Iranian Journal of Blood & Cancer

Iran J Blood Cancer, 2025, 17(3), 62-72

Original Study

Comparative Evaluation of Custom Convolutional Neural Networks and EfficientNet-B3 for Malaria Cell Image Classification: Impact of Targeted Data Augmentation on Model Performance

Reza Mohit[1], Emad Milani[2], Ata Amini[3], Mehrnaz Ahani[4], Elham Nazari[5*], Tahmineh Aldaghi[6*]

[1]Department of Anesthesia, School of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[2]Student Research Committee, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[3]Department of Health Information Management and Medical Informatics, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran.
[4]Department of Midwifery, School of Nursing and Midwifery, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[5]Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[6]Institute of Biophysics and Informatics, First Faculty of Medicine, Charles University, Prague, Czech Republic.

Scan and read the article online

Citation Mohit R, Milani E, Amini A, Ahani M, Nazari E, Aldaghi T. Comparative Evaluation of Custom Convolutional Neural Networks and EfficientNet-B3 for Malaria Cell Image Classification: Impact of Targeted Data Augmentation on Model Performance. Iran J Blood Cancer. 2025 Sep 30;17(3): 62-72.

doi 0

Article info:

Received: 10 July 2025
Accepted: 10 Sep 2025
Published: 30 Sep 2025

Abstract

Background: Malaria diagnosis with thin blood smears remains labor-intensive and relies on the operator. Deep learning could enable accurate automation.

Objective: Compare four convolutional approaches for classifying parasitized versus uninfected erythrocytes and to evaluate whether targeted image-quality augmentations enhance performance.

Materials and Methods: We used the balanced NIH/Kaggle dataset, which included 13,780 parasitized and 13,780 uninfected samples. Data were split stratified into training, validation, and test sets (70/15/15). Images were resized to 256×256 and normalized. Four experiments were conducted: (1) a custom CNN; (2) the same CNN with targeted augmentation applied to 20% of training samples per class—using Contrast Limited Adaptive Histogram Equalization [CLAHE] and controlled brightness adjustment—and augmented images were added back to the training set (totaling 30,864 images); (3) a soft-attention parallel CNN (SPCNN); and (4) transfer learning with EfficientNet-B3 on 300×300 inputs with full fine-tuning. Evaluation metrics included accuracy, precision, recall, F1 score, and AUC-ROC.

Results: EfficientNet-B3 achieved the highest performance with a validation accuracy of 0.9741, 98% precision, 96% recall, an F1 score of 0.97, and an AUC-ROC of 0.9964. SPCNN was competitive but slightly lower, with a validation accuracy of 0.9652, 98% precision, 95% recall, an F1 score of 0.96, and an AUC-ROC of 0.9909. The baseline CNN had a validation accuracy of 0.9649, 97% precision, 94% recall, an F1 score of 0.96, and an AUC-ROC of 0.9910. Targeted augmentation resulted in negligible change compared to the baseline CNN, with a validation accuracy of 0.9647, an F1 score of 0.96, and an AUC-ROC of 0.9908, indicating limited added discriminative value for this dataset.

Conclusion: EfficientNet-B3 outperformed SPCNN and custom CNNs. The CLAHE/brightness strategy applied to 20% of training images and added back to the dataset did not significantly improve generalization. External validation and prospective field testing are necessary before clinical deployment.

Keywords:

Malaria
Deep learning
EfficientNet-B3
Albumentations
Parasitized cell images
Medical imaging AI
Soft-attention parallel CNN

* Corresponding Author:
Elham Nazari
E-mail: **Elham.Nazari@sbmu.ac.ir**

Tahmine Aldaghi
E-mail: **Tahmineh1989@gmail.com**

[DOI: 10.61882/ijbc.17.3.62]

1. INTRODUCTION

Malaria is still one of the most critical global health issues, particularly in tropical and subtropical regions, where it affects hundreds of millions of people annually. In 2024, the World Health Organization released a report that mentioned there were 263 million cases and 597,000 deaths from malaria worldwide in 2023—that is 11 million more cases than in 2022 [1,2]. Most malaria deaths happen in the African Region, including Nigeria, the Democratic Republic of Congo, and Uganda [1,3]. The disease mainly affects vulnerable groups, especially children under five and pregnant women [3].

Recent advances in malaria vaccines mark significant public health achievements. The RTS, S/AS01 vaccine, which was introduced in 2021, was the first malaria vaccine recommended by the WHO against a parasitic disease. At first, the vaccine showed an average efficacy of around 50% in large-scale trials, but over time, its protective effect decreased, which shows the need for booster doses [4]. In 2023, a second vaccine—R21/Matrix-M—gained WHO prequalification. In phase 2b trials in Burkina Faso, it showed 75% efficacy after a three-dose regimen [5]. However, relying solely on vaccine-based strategies is not enough to eliminate malaria. Effective disease control also depends on timely and accurate diagnosis, especially in resource-limited areas [6,7].

Traditional malaria diagnostic methods include blood smear microscopy, rapid diagnostic tests (RDTs), and polymerase chain reaction (PCR). Microscopy remains the gold standard, but its sensitivity depends on the operator's skill and parasite density [8]. RDTs are widely used because of their simplicity but can fail in areas with HRP2 gene deletions or low parasitemia [9]. PCR provides high diagnostic accuracy but is not feasible for routine use in endemic regions because of costs and infrastructure requirements [10].

Machine learning (ML), especially computer vision techniques, provides scalable options for automated malaria diagnosis. Convolutional neural networks (CNNs) have shown excellent performance in identifying Plasmodium-infected red blood cells (RBCs) in thin smear images. [11,12]. These models can learn hierarchical features directly from raw images, outperforming traditional handcrafted feature methods. Additionally, CNN-based systems have been developed for tasks beyond binary classification, such as parasite staging and morphological differentiation—critical for clinical decision-making. [13].

