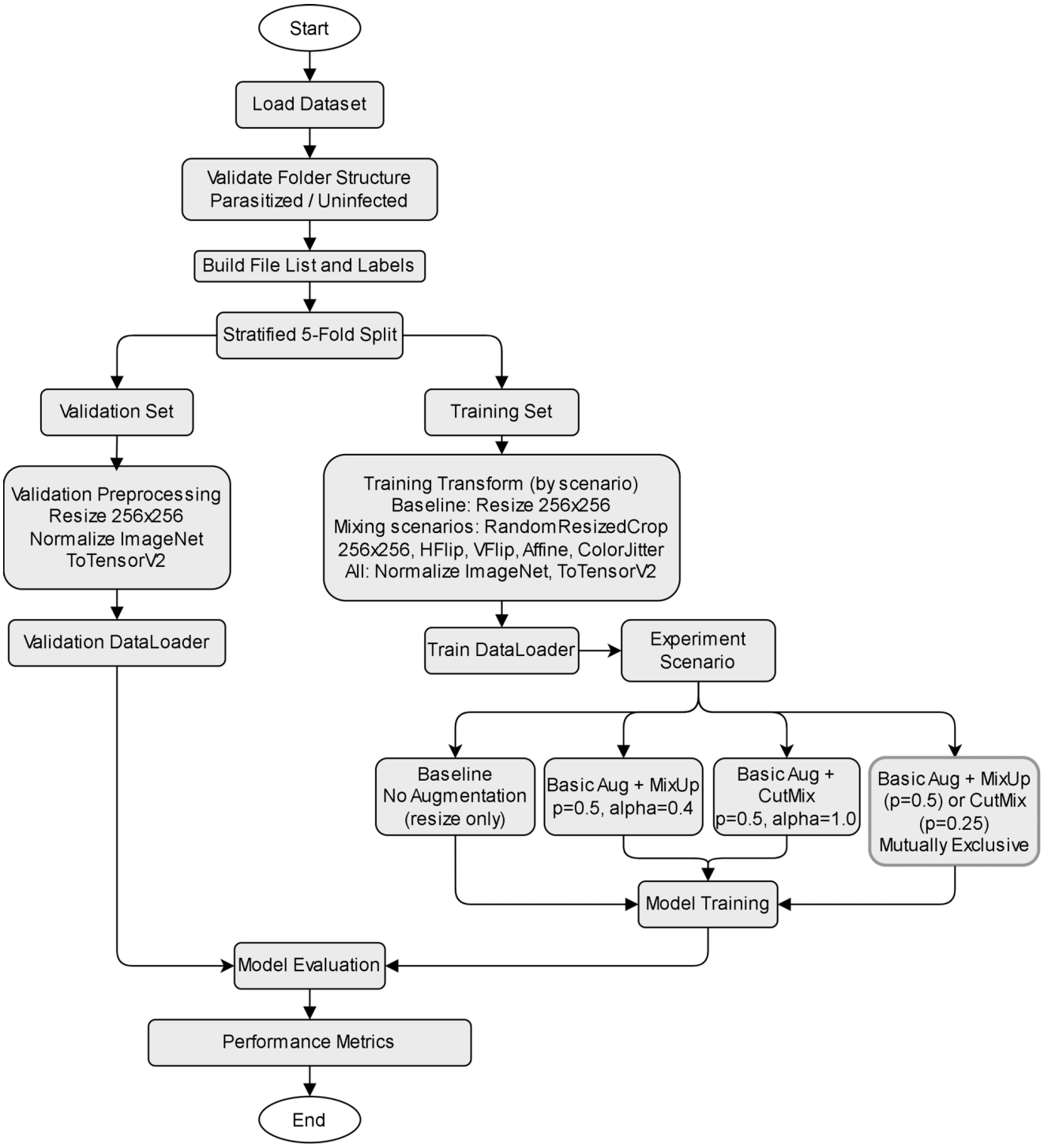


FIGURE 1: Flow diagram of the preprocessing and augmentation pipeline with four scenarios applied to malaria cell images.

CNN backbone architectures

Five CNN backbones were selected to represent different generations of convolutional architectures: VGG16, VGG19, ResNet34, ResNet50, and EfficientNetV2-S. VGG16 and VGG19 were included as classical CNN baselines because of their simple sequential architecture, consistent 3 × 3 convolutional filters, and long-standing use as transferable feature extractors [30]. ResNet34 and ResNet50 were selected to represent residual learning architectures, which use shortcut

connections to improve gradient flow and reduce optimization difficulty in deeper networks [31]. EfficientNetV2-S was selected as the main modern efficient backbone because it combines compound scaling with Fused-MBConv blocks to improve the trade-off between representational capacity and computational efficiency [32].

All backbones were initialized with ImageNet-1K pretrained weights. The original classifier layer of each model was disabled and replaced with the same binary classification head consisting of Dropout 0.4, a 512-unit fully connected layer, GELU activation, layer normalization, Dropout 0.4, and a final one-unit linear output. The final output was treated as a logit during training, and sigmoid activation was applied only during inference to obtain class probabilities. The same input size, preprocessing pipeline, training protocol, and evaluation metrics were applied across all backbones to keep the comparison as controlled as possible.

Architecture	Architectural family	Main characteristic	Role in this study
VGG16	Sequential CNN	Stacked 3 × 3 convolutions	Classical baseline
VGG19	Sequential CNN	Deeper VGG variant	Classical baseline
ResNet34	Residual CNN	Basic residual blocks	Residual baseline
ResNet50	Residual CNN	Bottleneck residual blocks	Deeper residual baseline
EfficientNetV2-S	Efficient CNN	Fused-MBConv and compound scaling	Main efficient backbone

TABLE 2: CNN backbone architectures evaluated in this study.

Training protocol

All models were trained using ImageNet-1K pretrained weights to leverage general visual representations learned from large-scale natural image data. The final classification layer of each backbone was replaced with a new binary classification head to distinguish Parasitized and Uninfected cell images. Although ImageNet pretraining is useful for transfer learning, natural-image pretraining may not fully match microscopy images because of domain differences in color, texture, scale, and object morphology. Therefore, all pretrained models were fine-tuned on malaria cell images using the same training protocol to reduce domain mismatch and allow a controlled comparison across backbones.

This study used Focal Loss [33], with $\alpha = 0.25$ and $\gamma = 2.0$, defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

Although the dataset is class-balanced, some samples may still be difficult to classify due to visual ambiguity. In this study, difficult samples refer to images such as Parasitized cells with small or low-contrast parasite regions, Uninfected cells with staining artifacts resembling parasites, or cells affected by variations in illumination, focus, and staining quality. The modulating factor $(1 - p_t)^{\gamma}$ reduces the contribution of easily classified samples and allows the model to focus more on these visually ambiguous samples. The alpha and gamma values were fixed at $\alpha = 0.25$ and $\gamma = 2.0$ following the original Focal Loss formulation [33]. These parameters were not tuned separately, so the selected values should be interpreted as fixed experimental settings rather than optimized hyperparameters.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

The optimizer used in this study was AdamW [34] with a weight decay of 1×10^{-4}. AdamW applies weight decay separately from the adaptive gradient update, which improves regularization stability during fine-tuning. The initial learning rate was set to 3×10^{-4} for ResNet34, ResNet50, and EfficientNetV2-S, and 1×10^{-4} for VGG16 and VGG19 to maintain fine-tuning stability in the larger VGG backbones. A cosine annealing learning rate scheduler was used to gradually reduce the learning rate until the end of training.

Automatic Mixed Precision with float16 computation was enabled through torch.cuda.amp.autocast() to reduce GPU memory usage and accelerate matrix operations. PyTorch automatic gradient scaling was used during mixed-precision training to reduce the risk of numerical underflow. Gradient accumulation was applied specifically to VGG16 and VGG19 using an actual batch size of 16 with two accumulation steps, resulting in an effective batch size of 32. ResNet34, ResNet50, and EfficientNetV2-S were trained with a batch size of 32 without gradient accumulation.

Training was conducted for eight epochs per fold under the 5-fold stratified cross-validation scheme, without early stopping. This epoch setting was selected as a compromise between fine-tuning effectiveness and computational feasibility. Since all backbones were initialized with ImageNet-1K pretrained weights, training was intended to adapt pretrained visual representations to malaria cell images rather than learn features from scratch. The full experimental design consisted of five backbones, four augmentation scenarios, and five folds, resulting in 100 training runs. Therefore, eight epochs provided a consistent and manageable training budget for a controlled comparison across all configurations.

During training, the training loss, validation loss, training area under each curve (AUC), and validation AUC were recorded at each epoch. These values were used to construct learning curves for monitoring convergence behavior, training stability, and possible overfitting or underfitting. The same training protocol was applied to all augmentation scenarios and backbone architectures, allowing performance differences to be more directly associated with the evaluated model and augmentation configuration.

Evaluation metrics and statistical analysis

Model performance was evaluated using accuracy, precision, recall, F1-score, and the area under the ROC-AUC. For class-based metrics, the sigmoid output was converted into a binary prediction using a fixed threshold of 0.5. Predictions with probability ≥ 0.5 were assigned to the Parasitized class, while predictions with a probability of < 0.5 were assigned to the Uninfected class. This threshold was used for accuracy, precision, recall, F1-score, and confusion matrix calculation. The threshold of 0.5 was selected because the dataset was class-balanced and no threshold optimization was performed. In clinical deployment, however, the decision threshold should be adjusted according to the desired sensitivity-specificity trade-off. In contrast, ROC-AUC was computed across multiple thresholds and therefore did not depend on a single classification threshold.

In binary classification, predictions were categorized into True Positive, True Negative, False Positive, and False Negative. In these equations, TP, TN, FP, and FN denote the number of True Positive, True Negative, False Positive, and False Negative predictions, respectively. Accuracy measures the proportion of correct predictions among all evaluated samples:

$$\mathit{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \tag{3}$$

Precision measures the proportion of positive predictions that are truly positive:

$$\mathit{Precision} = TP/(TP + FP) \tag{4}$$

Recall, also known as sensitivity, measures the proportion of positive samples correctly identified by the model:

$$\mathit{Recall} = TP/(TP + FN) \tag{5}$$

F1-score provides the harmonic mean of precision and recall:

$$F1\text{-}\mathit{score} = 2 \times (\mathit{Precision} \times \mathit{Recall})/(\mathit{Precision} + \mathit{Recall}) \tag{6}$$

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

The ROC-AUC was constructed from the relationship between the True Positive Rate and the False Positive Rate:

$$TPR = TP/(TP + FN) \tag{7}$$

$$FPR = FP/(FP + TN) \tag{8}$$

The final performance of each model was reported as the mean and standard deviation across the five folds. The mean was calculated as:

$$\bar{x} = \frac{1}{K}\sum_{i=1}^{K} x_i \tag{9}$$

where $\bar{x}$ is the mean metric value, K is the number of folds, and xi is the metric value for the i-th fold. Since this study used 5-fold cross-validation, $K = 5$.

