Resolution-Aware Malaria Screening: Do Super-Resolved RBC Images Improve CNNs and Vision Transformers?

Igor Oliveira[1], **Arthur Negrão**[1], **Ederson N. F. G. Júnior**[1], **Guilherme Silva**[1], **Matheus Vieira**[1], **Pedro Silva**[2]

[1] Postgraduate Program in Computer Science – Federal University of Ouro Preto – Brazil
igor.machado@aluno.ufop.edu.br

[2]Computing Department – Federal University of Ouro Preto – Brazil
silvap@ufop.edu.br

Abstract. *Automated malaria screening from microscopy images can improve diagnostic scalability in resource-limited settings, but detecting infected red blood cells (RBCs) is challenging due to subtle morphological cues and resolution constraints. We study resolution-aware malaria classification on the NIH Malaria Dataset using EfficientNet-B0 and ViTs, evaluating transfer learning, optimization strategies, and ESRGAN-based super-resolution preprocessing. Results show that cosine decay and transfer learning are critical for robust performance, enabling EfficientNet-B0 to reach 96% accuracy and recall, while ESRGAN with ViT achieves 97% accuracy and 98% recall, matching SOTA results while reducing false negatives.*

1. Introduction

Malaria remains a persistent global health challenge, caused by *Plasmodium* parasites and transmitted by female *Anopheles* mosquitoes [Secretaria de Estado da Saúde do Pará 2023]. While effective therapies exist, clinical outcomes are tightly coupled to early and reliable diagnosis; delays can rapidly lead to severe complications and mortality [Shekar et al. 2020]. The practical bottleneck is not the absence of diagnostic tools, but the difficulty of delivering consistent, high-throughput screening in endemic settings where expert microscopists, laboratory infrastructure, and computational resources are limited.

Microscopy of stained blood smears is the clinical reference standard. However, this workflow is labor-intensive, subjective, and sensitive to operator expertise, often exhibiting substantial inter-observer variability [Reddy and Juliet 2019]. These limitations are amplified in real-world deployment scenarios: heterogeneous acquisition conditions (illumination, stain intensity, focus, and sensor characteristics), constrained throughput, and limited access to specialized personnel. As a result, there is strong motivation for computer-aided diagnosis systems that can scale screening while preserving clinically meaningful sensitivity.

Automated recognition of infected red blood cells (RBCs) is deceptively hard. The discriminative evidence typically appears as small, high-frequency intracellular structures whose visibility is strongly modulated by optical blur and sensor resolution. This creates a fundamental tension: models must be sensitive to subtle morphological cues, yet remain efficient enough for resource-constrained environments. In addition, training and evaluation can be confounded by "shortcut" signals (e.g., stain artifacts or acquisition biases),

which may inflate in-dataset performance while degrading robustness under distribution shift. These constraints motivate a design space that jointly considers (i) architectural inductive biases, (ii) optimization and transfer learning regimes, and (iii) input-resolution limitations as a first-class concern.

Most automated malaria classifiers on the NIH Malaria Dataset rely on transfer learning with deep CNN backbones such as ResNet50 and InceptionV3, typically reporting accuracies between 95% and 97% [Reddy and Juliet 2019, Minarno et al. 2023]. While these studies establish strong baselines, they largely remain within the CNN paradigm and rarely analyze architectural efficiency or compare different model families under controlled evaluation protocols. Other works investigate alternative training regimes (e.g., training from scratch, freezing, or fine-tuning) [Harahap et al. 2021, Shekar et al. 2020], but systematic comparisons between convolutional architectures and attention-based models remain limited, particularly under deployment constraints typical of endemic regions.

Pre-processing has also been explored as a way to improve recognition performance. Classical spatial filtering has shown benefits in some settings [Çinar and Yildirim 2020], highlighting the importance of input quality for microscopy-based diagnosis. More recently, neural super-resolution (SR) methods have emerged as a potential strategy to recover high-frequency details lost during image acquisition. However, adversarial SR approaches such as ESRGAN [Wang et al. 2018] may generate perceptually plausible textures that are not necessarily diagnostically faithful, and their impact on malaria RBC classification remains largely unexplored. In parallel, Vision Transformers have demonstrated strong performance in medical imaging tasks [Waseem Sabir et al. 2023], motivating investigation of their applicability to malaria microscopy, where discriminative features are small and often resolution-limited.