Several CNN strategies have been examined for malaria diagnosis, including custom models built from scratch, transfer learning with pretrained architectures like EfficientNet, and models improved with attention mechanisms [14,15]. Data augmentation, normalization, and pruning are frequently used to enhance generalization, especially in datasets with limited diversity [16,17]. More recently, architectures such as Soft Attention Parallel CNNs (SPCNNs) have been developed to improve feature localization and multiscale representation, resulting in high diagnostic accuracy [18].

This study aims to systematically evaluate three CNN-based pipelines for automated malaria detection using microscopic blood smear images: (i) a custom CNN trained from scratch, with and without Albumentations augmentation, (ii) a transfer learning model based on EfficientNet B3, and (iii) a Soft Parallel CNN (SPCNN) that includes attention-guided feature extraction. Each model is evaluated for diagnostic accuracy, computational efficiency, and suitability for deployment in low-resource clinical settings. By comparing these approaches on the same dataset, the study offers insights into how architectural choices and preprocessing methods influence the effectiveness of ML-driven malaria diagnosis.

2 LITERATURE REVIEW

Automated analysis of thin blood smear images to detect malaria has greatly improved over the past decade due to the high number of malaria cases in places with few resources, and it is hard for people to do manual checks reliably. Early methods relied on handcrafted features and traditional machine-learning models, but these were replaced by deep convolutional neural networks (CNNs) as labeled data and computing power became available [19,20].

Rajaraman et al. [19] showed that using pre-trained CNNs like VGG and ResNet can help classify malaria cells better than older methods. Later studies showed that combining different deep learning models can improve detection results and lower error rates in thin smear images [20].

The Kaggle malaria dataset, which comes from the National Institutes of Health (NIH), includes about 27,558 RGB images of cells, half of which are parasitized and uninfected classes [21]. This dataset is now a standard for testing and comparing CNNs, and it helps researchers conduct open evaluations. Based on the literature, there are two main ways to approach this problem:

1. Training custom CNNs from scratch, which are tailored to microscopy data but require large datasets for optimal generalization.

[DOI: 10.61882/ijbc.17.3.62]

2. Transfer learning, which involves fine-tuning pretrained models such as VGG, ResNet, MobileNet, and EfficientNet on malaria images [19,22].

When the dataset size is small, transfer learning approaches show faster convergence and better generalization compared to models trained from scratch [19,22]. For instance, studies using different versions of EfficientNet (B0–B7) have reported state-of-the-art accuracy, taking advantage of the architecture's compound scaling strategy [23].

Data augmentation is a common technique in malaria detection research. Shorten and Khoshgoftaar [24] reviewed various techniques and pointed out that geometric transformations, photometric adjustments, and histogram equalization (like CLAHE) can improve model robustness. In microscopy, CLAHE, moderate brightness and contrast adjustments, and small affine transformations are especially effective because they mimic real variations in image capturing [24]. However, the benefits heavily depend on the specific task and the quality of the original images.

Besides regular CNNs, attention mechanisms and parallel architectures have been introduced to capture both local parasite morphology and broader cellular context. Recent works on Soft Attention Parallel CNNs (SPCNN) have reported performance improvements over conventional transfer learning baselines while remaining efficient enough for deployment in resource-limited settings. [22].

In summary, previous research indicates that combining a well-structured experimental protocol—including the use of a balanced public dataset [21], baseline and augmented custom CNNs [19,20], attention-based models [22], and modern pretrained architectures like EfficientNet [24]—along with standardized evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC—represents the current best practice for classifying malaria cell images. This framework directly guides the experimental design of the present study.

3. MATERIALS AND METHODS

3.1. Dataset and Preprocessing

This study used a publicly available dataset called Cell Images for Malaria Detection, which is on Kaggle, and includes 27,560 labeled microscopic RGB images evenly split into two categories: Parasitized (n = 13,780) and Uninfected (n = 13,780) [25]. To maintain class balance, a stratified folder-based split was applied, resulting in 70% training, 15% validation, and 15% test sets, corresponding to 19,320, 4,140, and 4,140 samples, respectively.

All images were resized to 256×256 pixels to improve the original dataset's low resolution and blurring, as observed during initial visual inspection. Image normalization was performed per channel using a mean and standard deviation of (0.5, 0.5, 0.5). Data loading and batching were handled with PyTorch's DataLoader, using a batch size of 32.

3.2. Baseline Model: Custom CNN Architecture

The base model was a custom-designed convolutional neural network (CNN) with six layers, featuring the following channel progression: 32 → 64 → 128 → 128 → 256 → 512. Each layer used a 3×3 kernel, stride = 1, and padding = 1, followed by ReLU activation and dropout (p = 0.2). The output was flattened and fed into a fully connected layer with 512 neurons, followed by another dropout layer (p = 0.2), and a final sigmoid-activated neuron for binary classification.

The model was trained using Binary Cross-Entropy with Logits (BCEWithLogitsLoss) and the Adam optimizer with an initial learning rate of 0.001. A learning rate scheduler (patience = 3, factor = 0.2, min_lr = 1e-5) and early stopping (patience = 5) were employed to prevent overfitting. Training was carried out on a dual-GPU system (NVIDIA T4 ×2) using PyTorch's DataParallel module. The model achieved its best validation loss (0.1121) at epoch 16, with training ending at epoch 21 through early stopping.

3.3. Data Augmentation Experiments

To evaluate how targeted augmentation affects model performance, 20% of the training data was modified using the Albumentations library [26,27]. These augmentations were intentionally selected to address the inherent blurriness and low resolution seen in the original dataset. Enhancements included Contrast Limited Adaptive Histogram Equalization (CLAHE) and controlled brightness adjustments, with the goal of improving visual clarity, highlighting morphological features, and better simulating realistic imaging variations in both parasitized and uninfected cells. The augmented images were then added back into the original training set to enhance diversity and support better model learning.