The standard deviation was calculated as:

$$SD = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(x_i - \bar{x})^2} \tag{10}$$

Inferential statistical analysis was conducted using ROC-AUC as the primary variable because ROC-AUC measures discriminative ability independently of a single classification threshold. The ROC-AUC values from each combination of backbone model, augmentation scenario, and fold were collected into a master table for statistical analysis.

The Friedman test was used as a non-parametric test to determine whether significant performance differences existed among multiple treatments evaluated on the same cross-validation folds. Two statistical pipelines were applied. Pipeline A compared multiple backbone models within each augmentation scenario, using folds as blocks and backbone models as treatments. Pipeline B compared multiple augmentation scenarios within each backbone model, using folds as blocks and augmentation scenarios as treatments.

The Friedman statistic is defined as:

$$\chi_F^2 = \frac{12}{Nk(k+1)}\sum_{j=1}^{k} R_j^2 - 3N(k+1) \tag{11}$$

where N is the number of blocks or folds, k is the number of treatments being compared, and R_j is the sum of ranks for the j-th treatment across all blocks.

Kendall's W was reported as an effect size measure for the Friedman test:

$$W = \frac{\chi_F^2}{N(k-1)} \tag{12}$$

Kendall's W ranges from 0 to 1, where values closer to 1 indicate stronger agreement in rankings across folds. The interpretation used in this study was negligible if $W < 0.10$, small if $0.10 \leq W < 0.30$, medium if $0.30 \leq W < 0.50$, and large if $W \geq 0.50$.

After the Friedman test, the Nemenyi post-hoc test was applied to identify pairwise differences between treatments. The Nemenyi test compares the difference in mean ranks between treatments using the Critical Difference:

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{13}$$

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

where q_{alpha} is the critical value based on the Studentized range statistic, k is the number of treatments being compared, and N is the number of folds. Two treatments were considered significantly different when the rank difference exceeded the Critical Difference and the Nemenyi p-value was below the significance level of $\alpha = 0.05$. As the Nemenyi test is rank-based and can have limited statistical power with a small number of blocks, the post-hoc results were interpreted together with mean ranks, effect sizes, and descriptive performance metrics rather than as the sole basis for model selection.

The Friedman test and Nemenyi post-hoc test were chosen based on the experimental design, which used five cross-validation folds. With only five folds, the Wilcoxon signed-rank test was not used as the main statistical test because the smallest attainable p-value in an exact two-sided test with $n = 5$ is 0.0625, which never reaches the conventional significance level of $\alpha = 0.05$. Normality testing was considered but not used as a decision criterion because tests of normality have limited statistical power with very small sample sizes. For this reason, a rank-based non-parametric approach was considered more suitable for this experimental setting.

Because rank-based tests do not convey the magnitude of the differences, effect sizes were also reported. For each comparison between EfficientNetV2-S and a baseline, and between each mixing scenario and the no-augmentation scenario, the mean paired difference across the five folds was reported with a 95% confidence interval. As the training sets of different folds overlap, the fold-wise differences are not independent, and a standard interval would be too narrow. The interval was therefore computed with the corrected resampled variance of Nadeau and Bengio [35], in which the variance of the fold-wise differences is multiplied by $(1/k + n_{test}/n_{train})$ instead of $1/k$, with $k = 5$ and $n_{test}/n_{train} = 0.25$, using a t distribution with four degrees of freedom. These intervals are descriptive and were not adjusted for multiplicity. To control the family-wise error rate across the nine Friedman tests (four in Pipeline A and five in Pipeline B), the Holm procedure [36] was applied, and the Nemenyi post-hoc results were treated as confirmatory only when the Holm-adjusted Friedman p-value was below 0.05. Within each family, the Nemenyi test already accounts for all pairwise comparisons. Each configuration was trained once with a single random seed, so variability due to weight initialization, data ordering, and augmentation sampling was not estimated.

Grad-CAM explainability analysis

Grad-CAM was used to provide post-hoc visual explanations of model predictions [19]. Grad-CAM generates class-discriminative localization maps by using the gradients of the target class score flowing into the final convolutional layer. In this study, Grad-CAM was applied to EfficientNetV2-S to examine whether the model focused on visually meaningful regions associated with malaria infection. The Grad-CAM outputs were used primarily for qualitative visual inspection of a small number of images selected by a fixed rule. No expert annotations or quantitative localization metrics were used, and faithfulness was examined only through the deletion-insertion check described below.

Representative samples were selected from four confusion matrix categories: True Positive, False Positive, False Negative, and True Negative. True Positive refers to Parasitized images correctly predicted as Parasitized, False Positive refers to Uninfected images incorrectly predicted as Parasitized, False Negative refers to Parasitized images incorrectly predicted as Uninfected, and True Negative refers to Uninfected images correctly predicted as Uninfected. Samples were taken from the fold-1 validation predictions of the EfficientNetV2-S model trained under the CutMix scenario. To avoid subjective selection, three images were chosen per category: the cases at the 25th, 50th, and 75th percentiles of predicted probability within that category. Each heatmap was computed for the class predicted by the model, so the maps for cases predicted as Uninfected show the regions supporting the Uninfected decision. No hematology expert annotation was used for selecting or validating the highlighted regions. The resulting heatmaps were overlaid on the original images using a jet colormap to visualize the regions that contributed most strongly to the model decisions. Since saliency-based explanations may appear visually plausible without necessarily proving causal reasoning, Grad-CAM was treated as supporting qualitative evidence rather than standalone clinical validation [22,25].

To complement the visual inspection, a perturbation-based faithfulness check was performed using deletion and insertion curves [37]. For 200 images randomly sampled from the fold-1 validation set, pixels were ranked by their Grad-CAM value for the predicted class. In the deletion test, the highest-ranked pixels were progressively replaced with the

ImageNet mean color (zero after normalization) in steps of 5%. In the insertion test, they were progressively revealed on a Gaussian-blurred copy of the image ($\sigma = 10$ pixels). The probability of the predicted class was recorded at each step, and the AUC was computed. A faithful map should give a lower deletion AUC and a higher insertion AUC than an uninformative map. As a reference, the same procedure was applied to random maps with the same spatial resolution as the Grad-CAM maps (8×8, bilinearly upsampled), and paired differences were assessed with the Wilcoxon signed-rank test.

Figure *2* summarizes the complete research framework, from dataset preparation and multi-backbone training to statistical, computational, and Grad-CAM-based evaluation.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

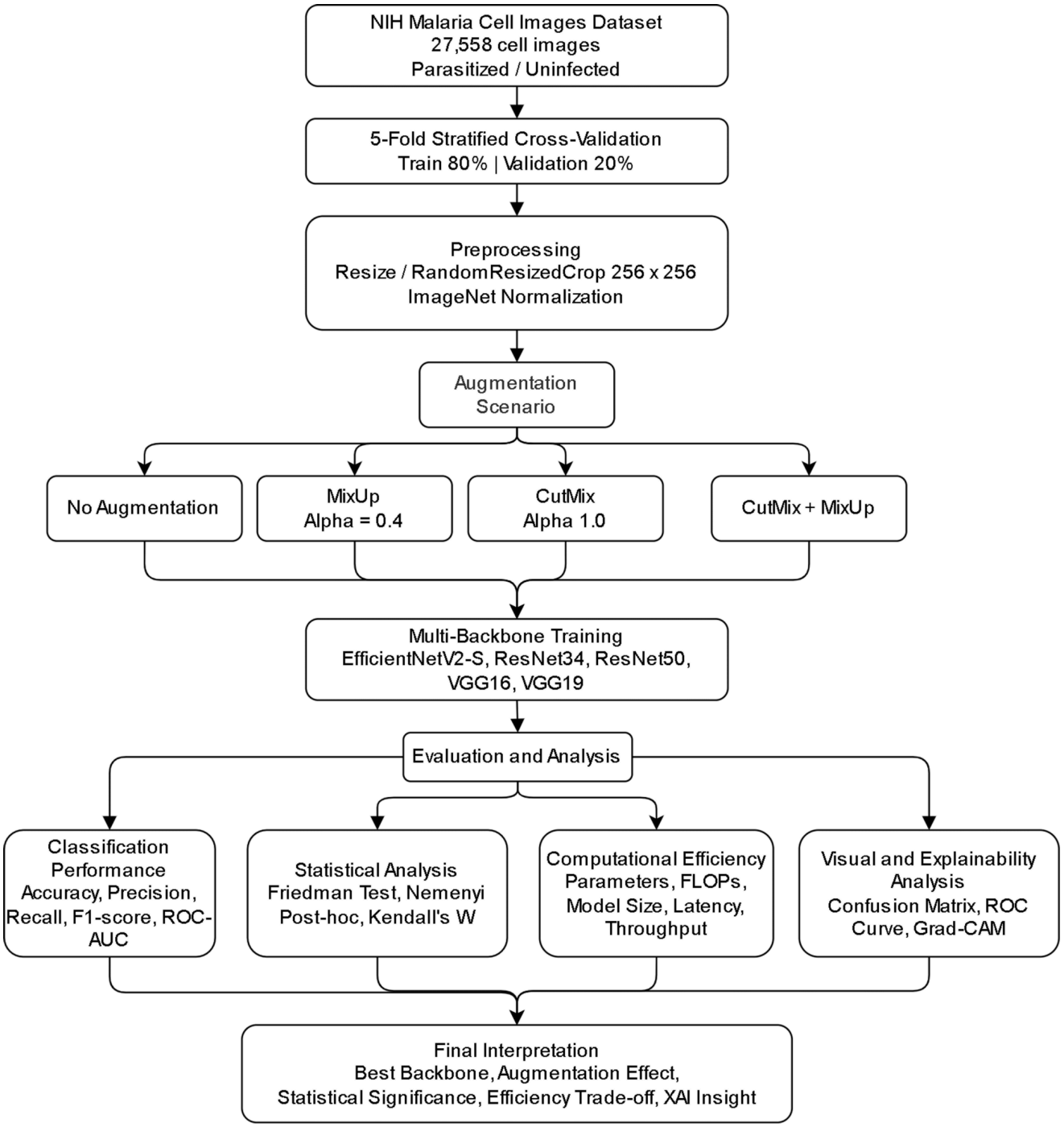


FIGURE 2: Overall research methodology framework, covering the data pipeline, multi-backbone training, and XAI-based evaluation.

Results And Discussion

Experimental setup

Model development and script preparation were carried out locally using Visual Studio Code, while training and evaluation were executed on the Kaggle notebook environment using dual NVIDIA Tesla T4 GPUs (T4 × 2) for all augmentation scenarios. Training was conducted on May 4, 2026, using an environment pinned to the March 20, 2026,

Kaggle Docker image to ensure library version consistency. The environment used Python 3.12.12 with PyTorch 2.10.0+cu128, torchvision 0.25.0+cu128, timm 1.0.25, albumentations 2.0.8, scikit-learn 1.6.1, OpenCV 4.13.0, statsmodels 0.14.6, NumPy 2.0.2, and pandas 2.3.3. The NIH Malaria Cell Images Dataset was accessed directly from the Kaggle dataset repository (iarunava/cell-images-for-detecting-malaria).