The following transformations were applied with specified parameters:

- RandomBrightnessContrast (limit = 0.2, p = 1.0)
- CLAHE (clip limit = 4.0, tile grid size = 8×8, p = 1.0)
- ShiftScaleRotate (shift = 0.02, scale = 0.05, rotation = ±10°, p = 0.5)

[DOI: 10.61882/ijbc.17.3.62]

Augmented samples were added to the original dataset instead of replacing them, increasing the training set to 30,864 images (15,432 per class). This aimed to encourage the model to generalize across different visual representations of similar samples.

The CNN architecture and training parameters remained the same as the baseline. However, the augmented dataset did not produce significant improvements in key metrics like validation accuracy, AUC, or F1-score. This suggests that the transformations used with this specific approach did not substantially boost the dataset's discriminative power.

3.4. Parallel Attention-Based CNN (SPCNN)

A third experimental model used a parallel CNN architecture called Soft Attention Parallel Convolutional Neural Network (SPCNN), designed to extract multiscale spatial features through dual convolutional streams.

The first stream had four convolutional layers with 3×3 kernels, increasing channel sizes from 32 to 64, 128, and 256, with MaxPooling (stride = 2) and Dropout2D (p = 0.2) for regularization. The second stream was similar but used 5×5 convolutional layers to capture a wider context. Each stream was followed by a soft attention module, made of two 1×1 convolutional layers: the first mapping 256 channels to 256, and the second producing a single-channel attention map using a sigmoid function, which then influenced the feature maps.

The outputs from both streams were pooled to create 256-dimensional vectors, which were combined and fed into a fully connected layer with 256 neurons and a ReLU activation. This was followed by dropout (p = 0.3) and a final sigmoid activation for binary output.

The training settings (optimizer, scheduler, loss function, batch size) for this model matched the baseline configuration. It was trained on dual NVIDIA T4 GPUs converged by epoch 17, with a training loss of 0.1088, training accuracy of 96.30%, validation loss of 0.1026, and validation accuracy of 96.52%. This model showed a slight improvement over the baseline CNN [28].

3.5. Transfer Learning with EfficientNet-B3

The final experimental model used transfer learning with EfficientNet-B3, a CNN that was pretrained on a large image dataset called ImageNet [29] [30] [31]. Several versions of EfficientNet (B0, B3, B4, B5, B7) were initially tested, and B3 was chosen for its balance of accuracy and computational efficiency.

Input images were resized to 300×300 pixels and normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]). No augmentations were applied during this phase. The model's original classification head was replaced with a custom head consisting of a fully connected layer (256 neurons, ReLU), dropout (p = 0.5), and a sigmoid output layer.

Training used the Adam optimizer (lr = 0.001, weight decay = 1e-5), BCEWithLogitsLoss, and a learning rate scheduler (factor = 0.3, patience = 3, min_lr = 1e-5). All layers were unfrozen for full fine-tuning. Training was performed on a single NVIDIA P100 GPU (Kaggle) for up to 15 epochs.

Early stopping was triggered at epoch 12, and the best model was found at epoch 7, with the following results:

- Training Loss: 0.0587
- Training Accuracy: 98.16%
- Validation Loss: 0.0753
- Validation Accuracy: 97.41%

The Results section reports evaluation metrics such as AUC-ROC, confusion matrix, precision, recall, and F1-score. Among all tested models, EfficientNet-B3 achieved the highest overall classification performance.

4. RESULTS

4.1 Dataset Characteristics

The balanced malaria cell image dataset from Kaggle included 27,560 images, with 13,780 images in each class (infected and uninfected). The images were divided into training, validation, and test sets in a 70:15:15 ratio, maintaining class balance across all subsets.

4.2 Experiment 1 — Baseline CNN Model

A custom convolutional neural network was trained on the original dataset without augmentation.

- Input size: (256 × 256)
- Batch size: 32
- Optimizer: Adam(lr=0.001)
- Loss function: BCEWithLogitsLoss

Performance on the test and validation set:

- Train_Accuracy: 0.9635
- Val_Accuracy: 0.9649
- Train_Loss: 0.1068
- Val_loss: 0.1121
- Precision: 97%
- Recall: 94%
- F1-score: 0.96
- AUC-ROC: 0.9910

[DOI: 10.61882/ijbc.17.3.62]

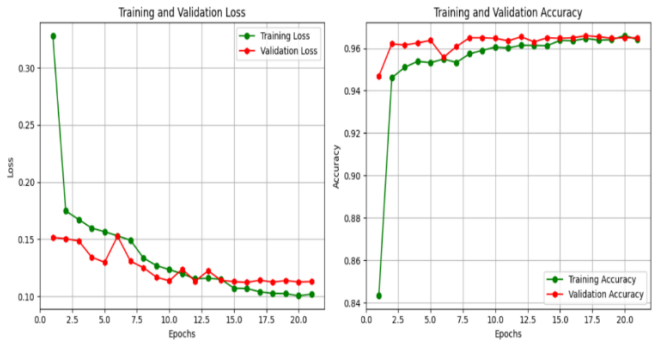
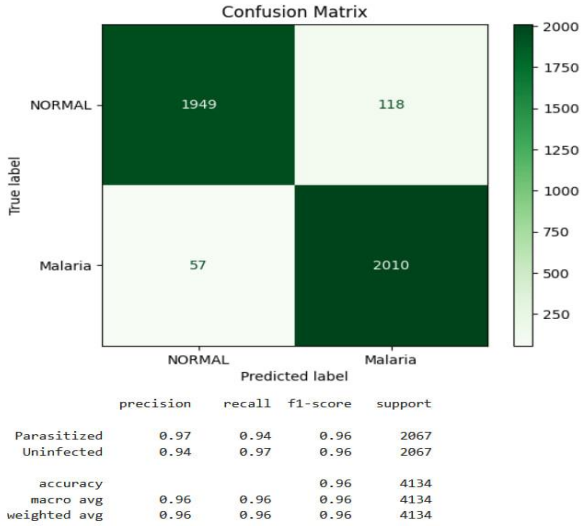


Figure 1. Shows the Learning Curve for Experiment 1.



	precision	recall	f1-score	support
Parasitized	0.97	0.94	0.96	2067
Uninfected	0.94	0.97	0.96	2067
accuracy			0.96	4134
macro avg	0.96	0.96	0.96	4134
weighted avg	0.96	0.96	0.96	4134

Figure 2. Shows the confusion matrix for Experiment 1.

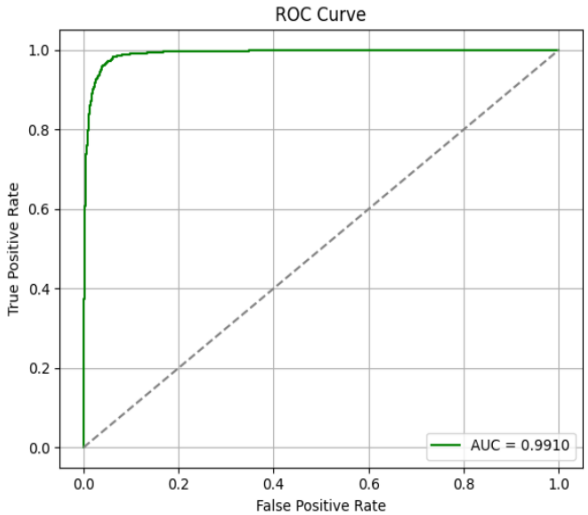


Figure 3. Shows the AUC_Roc Curve. demonstrating the model's ability to distinguish between classes.

4.3 Experiment 2 — CNN with Data Augmentation

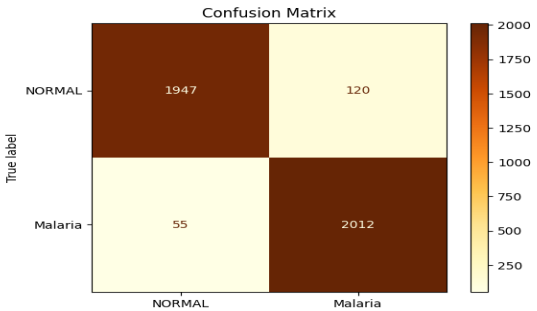
In this experiment, the same CNN architecture was trained with data augmentation applied to 20% of the training samples, then added to the original dataset. Augmentation techniques included: RandomBrightnessContrast (brightness_limit=0.2, contrast_limit=0.2, p=1.0), CLAHE (clip_limit=4.0, tile_grid_size=(8, 8), p=1.0), and ShiftScaleRotate (shift_limit=0.02, scale_limit=0.05, rotate_limit=10, p=0.5).

Performance on the test and validation set:

- Train_Accuracy: 0.9628
- Val_Accuracy: 0.9647
- Train_Loss: 0.1066
- Val_loss: 0.1111
- Precision: 97%
- Recall: 94%
- F1-score: 0.96
- AUC-ROC: 0.9908

Comparison of the baseline custom CNN (Experiment 1) with its augmented counterpart (Experiment 2) showed negligible differences in accuracy, precision, recall, F1-score, and AUC-ROC. The specific augmentation strategy applied to 20% of training samples did not lead to a measurable improvement in generalization performance on the test set. As a result, no augmentation was used in subsequent modeling experiments.

Figure 4. Shows the Learning Curve for Experiment 2.



	precision	recall	f1-score	support
Parasitized	0.97	0.94	0.96	2067
Uninfected	0.94	0.97	0.96	2067
accuracy			0.96	4134
macro avg	0.96	0.96	0.96	4134
weighted avg	0.96	0.96	0.96	4134

Figure 5. Shows a confusion matrix for Experiment 2.

[DOI: 10.61882/ijbc.17.3.62]

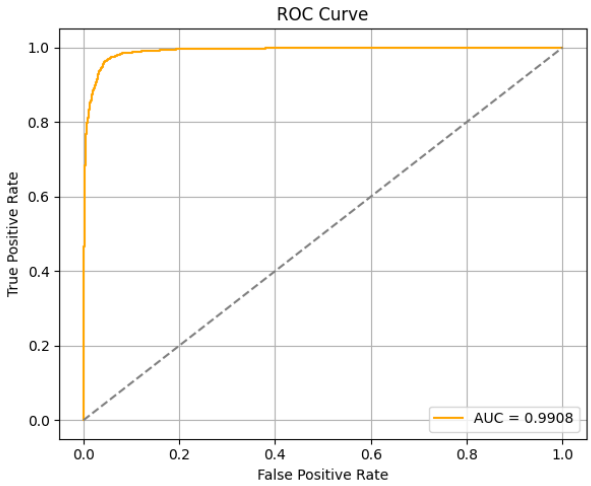


Figure 6. Shows the AUC_Roc Curve. demonstrating the model's ability to distinguish between classes.

4.4. Experiment 3 — Parallel Attention-Based CNN (SPCNN)

In this experiment, a Soft Attention Parallel CNN (SPCNN) architecture was implemented, consisting of two parallel convolutional streams with different kernel sizes to capture multi-scale features, followed by spatial attention blocks, global average pooling, and a fully connected classification head. Dropout layers were included for regularization. The model was trained using the Adam optimizer with a learning rate of 0.001, batch size of 32, and early stopping based on validation loss. Training was performed on an NVIDIA T4*2 GPU, with the lowest validation loss achieved at epoch 17, after which training was halted.