To manage the computational load of the full experimental design (five backbones × four augmentation scenarios × five folds = 100 training runs), the four augmentation scenarios were executed across separate Kaggle notebooks under two Kaggle accounts to enable parallel execution, using the same codebase except for the image transform module. The first account executed the CutMix-only and MixUp+CutMix scenarios, while the second account executed the MixUp-only and no-augmentation scenarios. All notebooks shared the same training, data loading, model, and loss implementation (crossval_train.py, dataset.py, engine.py, model_zoo.py, and losses.py). The MixUp, CutMix, and MixUp + CutMix scenarios used augmentations.py, which contains the basic augmentation pipeline, and were selected by passing --mixup, --cutmix, or both flags to run_all_models.py via run_all_experiments.sh. The no-augmentation scenario used no_augmentations.py in place of augmentations.py, which applies only resizing and normalization, and was run without mixing flags. Apart from these augmentation settings, the implementation and execution environment were identical across scenarios. This design ensured that performance differences across scenarios could be attributed to the augmentation strategy itself rather than to inconsistencies in implementation or execution environment.

Overall performance of EfficientNetV2-S

Table *3* presents the per-fold performance of EfficientNetV2-S under the CutMix augmentation scenario, which produced the best overall result among the evaluated configurations. Across the 5-fold stratified cross-validation, EfficientNetV2-S achieved accuracy values from 97.19% to 98.11%, ROC-AUC values from 99.67% to 99.83%, and F1-score values from 97.14% to 98.10%. The low standard deviation values across all evaluation metrics indicate that the model maintained stable generalization across different validation splits.

Fold	Model	Accuracy (%)	AUC (%)	F1-score (%)	Precision (%)	Recall (%)
Fold 1	EfficientNetV2-S	98.11	99.83	98.10	99.04	97.17
Fold 2	EfficientNetV2-S	97.46	99.67	97.42	98.99	95.90
Fold 3	EfficientNetV2-S	97.57	99.69	97.54	98.66	96.44
Fold 4	EfficientNetV2-S	97.48	99.68	97.46	98.23	96.70
Fold 5	EfficientNetV2-S	97.19	99.71	97.14	98.91	95.43
Mean ± SD	EfficientNetV2-S	97.56 ± 0.30	99.72 ± 0.06	97.53 ± 0.31	98.77 ± 0.30	96.33 ± 0.61

TABLE 3: Per-fold performance of EfficientNetV2-S with CutMix augmentation on the NIH Malaria Dataset.

Note. The values in the final row represent the mean ± standard deviation across the 5-fold cross-validation. Standard deviations across folds or runs were computed as the population standard deviation (ddof = 0) in all tables; the confidence intervals in Table 8 and the standard deviations across images in Table 14 use the sample standard deviation.

EfficientNetV2-S achieved a mean accuracy of 97.56 ± 0.30%, ROC-AUC of 99.72 ± 0.06%, F1-score of 97.53 ± 0.31%, precision of 98.77 ± 0.30%, and recall of 96.33 ± 0.61%. The relatively low standard deviation values indicate that the model was not strongly affected by a particular fold partition. In malaria screening, recall is especially important because

false negative predictions may cause infected samples to be missed. However, direct comparison with recall values reported in previous studies should be made cautiously because validation protocols, preprocessing, augmentation settings, and patient-level independence may differ across studies.

Backbone comparison on Fold 1 under the CutMix scenario

Table *4* presents the comparison of all evaluated CNN backbones on Fold 1 under the CutMix augmentation scenario. The evaluated metrics were accuracy, ROC-AUC, F1-score, precision, and recall. EfficientNetV2-S achieved the highest accuracy, ROC-AUC, F1-score, and recall in this fold, whereas ResNet50 had a slightly higher precision (99.10% vs. 99.04%). The accuracy margin between EfficientNetV2-S and the other backbones ranged from 0.72 to 1.41 percentage points, showing a clear numerical advantage in the best-performing fold.

Model	Accuracy (%)	AUC (%)	F1-score (%)	Precision (%)	Recall (%)
EfficientNetV2-S	98.11	99.83	98.10	99.04	97.17
ResNet50	97.39	99.70	97.34	99.10	95.65
VGG19	97.35	99.64	97.31	98.77	95.90
VGG16	96.97	99.64	96.91	98.98	94.92
ResNet34	96.70	99.67	96.62	99.01	94.34

TABLE 4: Performance comparison of all CNN backbones on Fold 1 under the CutMix augmentation scenario.

Note. The best value for each metric is shown in bold in the final manuscript.

Across the 5-fold cross-validation, EfficientNetV2-S with CutMix achieved a mean accuracy of 97.56 ± 0.30%, outperforming VGG19 at 96.94 ± 0.36%, VGG16 at 96.88 ± 0.30%, ResNet50 at 96.85 ± 0.43%, and ResNet34 at 96.40 ± 0.33%. This pattern shows that EfficientNetV2-S consistently produced the best numerical performance among the evaluated backbones under the CutMix setting.

The strong performance of EfficientNetV2-S can be associated with its architectural design. EfficientNetV2-S combines compound scaling with Fused-MBConv blocks, allowing the model to balance representational capacity and computational efficiency. Fused-MBConv blocks combine the expansion and spatial convolution operations more directly than the MBConv design used in earlier EfficientNet variants [38], which can improve training efficiency and local feature extraction while reducing unnecessary computational overhead in early layers [32]. This characteristic is relevant for malaria cell classification because the discriminative features between Parasitized and Uninfected cells often appear as small local structures, such as dark-stained spots, cytoplasmic texture changes, or abnormal staining patterns inside red blood cells.

ROC-AUC-based statistical significance

Statistical significance analysis was conducted using the Friedman test to determine whether the performance differences among CNN backbones were consistent across folds. The analysis used ROC-AUC values obtained from the five cross-validation folds because ROC-AUC reflects the model ability to distinguish between Parasitized and Uninfected classes without depending on a single classification threshold.

The Friedman test was used as the non-parametric statistical test, followed by the Nemenyi post-hoc test to identify pairs of models or augmentation scenarios with significant differences. Two statistical pipelines were applied. Pipeline A compared the five CNN backbones (EfficientNetV2-S, ResNet34, ResNet50, VGG16, and VGG19) within each augmentation scenario using five folds as blocks. Pipeline B compared the four augmentation scenarios (without augmentation, MixUp, CutMix, and MixUp + CutMix) within each backbone, also using five folds as blocks.

Before statistical testing, the ROC-AUC values of each model were summarized as the mean and standard deviation across the five folds. This summary provides an initial overview of the discriminative performance of each backbone under each augmentation scenario.

Scenario	Model	ROC-AUC ± SD
CutMix	EfficientNetV2-S	0.9972 ± 0.0006
CutMix	ResNet50	0.9959 ± 0.0006
CutMix	VGG16	0.9959 ± 0.0004
CutMix	VGG19	0.9955 ± 0.0005
CutMix	ResNet34	0.9954 ± 0.0008
MixUp	EfficientNetV2-S	0.9970 ± 0.0006
MixUp	VGG16	0.9961 ± 0.0005
MixUp	VGG19	0.9960 ± 0.0006
MixUp	ResNet50	0.9958 ± 0.0006
MixUp	ResNet34	0.9954 ± 0.0008
CutMix & MixUp	EfficientNetV2-S	0.9971 ± 0.0005
CutMix & MixUp	VGG16	0.9960 ± 0.0004
CutMix & MixUp	VGG19	0.9957 ± 0.0007
CutMix & MixUp	ResNet50	0.9956 ± 0.0008
CutMix & MixUp	ResNet34	0.9952 ± 0.0007
Without Augmentation	EfficientNetV2-S	0.9958 ± 0.0009
Without Augmentation	VGG16	0.9950 ± 0.0008
Without Augmentation	ResNet50	0.9948 ± 0.0011
Without Augmentation	VGG19	0.9945 ± 0.0010

Without Augmentation	ResNet34	0.9943 ± 0.0010

TABLE 5: Summary of ROC-AUC values of all CNN backbones under each augmentation scenario.

Note. No augmentation: training images were only resized to 256 × 256 pixels and normalized. MixUp, CutMix, and MixUp + CutMix: basic augmentation (RandomResizedCrop, horizontal/vertical flipping, affine transformation, and ColorJitter) combined with the indicated batch-level mixing strategy.

Based on Table 5, EfficientNetV2-S achieved the highest mean ROC-AUC value across all evaluation scenarios. Under the CutMix scenario, EfficientNetV2-S achieved a ROC-AUC of 0.9972 ± 0.0007, followed by ResNet50 and VGG16 at 0.9959, VGG19 at 0.9955, and ResNet34 at 0.9954. A similar pattern was observed under MixUp, CutMix & MixUp, and the without-augmentation scenario. These findings show that EfficientNetV2-S had the most consistent discriminative performance among the evaluated CNN backbones.

To determine whether these ROC-AUC differences were statistically significant, the Friedman test was conducted for each augmentation scenario. In Pipeline A, the cross-validation folds were treated as blocks, while the CNN backbones were treated as the compared treatments.

Scenario	χ^2	df	*p*-value	Holm-adjusted *p*	Kendall's *W*	Interpretation	Conclusion
CutMix	14.0800	4	0.0070	0.0423	0.7040	Large	Significant
MixUp	17.1200	4	0.0018	0.0165	0.8560	Large	Significant
CutMix & MixUp	14.8800	4	0.0050	0.0370	0.7440	Large	Significant
Without Augmentation	15.0400	4	0.0046	0.0370	0.7520	Large	Significant

TABLE 6: Friedman test results among CNN backbones based on ROC-AUC.

Note. Holm-adjusted *p*-values were computed across all nine Friedman tests (Tables 6 and 9). Augmentation scenarios are defined in the note to Table 5.

Based on Table 6, the Friedman test results showed significant global differences among the CNN backbones across all evaluation scenarios. Under CutMix, the test produced $\chi^2 = 14.08$, df = 4, $p = 0.0070$, and Kendall's $W = 0.704$. Under MixUp, the test produced $\chi^2 = 17.12$, df = 4, $p = 0.0018$, and Kendall's $W = 0.856$. Under CutMix & MixUp, the test produced $\chi^2 = 14.88$, df = 4, $p = 0.0050$, and Kendall's $W = 0.744$. Under the without-augmentation scenario, the test produced $\chi^2 = 15.04$, df = 4, $p = 0.0046$, and Kendall's $W = 0.752$. All Kendall's W values fell in the large effect size category ($W \geq 0.50$), indicating strong and consistent ranking differences among the backbones across the five folds. After Holm correction across the nine Friedman tests, all four tests remained significant (adjusted $p = 0.0423$ for CutMix, 0.0165 for MixUp, 0.0370 for CutMix & MixUp, and 0.0370 for the without-augmentation scenario).

Since the Friedman test indicated significant global differences, the analysis continued with the Nemenyi post-hoc test. This test was used to identify which backbone pairs had significantly different rankings.

Augmentation Scenario	Model A	Model B	Mean Rank A	Mean Rank B	Rank Difference	CD	*p*-value	Conclusion
CutMix	EfficientNetV2-S	ResNet34	5.0000	1.6000	3.4000	2.7280	0.0061	Significant
CutMix	EfficientNetV2-S	VGG19	5.0000	2.0000	3.0000	2.7280	0.0227	Significant
MixUp	EfficientNetV2-S	ResNet34	5.0000	1.0000	4.0000	2.7280	0.0006	Significant
CutMix & MixUp	EfficientNetV2-S	ResNet34	5.0000	1.4000	3.6000	2.7280	0.0029	Significant
Without Augmentation	EfficientNetV2-S	ResNet34	4.8000	1.2000	3.6000	2.7280	0.0029	Significant

TABLE 7: CNN backbone pairs with significant differences based on the Nemenyi post-hoc test.

Note. Augmentation scenarios are defined in the note to Table 5.

Based on Table *7*, EfficientNetV2-S showed a significant difference compared with ResNet34 across all scenarios: CutMix (rank difference = 3.40, CD = 2.728, $p = 0.0061$), MixUp (rank difference = 4.00, CD = 2.728, $p = 0.0006$), CutMix & MixUp (rank difference = 3.60, CD = 2.728, $p = 0.0029$), and without augmentation (rank difference = 3.60, CD = 2.728, $p = 0.0029$). Under CutMix, EfficientNetV2-S also showed a significant difference compared with VGG19 (rank difference = 3.00, CD = 2.728, $p = 0.0227$). These results indicate that EfficientNetV2-S ranked first in nearly all folds, although a rank-based test does not show how large the differences were. Because only five non-independent folds were available, these pairwise results should be regarded as suggestive rather than definitive.

However, the Nemenyi post-hoc results also showed that EfficientNetV2-S was not significantly different from all comparison backbones. In several comparisons, such as EfficientNetV2-S versus ResNet50 and EfficientNetV2-S versus VGG16, the rank differences did not exceed the Critical Difference value, and the Nemenyi *p*-values were not below 0.05. Therefore, the most appropriate statistical claim is that EfficientNetV2-S achieved the highest ROC-AUC performance and the most consistent ranking, but it cannot be claimed to significantly outperform all baselines in every pairwise comparison. This interpretation also acknowledges the limited statistical power of post-hoc rank tests when only five folds are used.

Table *8* reports the magnitude of these differences. The mean ROC-AUC of EfficientNetV2-S exceeded that of the baselines by 0.0008 to 0.0019 across the 16 comparisons, and the corrected 95% confidence intervals excluded zero in 14 of them; the exceptions were VGG16 and VGG19 in the no-augmentation scenario. In absolute terms, however, all differences were below 0.002, and all five backbones reached a ROC-AUC above 0.994, so the practical relevance of the difference in discriminative ability is limited. At the fixed threshold of 0.5, the differences were more visible in the mixing scenarios. Under CutMix, EfficientNetV2-S had higher accuracy by 0.62 to 1.16 percentage points and higher recall by 1.31 to 2.48 percentage points, which corresponds to roughly 13 to 25 fewer missed parasitized cells per 1,000 parasitized

cells. In the no-augmentation scenario, the accuracy and recall differences were small, and all their confidence intervals included zero. The advantage of EfficientNetV2-S is therefore consistent in direction but small in magnitude, and whether it is clinically relevant cannot be determined from cell-level results on a single dataset.

Scenario	Baseline	ΔROC-AUC [95% CI]	ΔAccuracy, pp [95% CI]	ΔRecall, pp [95% CI]
CutMix	ResNet34	0.0018 [0.0005, 0.0030]	1.16 [0.70, 1.62]	2.48 [1.77, 3.20]
CutMix	ResNet50	0.0013 [0.0007, 0.0019]	0.71 [0.26, 1.17]	1.65 [0.57, 2.72]
CutMix	VGG16	0.0013 [0.0003, 0.0023]	0.69 [0.06, 1.31]	1.42 [0.11, 2.74]
CutMix	VGG19	0.0017 [0.0008, 0.0025]	0.62 [0.16, 1.09]	1.31 [-0.11, 2.74]
MixUp	ResNet34	0.0016 [0.0011, 0.0020]	1.05 [0.55, 1.54]	2.33 [1.66, 3.00]
MixUp	ResNet50	0.0012 [0.0011, 0.0013]	0.61 [0.02, 1.19]	1.24 [-0.02, 2.50]
MixUp	VGG16	0.0009 [0.0002, 0.0016]	0.48 [-0.11, 1.07]	0.81 [-0.59, 2.20]
MixUp	VGG19	0.0011 [0.0007, 0.0014]	0.48 [0.15, 0.82]	0.95 [-0.09, 1.99]
CutMix & MixUp	ResNet34	0.0019 [0.0010, 0.0028]	1.12 [0.52, 1.73]	2.50 [1.22, 3.77]
CutMix & MixUp	ResNet50	0.0015 [0.0003, 0.0026]	0.65 [-0.21, 1.51]	1.41 [-0.64, 3.45]
CutMix & MixUp	VGG16	0.0011 [0.0005, 0.0017]	0.36 [-0.23, 0.94]	0.46 [-1.04, 1.97]
CutMix & MixUp	VGG19	0.0013 [0.0002, 0.0025]	0.41 [-0.32, 1.13]	0.68 [-0.94, 2.31]
Without Augmentation	ResNet34	0.0015 [0.0007, 0.0022]	0.21 [-0.54, 0.97]	0.15 [-1.84, 2.13]
Without Augmentation	ResNet50	0.0009 [0.0001, 0.0018]	0.44 [-1.03, 1.91]	0.99 [-2.69, 4.67]
Without Augmentation	VGG16	0.0008 [-0.0005, 0.0020]	0.14 [-0.48, 0.75]	0.35 [-1.45, 2.15]
Without Augmentation	VGG19	0.0012 [-0.0004, 0.0029]	0.26 [-0.10, 0.63]	0.44 [-1.55, 2.44]

TABLE 8: Mean paired differences between EfficientNetV2-S and each baseline across the five folds, with corrected 95% confidence intervals.

Note. Positive values favor EfficientNetV2-S. pp, percentage points. Confidence intervals use the corrected resampled variance of Nadeau and Bengio [35] and were not adjusted for multiplicity. Augmentation scenarios are defined in the note to Table 5.

Effect of augmentation strategies

In addition to comparing the backbones, this study evaluated whether the augmentation scenarios had a significant effect on ROC-AUC for each model. The evaluated scenarios were without augmentation, MixUp, CutMix, and MixUp + CutMix. In Pipeline B, the cross-validation folds were treated as blocks, while the augmentation scenarios were treated as the compared treatments.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

Model	χ^2	df	*p*-value	Holm-adjusted *p*	Kendall's *W*	Interpretation	Conclusion
EfficientNetV2-S	9.7200	3	0.0211	0.0844	0.6480	Large	Not significant after correction
ResNet34	9.7200	3	0.0211	0.0844	0.6480	Large	Not significant after correction
ResNet50	5.8800	3	0.1176	0.1176	0.3920	Medium	Not significant
VGG16	9.7200	3	0.0211	0.0844	0.6480	Large	Not significant after correction
VGG19	10.6800	3	0.0136	0.0679	0.7120	Large	Not significant after correction

TABLE 9: Friedman test results among augmentation scenarios for each CNN backbone.

Note. Holm-adjusted *p*-values were computed across all nine Friedman tests (Tables 6 and 9). Augmentation scenarios are defined in the note to Table 5.

Based on Table 9, the unadjusted Friedman tests indicated an effect of the augmentation scenario on ROC-AUC for EfficientNetV2-S, ResNet34, VGG16, and VGG19 ($p = 0.0136$ to 0.0211, with large Kendall's *W*), but not for ResNet50 ($p = 0.1176$). After Holm correction across the nine Friedman tests, none of the five tests remained significant (adjusted $p = 0.0844$ for EfficientNetV2-S, ResNet34, and VGG16, 0.1176 for ResNet50, and 0.0679 for VGG19). The effect of augmentation on ROC-AUC is therefore suggestive rather than statistically confirmed, and the post-hoc comparisons below are reported as exploratory.

To identify which augmentation scenario pairs differed significantly, the Nemenyi post-hoc test was conducted in Pipeline B.

Model	Scenario A	Scenario B	Mean Rank A	Mean Rank B	Rank Difference	CD	*p*-value	Conclusion
EfficientNetV2-S	CutMix	Without Augmentation	3.4000	1.0000	2.4000	2.0976	0.0173	Significant

ResNet34	CutMix	Without Augmentation	3.2000	1.0000	2.2000	2.0976	0.0355	Significant
ResNet34	MixUp	Without Augmentation	3.2000	1.0000	2.2000	2.0976	0.0355	Significant
VGG16	MixUp	Without Augmentation	3.6000	1.2000	2.4000	2.0976	0.0173	Significant
VGG19	MixUp	Without Augmentation	3.6000	1.0000	2.6000	2.0976	0.0079	Significant

TABLE 10: Augmentation scenario pairs with significant differences based on the Nemenyi post-hoc test.

Note. Augmentation scenarios are defined in the note to Table 5.

In this exploratory analysis (Table *10*), compared with the no-augmentation scenario, the CutMix scenario (basic augmentation combined with CutMix) yielded significantly higher ROC-AUC ranks for EfficientNetV2-S (rank difference = 2.40, CD = 2.098, $p = 0.0173$) and ResNet34 (rank difference = 2.20, CD = 2.098, $p = 0.0355$). Similarly, the MixUp scenario (basic augmentation combined with MixUp) yielded significantly higher ROC-AUC ranks than the no-augmentation scenario for ResNet34 (rank difference = 2.20, CD = 2.098, $p = 0.0355$), VGG16 (rank difference = 2.40, CD = 2.098, $p = 0.0173$), and VGG19 (rank difference = 2.60, CD = 2.098, $p = 0.0079$). These findings indicate that the effect of augmentation was model-dependent, meaning that the most beneficial augmentation strategy was not the same for every backbone. The corresponding mean ROC-AUC gains over the no-augmentation scenario were small, ranging from 0.0008 to 0.0014 across backbones and mixing scenarios.

A focused ablation study was also conducted on EfficientNetV2-S to isolate the contribution of each augmentation strategy to the final model performance.

Augmentation Scenario	Accuracy (%) ± SD	AUC (%)	F1-score (%)	Precision (%)	Recall (%)
Without Augmentation	97.11 ± 0.19	99.58 ± 0.09	97.08 ± 0.19	98.11 ± 0.62	96.08 ± 0.66
MixUp	97.49 ± 0.28	99.70 ± 0.06	97.45 ± 0.30	98.85 ± 0.33	96.09 ± 0.71
CutMix	97.56 ± 0.30	99.72 ± 0.06	97.53 ± 0.31	98.77 ± 0.30	96.33 ± 0.61
MixUp + CutMix	97.37 ± 0.27	99.71 ± 0.05	97.33 ± 0.27	98.86 ± 0.18	95.85 ± 0.50

TABLE 11: Data augmentation ablation study on EfficientNetV2-S, reported as mean ± SD across the 5-fold cross-validation.

Note. Augmentation scenarios are defined in the note to Table 5.