Performance on the test and validation set:

- Train_Accuracy: 0.9630
- Val_Accuracy: 0.9652
- Train_Loss: 0.1088
- Val_loss: 0.1026
- Precision: 98%
- Recall: 95%
- F1-score: 0.96
- AUC-ROC: 0.9909

Compared to both the baseline CNN (Experiment 1) and the augmented CNN (Experiment 2), SPCNN showed slightly improved performance in some metrics, suggesting that using parallel multi-scale feature extraction combined with attention may slightly enhance feature discrimination for malaria parasite detection in this dataset.

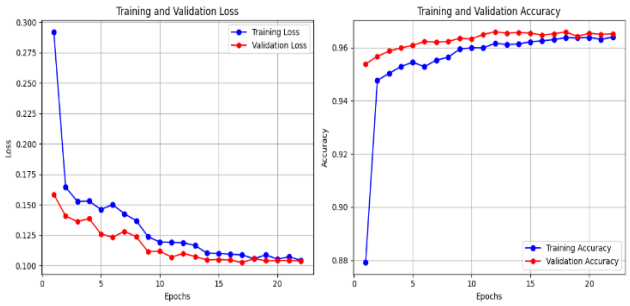
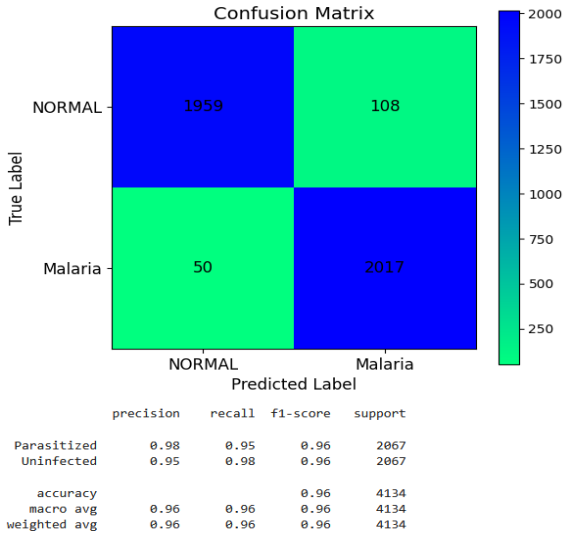


Figure 7. Shows the Learning Curve for Experiment 3.



	precision	recall	f1-score	support
Parasitized	0.98	0.95	0.96	2067
Uninfected	0.95	0.98	0.96	2067
accuracy			0.96	4134
macro avg	0.96	0.96	0.96	4134
weighted avg	0.96	0.96	0.96	4134

Figure 8. Shows the confusion matrix for Experiment 3.

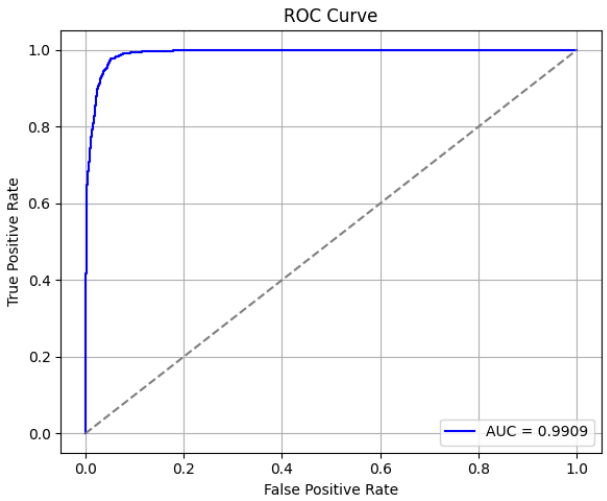


Figure 9. Shows the AUC_Roc Curve. demonstrating the model's ability to distinguish between classes.

[DOI: 10.61882/ijbc.17.3.62]

4.5. Experiment 4 — Transfer Learning with EfficientNet-B3

The final experiment employed **EfficientNet-B3**, initialized with ImageNet weights. Images were resized to **300×300**, normalized, and fed into the network.

- Loss function: BCEWithLogitsLoss
- Optimizer: Adam
- Learning rate: 0.001
- Early stopping: patience = 5
- Fine-tuning: All layers unfrozen and Trainable.

Performance on the test and validation set:

- Train_Accuracy: 0.9816
- Val_Accuracy: 0.9741
- Train_Loss: 0.0587
- Val_loss: 0.0753
- Precision: 98%
- Recall: 96%
- F1-score: 0.97
- AUC-ROC: 0.9964

This model achieved the highest performance among all experiments.

Figure 11 shows the confusion matrix for Experiment 3, while Figure 12 displays the ROC curve, emphasizing the model's superior classification ability.

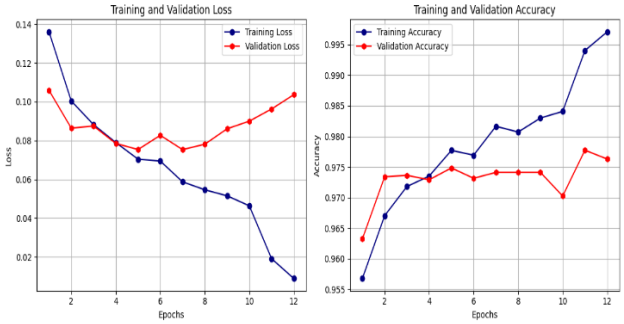


Figure 10. Show the Learning Curve for Experiment 4.