Based on Table *11*, the ablation results reveal several important findings. First, every augmentation strategy improved performance compared with the baseline without augmentation. As the mixing scenarios also included basic augmentation, these gains represent the combined effect of basic and mixing-based augmentation relative to the resize-only baseline. Among the three mixing scenarios, which share the same basic augmentation, CutMix achieved the highest mean accuracy. CutMix alone produced the largest improvement, increasing accuracy from 97.11 ± 0.19% to 97.56 ± 0.30%, corresponding to a gain of 0.45 percentage points. Second, MixUp alone also improved accuracy from 97.11 ± 0.19% to 97.49 ± 0.28% and AUC from 99.58 ± 0.09% to 99.70 ± 0.06% compared with the baseline, although the improvement in recall was marginal. Third, although MixUp + CutMix did not produce the highest numerical accuracy, it achieved the lowest AUC standard deviation among the evaluated scenarios, suggesting more stable discriminative performance across folds.

For EfficientNetV2-S, CutMix produced the best numerical performance, with a mean accuracy of 97.56 ± 0.30%, ROC-AUC of 99.72 ± 0.06%, and F1-score of 97.53 ± 0.31%. A possible explanation is that CutMix encourages the model to rely on local image regions, which is relevant for malaria cell classification because parasite-related structures often occupy only a small region within the cell image. However, this mechanism was not tested directly in this study, and the ROC-AUC differences among the three mixing scenarios were small (0.0001 to 0.0002). MixUp also improved performance compared with no augmentation, although the improvement was smaller than that achieved by CutMix. Meanwhile, the MixUp + CutMix combination remained competitive but was not the best scenario numerically. Therefore, combining augmentation methods does not always produce the highest performance, and augmentation strategy selection should consider both dataset characteristics and model architecture.

Computational efficiency

Computational efficiency evaluation was conducted to analyze the trade-off between model complexity and inference performance. All measurements were obtained using the exact model architecture employed during training, namely each backbone combined with the unified binary classification head (Dropout, a 512-unit fully connected layer, GELU activation, layer normalization, Dropout, and a final one-unit linear output), rather than the bare backbone alone. The evaluation was performed on hardware consisting of an Intel Core i5-11400H processor at 2.70 GHz, 16 GB RAM, and an NVIDIA GeForce RTX 3050 Laptop GPU with 4 GB VRAM. All measurements used the same 256 × 256 input size as the classification experiment, with a batch size of 1. Runtime measurements excluded disk I/O and were interpreted as hardware-specific estimates, since software libraries, CUDA/cuDNN kernels, warm-up procedures, and batching strategy can affect latency and throughput.

Model efficiency was assessed from two perspectives: static architectural metrics and practical inference metrics. Static architectural metrics included the number of parameters, MACs, and model size, all computed once per backbone since these values remain identical across folds and augmentation scenarios. Computational cost was measured with the thop profiling library on a single dummy input tensor of size 1 × 3 × 256 × 256. Because thop counts MACs rather than floating-point operations, the values are reported as MACs; one MAC corresponds to approximately two floating-point operations (FLOPs), while model size was estimated from the parameter count under the assumption of 32-bit floating-point storage. Practical inference metrics, namely latency and throughput, were measured separately on CPU and GPU using the trained checkpoint for each combination of backbone, augmentation scenario, and cross-validation fold. Each measurement used 20 warm-up iterations followed by 100 timed iterations; for GPU measurements, torch.cuda.synchronize() was called immediately before and after the timed loop to ensure accurate wall-clock timing of asynchronous CUDA operations. Throughput was derived as the reciprocal of the mean per-image latency. The architectural complexity summary of each model is presented in Table *12*.

Model	Parameters (Million)	MACs (Giga)	Model Size (MB)
EfficientNetV2-S	20.84	3.54	79.49
ResNet34	21.55	4.80	82.21
ResNet50	24.56	5.40	93.70
VGG16	134.53	20.04	513.18
VGG19	139.84	25.48	533.43

TABLE 12: Model architectural complexity of each CNN backbone, including the unified binary classification head.

Note. Parameters, multiply-accumulate operations (MACs), and model size are static architectural metrics computed once per backbone; they do not vary across folds or augmentation scenarios. MACs were measured with the thop profiling library on a single input tensor of size $1 \times 3 \times 256 \times 256$. Model size was estimated from the parameter count assuming 32-bit floating-point storage.

Based on Table *12*, EfficientNetV2-S remained the most lightweight model in terms of static architectural metrics, with the smallest number of parameters, the lowest MACs, and the smallest model size among the evaluated backbones. In contrast, the VGG family exhibited a substantially higher computational burden, with VGG19 requiring 25.48 GMACs and a model file size exceeding 500 MB.

To measure real-world inference performance, latency and throughput were evaluated across all four augmentation scenarios and five cross-validation folds for each backbone, since augmentation strategy does not alter the model architecture during inference. The values reported in Table *12* therefore represent the mean and standard deviation across all 20 measurements (four scenarios × five folds) per model, providing a more stable and representative estimate of inference performance than a single-run measurement.

Model	CPU Latency (ms)	CPU Throughput (img/s)	GPU Latency (ms)	GPU Throughput (img/s)
EfficientNetV2-S	63.16 ± 6.97	16.03 ± 1.82	23.87 ± 1.38	42.04 ± 2.48
ResNet34	43.49 ± 4.97	23.30 ± 2.72	6.12 ± 0.72	165.73 ± 19.76
ResNet50	56.24 ± 7.71	18.11 ± 2.45	8.08 ± 0.52	124.21 ± 7.95
VGG16	127.02 ± 11.25	7.94 ± 0.71	8.83 ± 0.03	113.22 ± 0.34
VGG19	158.39 ± 15.03	6.37 ± 0.64	10.80 ± 0.04	92.59 ± 0.32

TABLE 13: Average CPU and GPU inference performance (mean ± SD across all augmentation scenarios and folds).

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

Based on Table *13*, there was a clear discrepancy between theoretical computational efficiency and practical runtime performance. Although EfficientNetV2-S had the lowest theoretical MACs at 3.54 GMACs, ResNet34 achieved substantially faster inference on the tested hardware, with a mean GPU latency of 6.12 ± 0.74 ms and a mean throughput of 165.73 ± 20.28 images per second, compared with 23.87 ± 1.42 ms and 42.04 ± 2.55 images per second for EfficientNetV2-S. Notably, EfficientNetV2-S was even slower on GPU than both VGG16 (8.83 ± 0.03 ms) and VGG19 (10.80 ± 0.04 ms), despite having far fewer parameters and MACs than either VGG variant.

This result can be explained by the fact that standard convolutional operations, as used in ResNet and VGG architectures, are highly optimized on modern GPU hardware through mature cuDNN kernels. In contrast, EfficientNetV2-S relies on Fused-MBConv blocks and depthwise-separable-like operations whose theoretical operation count is low but whose practical runtime can be affected by memory access patterns, kernel launch overhead, and the relatively limited optimization of such operations on consumer-grade GPUs compared to standard convolutions. Therefore, theoretical efficiency, as reflected by parameter count and MACs, does not always translate directly into the fastest measured runtime performance.

Overall, GPU acceleration provided a substantial speed-up for all models compared with CPU-only computation. For systems prioritizing real-time inference on GPU hardware, ResNet34 demonstrated the best runtime efficiency, with the lowest GPU latency and the highest GPU throughput among all evaluated backbones. However, EfficientNetV2-S retained the strongest architectural efficiency, with the fewest parameters (20.84 million), the lowest MACs (3.54 GMACs), and the smallest model size (79.49 MB) among the evaluated backbones. This trade-off indicates that EfficientNetV2-S is best suited for deployment scenarios prioritizing storage footprint and architectural compactness, such as edge devices with limited memory, whereas ResNet34 may be preferable when GPU inference speed is the primary deployment requirement.

Confusion matrix and learning curve analysis

Figure *3* shows the confusion matrices of all CNN backbones on the Fold 1 validation subset of the NIH Malaria Dataset under the CutMix augmentation scenario. EfficientNetV2-S recorded the lowest number of false negatives among the evaluated backbones, with 78 false negatives. The false negative counts for ResNet34, ResNet50, VGG16, and VGG19 were 156, 120, 140, and 113, respectively. This characteristic is important in malaria screening because false negative predictions may cause infected samples to be missed, potentially delaying treatment.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

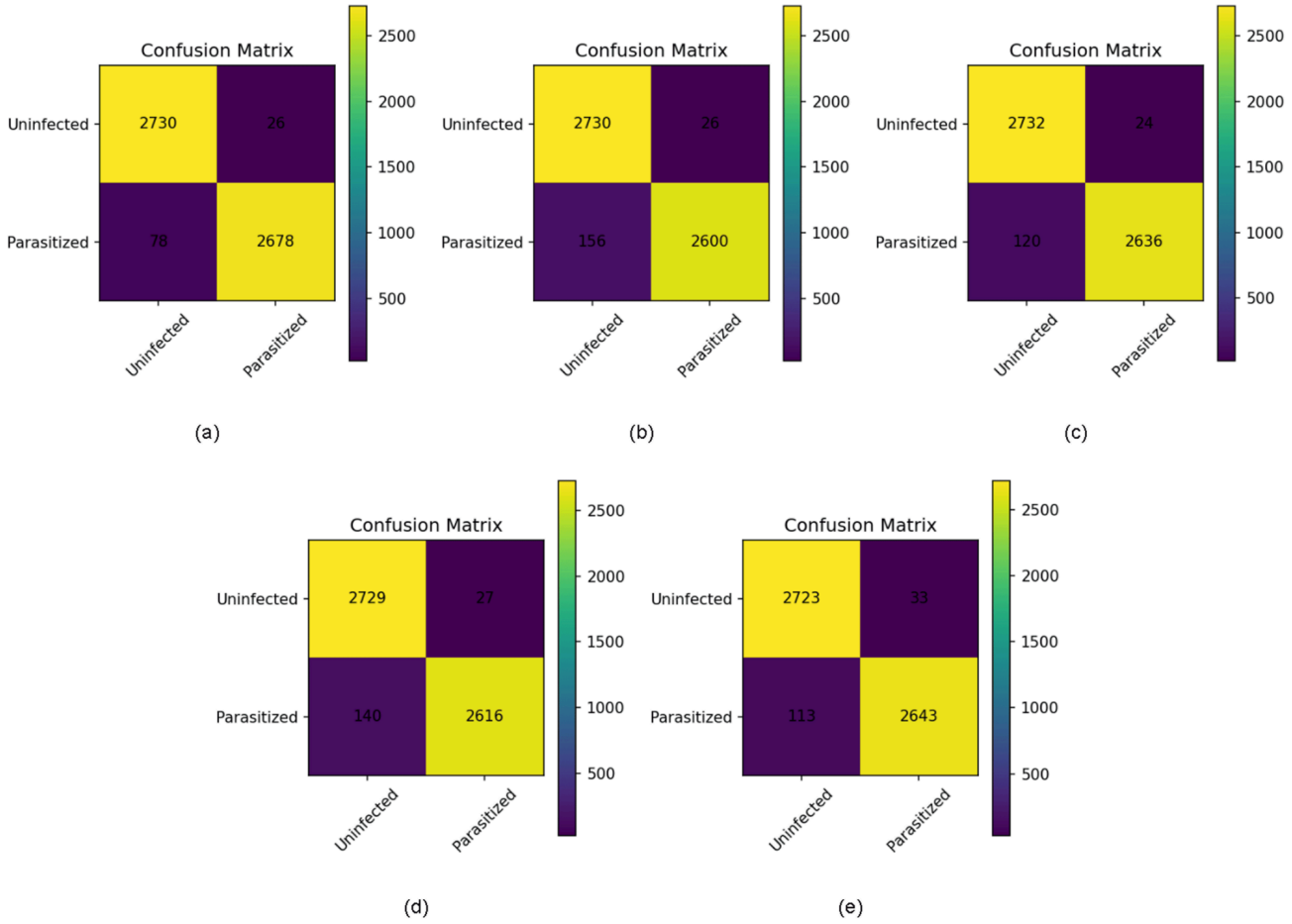


FIGURE 3: Confusion matrices of all CNN backbones on Fold 1 under the CutMix augmentation scenario: (a) EfficientNetV2-S, (b) ResNet34, (c) ResNet50, (d) VGG16, and (e) VGG19.

The confusion matrix results support the quantitative findings from the per-fold evaluation and backbone comparison. EfficientNetV2-S not only achieved the highest overall performance metrics but also showed a lower tendency to miss positive malaria samples in the analyzed fold. In Fold 1, EfficientNetV2-S produced 2678 true positives, 2730 true negatives, 26 false positives, and 78 false negatives. In diagnostic support systems, reducing false negatives is particularly relevant because the clinical risk of missing infected samples is generally higher than the risk of flagging additional samples for further review.

The lower number of false negatives was not limited to Fold 1. Under the CutMix scenario, EfficientNetV2-S produced the fewest false negatives among the five backbones in every fold (78, 113, 98, 91, and 126 in Folds 1 to 5). Summed over the five validation folds, which together contain all 13,779 Parasitized images, EfficientNetV2-S missed 506 infected cells (3.67%), compared with 687 for VGG19 (4.99%), 702 for VGG16 (5.09%), 733 for ResNet50 (5.32%), and 848 for ResNet34 (6.15%). This corresponds to 26% to 40% fewer missed infected cells than the baseline backbones. The reduction in false negatives came with a slightly higher number of false positives (166 for EfficientNetV2-S versus 136 to 159 for the baselines across 13,779 Uninfected images), a trade-off that is generally acceptable in screening, where flagged cells are subsequently reviewed by a microscopist. The corresponding recall differences and their confidence intervals are reported in Table *8*.

Figure *4* presents the learning curves of all CNN backbones on Fold 1 under the CutMix augmentation scenario. The curves show the training and validation performance across eight epochs for EfficientNetV2-S, ResNet34, ResNet50, VGG16, and VGG19. These learning curves were used to observe the convergence behavior of each model and to

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

examine whether the training process showed indications of overfitting or underfitting. Overall, the evaluated models showed stable training patterns, with validation performance improving along with the training process. EfficientNetV2-S reached the highest validation AUC (0.9983), while the other backbones reached 0.9964 to 0.9970, and the validation AUC of all backbones plateaued within the eight epochs.

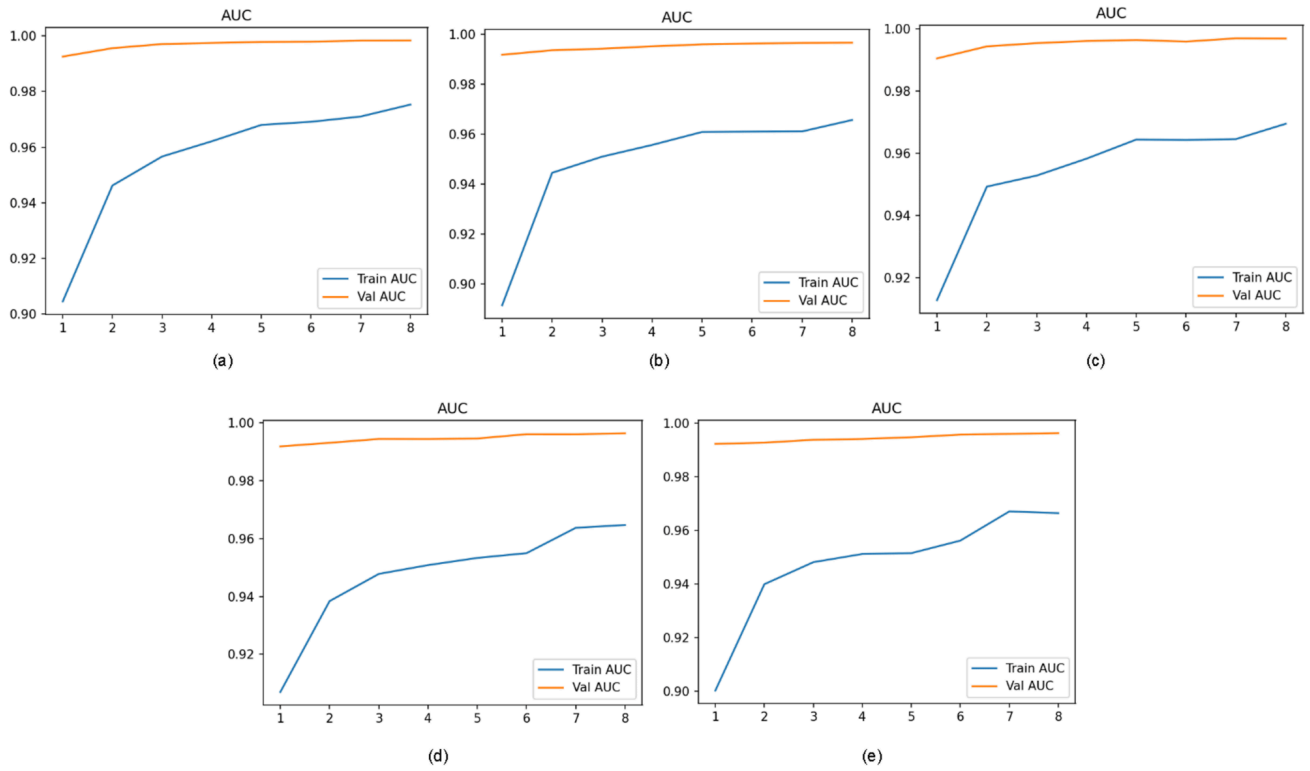


FIGURE 4: Learning curves of all CNN backbones on Fold 1 under the CutMix augmentation scenario: (a) EfficientNetV2-S (validation AUC = 0.9983), (b) ResNet34 (0.9967), (c) ResNet50 (0.9970), (d) VGG16 (0.9964), and (e) VGG19 (0.9964).

The reported values correspond to the best epoch and match Table 4. Training AUC is lower than validation AUC because training metrics were computed on augmented and mixed training batches.

The learning curve analysis provides additional evidence of the training stability of EfficientNetV2-S under the CutMix scenario. The validation performance remained stable relative to the training trend across the eight epochs, suggesting that the model converged effectively without clear signs of severe overfitting in the analyzed fold. This finding complements the cross-validation metrics and statistical analysis, which showed that EfficientNetV2-S achieved the highest overall performance and the most consistent ranking among the evaluated backbones.

Grad-CAM explainability analysis

Grad-CAM visualization was used to evaluate the image regions that contributed most strongly to the decisions made by EfficientNetV2-S. Figure *5* presents Grad-CAM visualizations for three cases in each of the four categories: True Positive, False Positive, False Negative, and True Negative. Red to yellow areas indicate regions with higher activation contribution to the model decision, while blue areas indicate regions with lower activation contribution. These visualizations should be interpreted cautiously because Grad-CAM can highlight spurious regions, may be sensitive to preprocessing or model architecture, and does not by itself prove that the highlighted area is clinically causal.

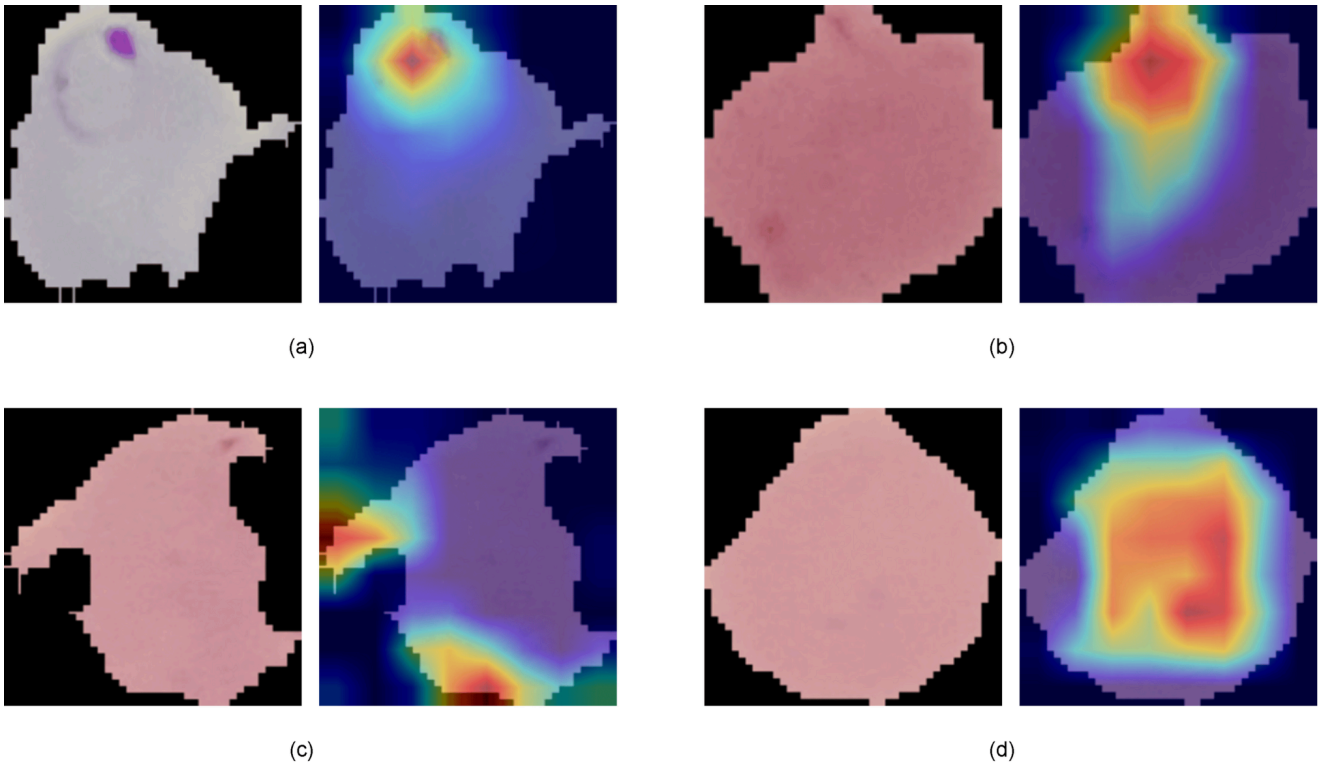


FIGURE 5: Grad-CAM visualization of EfficientNetV2-S (CutMix scenario, fold 1).