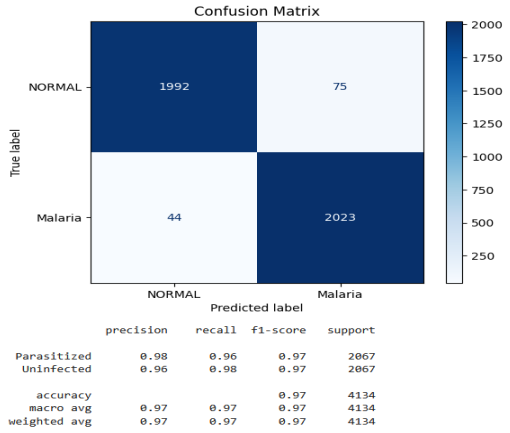


Figure 11. Shows a confusion matrix for Experiment 4.

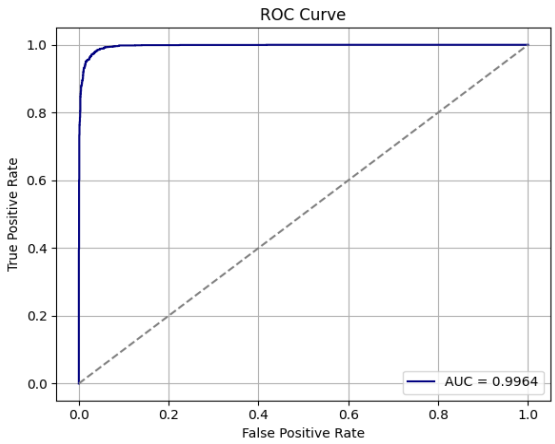


Figure 12. Shows the AUC_Roc Curve: Demonstrating the model's ability to distinguish between classes.

4.6. Comparative Summary

Table 1 summarizes the comparative performance of the four experimental setups. The baseline model (Exp1), trained without augmentation or pretraining, achieved strong results with a validation accuracy of 96.49% and an AUC-ROC of 0.991. Adding augmentation with CLAHE, brightness/contrast adjustments, and small affine transformations (Exp2) did not significantly change performance, indicating limited benefit of these strategies in this dataset. Incorporating early stopping (Exp3) slightly reduced overfitting, with a slightly lower validation loss (0.1026 vs. 0.1111) and similar accuracy (96.52%). The transfer learning approach using ImageNet initialization and resized input images (300×300) (Exp4) greatly outperformed the other setups, achieving the highest training and validation accuracy (98.16% and 97.41%, respectively), the lowest losses, and better generalization as shown by the AUC-ROC (0.9964). These results suggest that pretrained feature representations combined with careful regularization offer the most significant improvement for malaria cell classification in this study.

5. DISCUSSION

This study compared four CNN-based strategies for malaria parasite detection using the NIH/Kaggle dataset: a baseline custom CNN, a CNN with targeted augmentations, a soft-attention parallel CNN (SPCNN), and transfer learning with EfficientNet-B3. Among these, EfficientNet-B3 achieved the highest overall accuracy (val accuracy 0.9741; F1-score 0.97; AUC 0.9964). The other approaches produced competitive but slightly lower performance,

[DOI: 10.61882/ijbc.17.3.62]

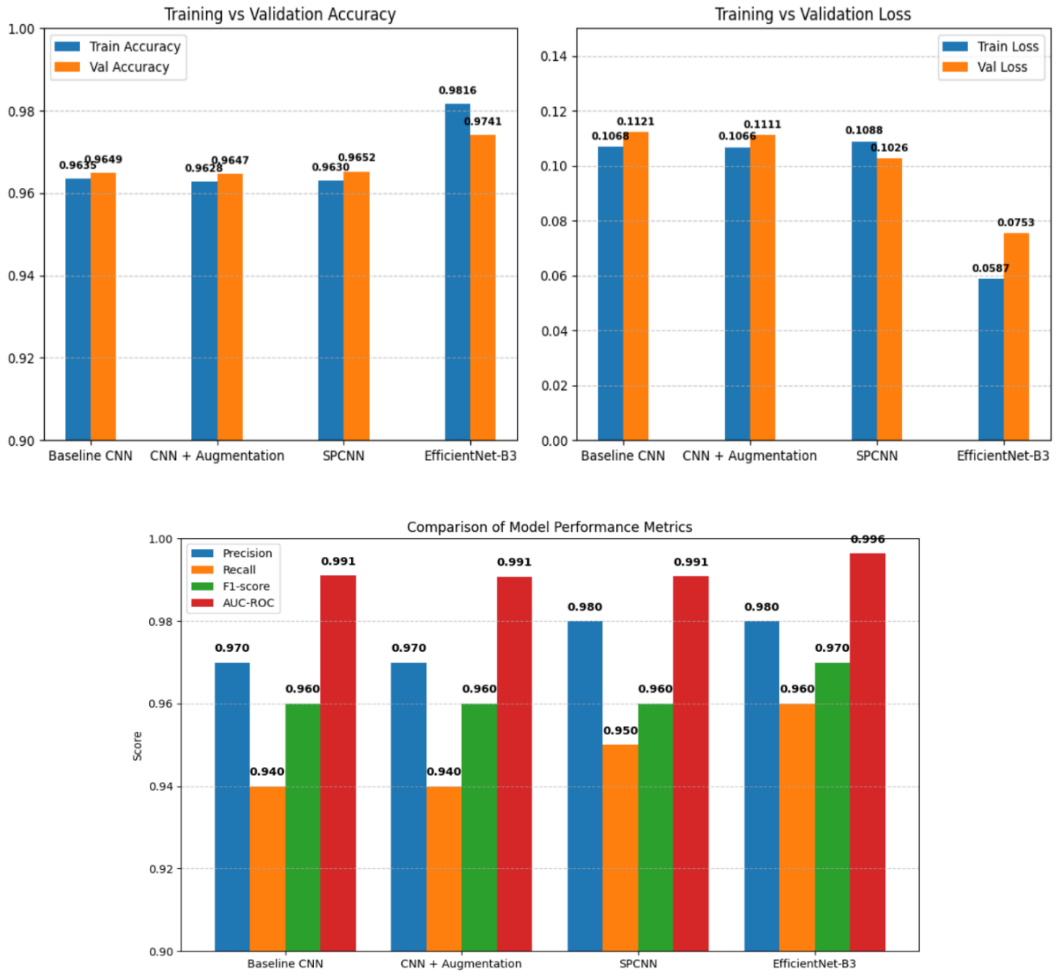


Figure 13. Summarizes the performance metrics for all four experiments.