Each panel shows the input image (left) and the Grad-CAM overlay for the predicted class (right): (a) TP: actual Parasitized, predicted Parasitized ($p = 0.89$); (b) FP: actual Uninfected, predicted Parasitized ($p = 0.62$); (c) FN: actual Parasitized, predicted Uninfected ($p = 0.41$); and (d) TN: actual Uninfected, predicted Uninfected ($p = 0.08$). p denotes the predicted probability of the Parasitized class.

In all three True Positive cases, the heatmap was concentrated on dark-stained inclusions inside the cell that are visually consistent with parasite-like structures. This interpretation was based on visual comparison with the cell morphology visible in the images, not on expert-annotated parasite masks. The three False Positive cases differed from one another: in one, the activation was centered on the cell interior without a visible inclusion; in another, it lay around a faint dark spot near the upper cell edge; and in the third, it was focused on a distinct dark-stained object at the cell margin. The last case suggests that some false positives may involve staining artifacts or possible labeling errors.

In the three False Negative cases, the cells contained only small or faint stained regions, and the maps for the Uninfected decision were concentrated at the cell boundary and adjacent background rather than on the cell interior. In the three True Negative cases, the activation was spread broadly over the homogeneous cell interior without a focal hotspot. All False Positive and False Negative examples had predicted probabilities relatively close to the 0.5 threshold, indicating low-confidence decisions. These observations are qualitative, are based on 12 images, and should be validated by hematology experts or laboratory analysts in future work.

The deletion-insertion check is summarized in Table *14*. For images predicted as Parasitized, Grad-CAM maps gave a lower deletion AUC than random maps in all 106 images (0.345 ± 0.108 vs. 0.645 ± 0.091) and a higher insertion AUC in 94.3% of images (0.816 ± 0.087 vs. 0.704 ± 0.114), indicating that the highlighted regions were the ones driving these predictions. For images predicted as Uninfected, the insertion AUC was also higher than that of random maps in 88.3% of images (0.814 ± 0.074 vs. 0.772 ± 0.082), but the deletion AUC was higher than that of random maps in 98.9% of images (0.785 ± 0.075 vs. 0.666 ± 0.058). Removing the regions highlighted by Grad-CAM therefore did not reduce the Uninfected probability. A plausible explanation is that an Uninfected decision reflects the absence of parasite-like features rather than the presence of a localized feature. Overall, this check supports the faithfulness of the Grad-CAM maps for Parasitized predictions, whereas the maps for Uninfected predictions should be interpreted with more caution.

Group	*n*	Deletion AUC (Grad-CAM)	Deletion AUC (random)	Insertion AUC (Grad-CAM)	Insertion AUC (random)
All images	200	0.552 ± 0.239	0.655 ± 0.078	0.815 ± 0.081	0.736 ± 0.106
Predicted Parasitized	106	0.345 ± 0.108	0.645 ± 0.091	0.816 ± 0.087	0.704 ± 0.114
Predicted Uninfected	94	0.785 ± 0.075	0.666 ± 0.058	0.814 ± 0.074	0.772 ± 0.082

TABLE 14: Deletion and insertion AUC of Grad-CAM maps compared with random maps on 200 randomly selected fold-1 validation images (mean ± SD).

Note. Lower deletion AUC and higher insertion AUC indicate a more faithful map. All paired differences between Grad-CAM and random maps: $p < 0.001$ (Wilcoxon signed-rank test).

Overall, in the examined cases, EfficientNetV2-S attended to internal cell regions resembling parasite-related structures in correctly classified Parasitized images, and the deletion-insertion check supported the faithfulness of these maps. However, the model could still be influenced by staining artifacts, texture variation, color intensity, or boundary regions, and the maps for Uninfected predictions were not supported by the deletion test. These findings show that Grad-CAM is useful for exploring correct predictions and possible sources of classification error, but its localization reliability is not guaranteed. Quantitative localization evaluation and expert assessment are still needed before these visual explanations can support clinical decision-making.

No medical or laboratory expert was involved in the XAI analysis of this study. The interpretation of the heatmaps, including the judgment of whether the highlighted regions correspond to parasite-like structures, was made by the authors, whose background is in informatics, on the basis of the cell morphology visible in the images. The explainability results should therefore be regarded as technical observations rather than clinically validated findings. Future validation should involve certified malaria microscopists or hematology experts in three ways: (i) annotating parasite regions in a sample of validation images, so that agreement between the heatmaps and expert annotations can be measured quantitatively, for example with intersection over union; (ii) independently rating whether each heatmap highlights diagnostically relevant structures, with inter-rater agreement reported; and (iii) reviewing the False Positive and False Negative cases to distinguish model errors from possible labeling errors or staining artifacts.

From a clinical perspective, the value of these explanations lies in supporting human review rather than replacing it. In a screening workflow, the classifier could flag cells as Parasitized, and the corresponding Grad-CAM map would show the microscopist where the evidence for that decision lies. A heatmap centered on an intracellular dark-stained inclusion gives the reviewer a specific region to confirm, whereas a heatmap concentrated on the cell edge or on a staining artifact, as seen in some of the False Positive examples, signals that the prediction should be treated with caution. Recurrent artifact-driven activations across many cells could also point to slide-preparation or staining-quality problems. In this way, explainability can make model errors easier to recognize, support appropriate trust, and keep the final diagnostic decision with the human expert. These potential benefits were not measured in this study and should be confirmed through reader studies involving hematology experts.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

Clinical implications and limitations

The results of this study indicate the potential of EfficientNetV2-S as an initial component in a microscopic image-based malaria screening support system. The recall value of 96.33% under the CutMix scenario suggests that the model had high sensitivity in identifying malaria-positive samples. This is relevant for preliminary screening because false negative predictions may cause infected samples to be missed. However, the model cannot yet be considered ready for direct clinical deployment. External validation on more diverse clinical datasets is still required, especially because real-world microscopic images may involve variations in staining, focus quality, illumination, artifacts, and parasite species that differ from the training dataset. The burden of malaria also falls most heavily on vulnerable communities [1], and climate change may further widen the health inequities affecting such populations [39]. Diagnostic support tools should therefore be developed and evaluated so that they serve these groups fairly, and future validation should include data from endemic, low-resource settings where diagnostic support is needed most, such as the eastern regions of Indonesia described in the Introduction.

Compared with previous studies, the main contribution of this study lies in the integration of five-backbone benchmarking, four augmentation scenarios, ROC-AUC-based non-parametric statistical testing, computational efficiency evaluation, and visual interpretability using Grad-CAM. The benchmarking used EfficientNetV2-S, ResNet34, ResNet50, VGG16, and VGG19 under the same preprocessing within each augmentation scenario and 5-fold stratified cross-validation. The augmentation analysis compared no augmentation, MixUp, CutMix, and MixUp + CutMix. Grad-CAM was applied to the best-performing EfficientNetV2-S configuration to inspect representative TP, FP, FN, and TN predictions. Several previous studies have reported high performance on malaria datasets, but they did not always include statistical validation among backbones, modern augmentation analysis, computational efficiency, and explicit XAI evaluation in a single framework.

Several limitations should be acknowledged. First, the evaluation was conducted on the NIH Malaria Cell Images Dataset, which contains single-cell images acquired under relatively controlled conditions. This dataset does not fully represent the variability of microscopic images in real healthcare facilities, particularly in endemic regions with limited equipment and diverse slide preparation quality. Second, data splitting was performed at the sample level with class stratification, and patient-level splitting was not applied. Cells from the same patient may therefore appear in both the training and validation folds, so the model's ability to generalize to independent patients, institutions, or acquisition devices cannot yet be confirmed. Third, the dataset used in this study focused on binary classification between Parasitized and Uninfected classes, so the model has not yet been evaluated for distinguishing Plasmodium species or estimating parasitemia levels. Fourth, the statistical evaluation used five cross-validation folds, meaning that *p*-values should still be interpreted together with effect size, mean rank, and performance consistency rather than being used as the sole basis for conclusions. Fifth, Grad-CAM interpretation remains largely qualitative. It was examined on 12 images and supported by a deletion-insertion check on 200 images, which indicated faithful maps for Parasitized predictions but not for Uninfected predictions, and it has not yet been validated by hematology experts, laboratory analysts, or quantitative localization metrics. Sixth, a scenario with basic augmentation alone was not evaluated, so the separate contributions of basic and mixing-based augmentation relative to the resize-only baseline could not be isolated. Seventh, each configuration was trained once with a single random seed and without repeated cross-validation, so run-to-run variability was not estimated. The reported confidence intervals account for the overlap between training folds but not for seed-related variability. Eighth, hyperparameters were not tuned for each backbone: all models were trained for eight epochs with fixed Focal Loss settings, and VGG16 and VGG19 used a lower learning rate and gradient accumulation. Although the validation AUC of all backbones plateaued within the eight epochs (Figure *4*), these choices may favor architectures that converge quickly, and per-backbone tuning could change the size of the observed differences.

Future research should conduct external validation on more diverse clinical datasets, including data from multiple locations, different microscope devices, staining variations, and broader *Plasmodium* species. System development can also be directed toward slide-level or field-of-view image analysis to better align with the microscopic diagnostic workflow in healthcare facilities. In addition, XAI evaluation should involve medical experts and quantitative validation

frameworks to assess whether model activation regions correspond to clinically relevant morphological features. Through these developments, the EfficientNetV2-S-based deep-learning approach can be further evaluated as a more transparent, efficient, and reliable support tool for malaria diagnosis.

Conclusions

This study presented an explainable deep-learning benchmark for automated malaria cell classification on the NIH Malaria Cell Images Dataset, comparing EfficientNetV2-S with four classical CNN backbones (VGG16, VGG19, ResNet34, and ResNet50) under a unified preprocessing, training, and evaluation protocol. Its main contribution is methodological: five backbones, four augmentation scenarios, non-parametric statistical testing with effect sizes, computational efficiency, and Grad-CAM explainability with a faithfulness check were evaluated on the same folds, with the same training settings and the same trained models. To our knowledge, this combination has not been reported in previous malaria cell classification studies; among the 15 earlier studies summarized in Table *1*, none evaluated mixing-based augmentation, and only one reported saliency-based explanations. Evaluating these aspects together revealed relationships that separate analyses would have missed. Under the CutMix scenario, the backbone with the best discriminative performance also had the fewest false negatives and the smallest model, but the slowest GPU inference, and its Grad-CAM maps were faithful for Parasitized predictions but not for Uninfected predictions. Across all augmentation scenarios, EfficientNetV2-S provided the most consistent discriminative performance and the highest mean rank among the evaluated backbones, achieving its best result under the CutMix scenario. The Friedman test confirmed significant global differences among the backbones, although the Nemenyi post-hoc analysis indicated that this superiority was statistically significant mainly relative to ResNet34 and, under CutMix, to VGG19. EfficientNetV2-S should therefore be regarded as the best-performing and most consistently ranked backbone rather than one that significantly outperforms every baseline in all pairwise comparisons. The absolute ROC-AUC differences among backbones were small (below 0.002). The ablation suggested that augmentation increased ROC-AUC in a model-dependent manner, with CutMix numerically best for EfficientNetV2-S, although these effects were not statistically significant after correction for multiple comparisons, suggesting that augmentation strategy should be matched to the dataset and architecture rather than assumed to benefit from combination.

From a deployment perspective, the results show a clear trade-off between architectural efficiency and measured runtime. EfficientNetV2-S had the fewest parameters (20.84 million), the lowest MACs (3.54 G), and the smallest model size (79.49 MB). This advantage was large relative to VGG16 and VGG19, which were about 6.5 times larger, but small relative to ResNet34 and ResNet50, which had 3% to 18% more parameters. In contrast, EfficientNetV2-S was the slowest backbone on the tested GPU (23.87 ms per image, 2.2 to 3.9 times slower than the baselines) and was also slower than ResNet34 and ResNet50 on CPU (63.16 vs. 43.49 and 56.24 ms per image), so its lower MAC count did not translate into faster inference. In practice, EfficientNetV2-S is preferable when storage is limited and missing infected cells is the main concern, whereas ResNet34 is preferable when inference speed is the priority. Even at its slowest measured speed (about 16 images per second on CPU), EfficientNetV2-S would classify 1,000 cell images in roughly one minute, although this estimate excludes cell segmentation and the other steps of a complete screening pipeline. Grad-CAM analysis provided qualitative interpretability, indicating that, in the examined cases, the model attended to parasite-like stained inclusions in correctly classified infected cells, while misclassified cells showed low-confidence predictions with activations on faint staining irregularities or cell-boundary regions; these explanations remain supportive rather than clinically conclusive. In conclusion, EfficientNetV2-S is a promising backbone for automated malaria cell classification, combining strong discriminative performance, the fewest missed infected cells, consistent statistical ranking, and a compact model size, at the cost of slower inference, together with qualitative Grad-CAM observations consistent with parasite-related regions in correct predictions. Nevertheless, the findings are limited by the use of a single public dataset, sample-level validation, fixed hyperparameter settings, and qualitative explainability. Future work should prioritize external validation on clinically diverse datasets, evaluation across Plasmodium species, slide-level or field-of-view analysis, expert and quantitative validation of explainability outputs, threshold optimization for the intended clinical use, and deployment testing on realistic edge devices before clinical translation.

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Purna Siswantomo, Ahmad Ilham, Wendy Sarasjati

Acquisition, analysis, or interpretation of data: Purna Siswantomo

Drafting of the manuscript: Purna Siswantomo

Critical review of the manuscript for important intellectual content: Ahmad Ilham, Wendy Sarasjati

Supervision: Ahmad Ilham, Wendy Sarasjati

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in GitHub at https://github.com/Purna-Siswantomo/CNN-multi-backbone.git, reference N/A.

Acknowledgements

The authors would like to express sincere gratitude to Ahmad Ilham and Wendy Sarasjati for their guidance, supervision, and constructive input throughout the development of this research. This study was conducted as part of the academic supervision process and prepared as a publication output to fulfill the requirements for completing the Bachelor's degree in Informatics at Universitas Muhammadiyah Semarang. The author also acknowledges that this research used a publicly available dataset, the NIH Malaria Cell Images Dataset. The data used in this study are publicly accessible, and the analytical methods and study materials are available from the corresponding author upon reasonable request.

References

1. World Health Organization: World malaria report 2025. (2025). Accessed: April 25, 2026: https://www.who.int/publications/i/item/9789240117822.
2. Lusiyana N, Rochmah SFN : Factors associated with active multiple episode of malaria in high high-endemicity area of malaria. Journal of Community Health. 2025, 11:410-418. 10.25311/keskom.Vol11.Iss2.2207
3. Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G: Image analysis and machine learning for detecting malaria. Translational Research. 2018, 194:36-55. 10.1016/j.trsl.2017.12.004

4. Litjens G, Kooi T, Bejnordi BE, et al.: A survey on deep learning in medical image analysis. Medical Image Analysis. 2017, 42:60-88. 10.1016/j.media.2017.07.005

5. Muhandisin A, Azhar Y: Malaria blood cell image classification using transfer learning with fine-tune ResNet50 and data augmentation. Jurnal RESTI. 2022, 6:891-897. 10.29207/resti.v6i5.4322

6. Jameela T, Athota K, Singh N, Gunjan VK, Kahali S: Deep learning and transfer learning for malaria detection. Computational Intelligence and Neuroscience. 2022, 2022:2221728. 10.1155/2022/2221728

7. Laghari AA, Muhammad W, Memon ML, Hussain A, Kumar A: Malaria parasite cell classification using transfer learning with state-of-the-art CNN architectures. Biology. 2025, 14:1792. 10.3390/biology14121792

8. Hemachandran K, Alasiry A, Marzougui M, Ganie SM, Pise AA, Alouane MTH, Chola C: Performance analysis of deep learning algorithms in diagnosis of malaria disease. Diagnostics. 2023, 13:534. 10.3390/diagnostics13030534

9. Mujahid M, Rustam F, Shafique R, Montero EC, Alvarado ES, de la Torre Diez I, Ashraf I: Efficient deep learning-based approach for malaria detection using red blood cell smears. Scientific Reports. 2024, 14:13249. 10.1038/s41598-024-63831-0

10. Boit S, Patil R: An efficient deep learning approach for malaria parasite detection in microscopic images. Diagnostics. 2024, 14:2738. 10.3390/diagnostics14232738

11. Kiki MRM, Sari Z, Rizki D: Comparison of EfficientNetB0 and EfficientNetB7 models in classifying malaria based on blood cells. Jurnal Online Informatika. 2024, 9:219-227. 10.15575/join.v9i2.1195

12. Shahin OR, Alshammari HH, Alabdali RN, Salaheldin AM, Saleh N: Automated multi-model framework for malaria detection using deep learning and feature fusion. Scientific Reports. 2025, 15:25672. 10.1038/s41598-025-04784-w

13. Bhuiyan M, Islam MS: A new ensemble learning approach to detect malaria from microscopic red blood cell images. Sensors International. 2023, 4:100209. 10.1016/j.sintl.2022.100209

14. Reddy CKK, Anisha PR, Almushharaf A, Talla R, Baili J, Cho Y, Nam Y: An optimized transfer learning approach integrating deep convolutional feature extractors for malaria parasite classification in erythrocyte microscopy. Frontiers in Medicine. 2025, 12:1684973. 10.3389/fmed.2025.1684973

15. Siłka W, Wieczorek M, Siłka J, Woźniak M: Malaria detection using advanced deep learning architecture. Sensors. 2023, 23:1501. 10.3390/s23031501

16. Acherar A, Tantaoui I, Thellier M, Lampros A, Piarroux R, Tannier X: Real-life evaluation of deep learning models trained on two datasets for Plasmodium falciparum detection with thin blood smear images at 500x magnification. Informatics in Medicine Unlocked. 2022, 35:101132. 10.1016/j.imu.2022.101132

17. Sukumarran D, Hasikin K, Mohd Khairuddin AS, Ngui R, Wan Sulaiman WY, Vythilingam I, Divis PCS: Machine and deep learning methods in identifying malaria through microscopic blood smear: A systematic review. Engineering Applications of Artificial Intelligence. 2024, 133:108529. 10.1016/j.engappai.2024.108529

18. Swastika W, Pradana BJ, Widodo RB, Sitepu R, Putra GGS: CNN based transfer learning for malaria parasite detection using thin-blood smear images. SIET '22: Proceedings of the 7th International Conference on Sustainable Information Engineering and Technology. 2022, 72-78. 10.1145/3568231.3568243

19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision. 2020, 128:336-359. 10.1007/s11263-019-01228-7

20. Houssein EH, Gamal AM, Younis EMG, Mohamed E: Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. Cluster Computing. 2025, 28:469. 10.1007/s10586-025-05281-5

21. Parveen R, Qui B, Song W, et al.: Trustworthy deep learning for malaria diagnosis using explainable artificial intelligence. Scientific Reports. 2025, 15:45037. 10.1038/s41598-025-28387-7

22. Muhammad D, Bendechache M: Unveiling the black box: a systematic review of Explainable Artificial Intelligence in medical image analysis. Computational and Structural Biotechnology Journal. 2024, 24:542-560. 10.1016/j.csbj.2024.08.005

23. Simonyan K, Vedaldi A, Zisserman A: Deep inside convolutional networks: visualising image classification models and saliency maps [PREPRINT]. arXiv. 2014, 10.48550/arXiv.1312.6034

24. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR: Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications. 2019, 10:1096. 10.1038/s41467-019-08987-4

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9

25. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B: Sanity checks for saliency maps. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada. 2018, 31:1-15.

26. Rajaraman S, Antani SK, Poostchi M, et al.: Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ. 2018, 6:e4568. 10.7717/peerj.4568

27. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D: mixup: beyond empirical risk minimization. International Conference on Learning Representations (ICLR). 2018, 1-13.

28. Yun S, Han D, Chun S, Oh SJ, Yoo Y, Choe J: CutMix: regularization strategy to train strong classifiers with localizable features. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South). 2019, 6022-6031. 10.1109/ICCV.2019.00612

29. Kim J-H, Choo W, Jeong H, Song HO: Co-Mixup: saliency guided joint mixup with supermodular diversity. International Conference on Learning Representations (ICLR). 2021, 1-21.

30. Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR). 2015, 1-14.

31. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, 10.1109/CVPR.2016.90

32. Tan M, Le Q: EfficientNetV2: smaller models and faster training. Proceedings of Machine Learning Research. 2021, 139:10096-10106.

33. Lin TY, Goyal P, Girshick R, He K, Dollár P: Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy. 2017, 2999-3007. 10.1109/ICCV.2017.324

34. Loshchilov I, Hutter F: Decoupled weight decay regularization. International Conference on Learning Representations (ICLR). 2019, 1-18.

35. Nadeau C, Bengio Y: Inference for the generalization error. Machine Learning. 2003, 52:239-281. 10.1023/a:1024068626366

36. Holm S: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979, 6:65-70.

37. Petsiuk V, Das A, Saenko K: RISE: randomized input sampling for explanation of black-box models. British Machine Vision Conference. 2018, 1-13.

38. Tan M, Le QV: EfficientNet: rethinking model scaling for convolutional neural networks. International Conference on Machine Learning. 2019, 1-10.

39. Shah I, Manda VK, Saruparia C, Younus M: Assessing the impact of climate change adaptation measures on health equity across vulnerable populations. Sustainability Education, Youth and Innovation. Dsilva J (ed): Springer, Cham; 2026. 275-290. 10.1007/978-3-032-21948-0_32

How to cite this article:
Siswantomo P, Ilham A, Sarasjati W (November 19, 2026) Explainable Deep Learning for Automated Malaria Cell Classification Using EfficientNetV2-S, Classical CNN Backbones, and Grad-CAM. Cureus J Comput Sci 3 : es44389-026-00380-9. DOI https://doi.org/10.7759/s44389-026-00380-9