Table 1. Summary of the performance metrics for all four experiments.

Metric/Setting	Exp1‡	Exp2‡	Exp3†‡	Exp4‡
Init	Rand	Rand	Rand	ImageNet
Input size (px)	256×256	256×256	256×256	300×300
Augment (train)	None	RBC+CLAHE+SSR	None	None
Optimizer (lr)	Adam(1e-3)	Adam(1e-3)	Adam(1e-3)	Adam(1e-3)
Loss	BCEwLogits	BCEwLogits	BCEwLogits	BCEwLogits
Early stopping	No	No	Yes	Yes(p5)
Train Acc (%)	96.35	96.28	96.3	98.16
Train Loss	0.1068	0.1066	0.1088	0.0587
Val Acc (%)	96.49	96.47	96.52	97.41
Val Loss	0.1121	0.1111	0.1026	0.0753
Test Precision (%)	97	97	98	98
Test Recall (%)	94	94	95	96
F1 Score	0.96	0.96	0.96	0.97
Test AUC-ROC	0.991	0.9908	0.9909	0.9964

[DOI: 10.61882/ijbc.17.3.62]

consistent with findings from prior studies emphasizing the value of pretrained ImageNet backbones [14,19].

EfficientNet-B3 likely performed best because of several factors. First, pretrained networks capture strong low- and mid-level features that adapt well to microscopy tasks, particularly when labeled data is limited [14,19]. Second, the compound scaling and architectural design of EfficientNet variants create an efficient balance between representational capacity and parameter efficiency, resulting in lower training loss and higher discriminative power when fully fine-tuned, as seen in the significantly lower training and validation losses for EfficientNet-B3 (train loss 0.0587; val loss 0.0753) compared to the custom models. Third, the EfficientNet experiment utilized a larger input resolution (300×300) and full fine-tuning of all layers, which likely preserved and adapted richer spatial features relevant to parasite detection morphology.

The SPCNN showed modest improvements in precision 98% and recall 95% compared to the baseline CNN for 97% and 94%, indicating that parallel multi-scale streams combined with spatial attention can improve localization of parasite-relevant regions and reduce background variability. This aligns with recent reports that attention and multi-scale processing enhance the detection of small, localized structures in microscopy images [18]. However, the overall improvements were incremental rather than transformative on this dataset, suggesting that attention-based architectural complexity may mainly offer benefits in cases where parasite appearances are highly variable or when background clutter is severe.

The augmentation experiment—in which CLAHE, moderate brightness and contrast adjustments, and small affine transformations were applied to 20% of the training samples from each class and then added back into the original dataset [17]—had a negligible impact on performance (validation accuracy 0.9647 versus 0.9649 for the baseline), indicating that this specific augmentation approach did not significantly enhance generalization for this dataset and training protocol. Possible reasons include (a) the original dataset's relative homogeneity after resizing and normalization, so the chosen transformations added little new, task-relevant variation; (b) the limited percentage of samples augmented (20%), which may have been too small to alter the learned decision boundary; or (c) that the augmentations used were not the most effective perturbations for highlighting discriminative morphological features in these cell crops. This finding is consistent with augmentation literature showing that improvements depend on the dataset and transformations used, and that inappropriate or insufficient augmentation can fail to improve—and sometimes even harm—performance [17].

Several practical and methodological implications follow. First, for malaria thin-smear classification on similar curated datasets, transfer learning with a well-chosen modern backbone (such as the EfficientNet family or its successors) serves as a strong baseline and may eliminate the need for extensive architecture engineering in many cases [14,19]. Second, attention-based or parallel architectures (e.g., SPCNN) remain valuable, especially when localization, interpretability, or robustness to diverse imaging conditions is required; such models could be prioritized for deployment on varied field microscopes or when integrating explainability modules is desired [14,18]. Third, augmentation strategies need careful tuning (types, magnitudes, and proportion of augmented samples) and should be validated empirically rather than assumed to be beneficial [17].

The limitations of this study restrict how broadly we can draw conclusions. All experiments used a single publicly available dataset (derived from NIH images and hosted on Kaggle) that, while common for benchmarking, might not reflect the full range of field image variability (such as different staining protocols, microscopes, camera sensors, and slide preparation artifacts). Therefore, external validation with independent datasets and prospective testing on field-collected slides are crucial before using these models in clinical or point-of-care settings [32]. Because the main goal of this work was to compare various deep learning architectures and identify the most promising model, we did not perform formal statistical significance tests like the DeLong test for AUC differences or McNemar/paired bootstrap tests for classification metrics. Our purpose was not to deploy the models in a clinical setting but to set a performance baseline for future research. For future studies targeting clinical use or regulatory approval, we recommend supplementing standard performance metrics (such as accuracy, F1-score, loss curves, confusion matrices, and ROC analysis) with formal statistical comparisons, including paired tests and appropriate corrections for multiple comparisons, to confidently validate performance claims.

Finally, computational cost and latency—crucial for low-resource deployments—were not thoroughly examined here; although EfficientNet-B3 achieved the highest accuracy, lightweight reparameterized models (such as RepVGG) or distilled variants might provide better trade-offs for mobile or embedded inference. Future work

[DOI: 10.61882/ijbc.17.3.62]

should therefore (1) validate top-performing models on external and prospectively collected microscopy images, (2) systematically explore augmentation strategies (including stronger mixes like CutMix/AugMix and domain-specific perturbations), (3) investigate model calibration and uncertainty quantification for safer clinical decision support, (4) evaluate inference latency and memory usage on target edge devices, and (5) incorporate interpretability methods (attention maps, gradient-based saliency) to boost clinician trust and support error analysis. Comparing EfficientNet-based models with newer backbones (such as EfficientNetV2, ConvNeXt, and hybrid Conv+Transformer models) and applying model compression or pruning techniques are additional promising directions to improve the balance between accuracy and deployability.

In conclusion, transfer learning with EfficientNet-B3 achieved the highest diagnostic accuracy, while SPCNN and custom CNNs delivered competitive results with small differences. The augmentation strategy tested here did not enhance generalization and was therefore excluded from later experiments. These results support using modern pretrained backbones as effective, practical tools for automated malaria microscopy, while highlighting the importance of careful validation, efficient deployment, and task-specific augmentation and architecture choices before clinical implementation.

6. CONCLUSION

We compared four convolutional network architectures for automated malaria parasite detection on segmented erythrocyte images: a baseline CNN, the same CNN with 20% targeted augmentation per class, a stacked parallel CNN with soft attention (SPCNN), and a fully fine-tuned EfficientNet-B3. In terms of accuracy, F1-score, and AUC, EfficientNet-B3 showed the best overall performance, while SPCNN achieved competitive results, and the augmentation protocol had little effect. These findings identify EfficientNet-B3 as a promising backbone for future malaria microscopy models and emphasize the importance of customizing augmentation strategies for dataset characteristics. The results establish a benchmark for future research focused on external validation and optimization for clinical or field use.

Acknowledgment

The authors acknowledge the financial support from Shahid Beheshti University of Medical Sciences.

Conflict of interest

The authors declared no conflict of interest.

Funding

This study was funded by Shahid Beheshti University of Medical Sciences.

Ethical statement

Not applicable.

References

1. World Health Organization. World Malaria Report 2024. Geneva: WHO; 2024.
2. Venkatesan P. WHO World Malaria Report 2024. Lancet Microbe. 2025;6(1).
3. Jones S, et al. Trends in Plasmodium burden in children and pregnant women in the WHO African Region. Lancet Glob Health. 2024.
4. RTS,S Clinical Trials Partnership. Efficacy and safety of RTS,S/AS01 malaria vaccine. Lancet. 2015;386(9988):31-45.
5. Datoo MS, et al. Efficacy of R21/Matrix-M malaria vaccine. Lancet. 2021;397(10287):1809–18.
6. Moody A. Rapid diagnostic tests for malaria parasites. Clin Microbiol Rev. 2002;15(1):66–78.
7. WHO. Malaria Microscopy Quality Assurance Manual, 2023.
8. Hopkins H, et al. Microscopy in malaria diagnosis. Malaria J. 2007;6:115.
9. Gamboa D, et al. A large proportion of P. falciparum in Peru lack pfhrp2 and pfhrp3. J Clin Microbiol. 2010;48(6):2055–7.
10. Snounou G. PCR diagnosis of malaria. Clin Microbiol Rev. 1993;6(1):15–28.
11. Rajaraman S, et al. Transfer learning for malaria parasite detection in thin smear images. PeerJ. 2018;6:e4568.
12. Pattanaik D, et al. Comparison of CNN frameworks for malaria diagnosis. arXiv preprint arXiv:1909.02829.
13. Liang Z, et al. CNN-based parasite stage classification. Comput Biol Med. 2021;134:104524.
14. Ahuja S, et al. EfficientNet B3-based malaria parasite detection. Biomed Signal Process Control. 2020;62:102093.
15. Tan M, Le Q. EfficientNet: Rethinking model scaling. ICML. 2019.
16. Shorten C, Khoshgoftaar TM. A survey on image data augmentation. J Big Data. 2019;6:60.
17. Perez L, Wang J. Effectiveness of data augmentation. arXiv preprint arXiv:1712.04621.
18. Ahamed F, et al. SPCNN for malaria detection. Sci Rep. 2025;15:6484.
19. 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.
20. Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. Trans R Soc Trop Med Hyg. 2018;112(4):170-182.

[DOI: 10.61882/ijbc.17.3.62]

21. Kaggle. Cell Images for Malaria Detection dataset. Available at: https://www.kaggle.com/datasets/iarunava/cell-images-for-malaria. Accessed 2025.

22. Li X, Wang Y, Zhang J, et al. Attention-based parallel CNN architectures for interpretable malaria diagnosis. Sci Rep. 2024;14:12345.

23. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60.

24. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proc Int Conf Mach Learn. 2019:6105–6114.

25. (Dataset): https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria

26. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):60.

27. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

28. Ahamed F, et al. SPCNN vs transfer learning on malaria detection. Sci Rep. 2025;15:6484.

29. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. ICML. 2019.

30. Silva, R. R. et al. (2022). Malaria Parasite Detection using EfficientNet Models. Biomedical Signal Processing and Control, 74, 103557.

https://doi.org/10.1016/j.bspc.2021.103557

31. Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, 29(2), 102–127.

https://doi.org/10.1016/j.zemedi.2018.11.002

32. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. Radiol Artif Intell. 2022;4(1):e210064. doi:10.1148/ryai.210064

[DOI: 10.61882/ijbc.17.3.62]