In this work, we revisit malaria screening through the lens of *resolution-aware* recognition. Using the NIH Malaria Dataset [Rajaraman et al. 2018] (a nearly balanced binary benchmark - 13,775 infected vs. 13,783 uninfected) we conduct a controlled comparison between two complementary model families: EfficientNet-B0 [Tan and Le 2019], a compute-efficient CNN that is widely adopted for edge-feasible vision, and a Vision Transformer (ViT) [Dosovitskiy et al. 2020], which models global relationships via self-attention and has shown strong transfer behavior in medical imaging. Crucially, we evaluate neural super-resolution (SR) as a pre-processing module intended to recover high-frequency details that may be attenuated during acquisition, thereby increasing the signal available to downstream classifiers. We focus on ESRGAN [Wang et al. 2018], an adversarial SR approach designed to reconstruct perceptually plausible textures. To the best of our knowledge, ESRGAN has not been systematically investigated as a pre-processing component for malaria RBC classification; prior SR-enhanced pipelines in related settings more commonly report non-adversarial architectures such as RCAN, rather than GAN-based SR.

Our guiding hypothesis is that *efficient recognition models, combined with transfer learning and resolution enhancement, can improve clinically relevant sensitivity without sacrificing deployability*. We test this hypothesis under two evaluation protocols that address complementary goals: (i) a literature-aligned 70/15/15 split to enable direct comparability with prior work [Minarno et al. 2023], and (ii) stratified k-fold cross-validation

($k = 5$) to quantify robustness across partitions while preserving class balance. Across experiments we emphasize not only accuracy but also infected-class recall, reflecting the disproportionate clinical cost of false negatives in screening workflows.

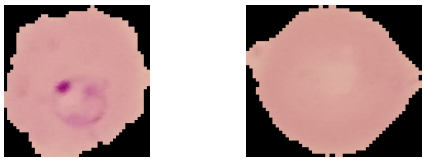
The contributions of this paper are: **(i) Resolution-aware malaria classification:** We evaluate the impact of spatial resolution on RBC classification and evaluate super-resolution within the recognition pipeline. **(ii) CNN–Transformer comparison:** We compare EfficientNet-B0 and ViT under matched protocols to analyze accuracy–efficiency trade-offs. **(iii) ESRGAN evaluation.** We investigate ESRGAN as a preprocessing step for malaria RBC classification. **(iv) Comparison with literature:** We report accuracy, precision, recall, and F1-score using stratified cross-validation and a standard 70/15/15 split, enabling direct comparison with prior work [Minarno et al. 2023].

2. Materials and Methods

This section describes the dataset and the methodology used for automated malaria classification. Our goal is to evaluate EfficientNet-B0 and a Vision Transformer under different training strategies and pre-processing choices, enabling controlled comparisons with prior work.

2.1. Dataset

We use the NIH Malaria Dataset [Rajaraman et al. 2018], which contains 27,558 segmented RBC images acquired from smartphone-attached microscopes: 13,775 infected and 13,783 uninfected (i.e., an effectively balanced binary classification benchmark). Cells were segmented by [Rajaraman et al. 2018] using a level-set method initialized with a Laplacian of Gaussian (LoG) filter, followed by morphological post-processing to remove segmentation artifacts.



(a) infected cell (b) not infected cell

Figure 1. Example of infected and not infected cell.

All images were loaded via *TensorFlow Datasets*, resized, and normalized to $[0, 1]$. We used 32×32 inputs for EfficientNet-B0 and 224×224 inputs for ViT. Binary labels were preserved without re-encoding.

2.2. Models and Resolution Enhancement

We used two models: EfficientNet-B0, a CNN-based architecture, and Vision Transformer. The EfficientNet-B0 [Tan and Le 2019] is initialized from ImageNet weights (transfer learning) and randomly (training from scratch). We replace the classifier head with *GlobalAveragePooling2D* followed by three dense layers with 64, 32, and 32 neurons respectively, all three with ReLU, *BatchNormalization*, 0.3 *Dropout*, and L_2 regularization, and a final sigmoid unit.

For the Vision Transformer, we evaluate the ViT-base-patch16-224-in21k model [Hugging Face 2020], which tokenizes the image into 16×16 patches and applies stacked self-attention blocks. We only adapted the final head for binary classification. Data splits, metrics, and reporting are matched to EfficientNet-B0 for comparability.

We assess two complementary strategies: (i) *data augmentation* (horizontal and vertical flips, rotations up to 90 degrees, and height/width/zoom shift range up to 0.2) and (ii) super-resolution using ESRGAN [Wang et al. 2018] as the ViT model needs a 224×224 input image. ESRGAN is applied as a deterministic pre-processing step to enhance high-frequency content; super-resolved images are used consistently for both training and testing to isolate the effect of input resolution enhancement on downstream recognition.

2.3. Evaluation

We evaluate under two protocols. First, we follow the common 70/15/15 train/validation/test split to enable direct comparison with prior NIH Malaria reports [Minarno et al. 2023]. Second, we perform stratified k-fold cross-validation with $k = 5$ to quantify robustness across partitions while maintaining class balance within each fold.

To compare the approaches, we report accuracy, precision, recall, and F1-score. In screening contexts, infected-class recall is emphasized due to the high clinical cost of false negatives.

3. Experiments and Results

We evaluate EfficientNet-B0 and ViT under ablations that target three axes: (i) transfer learning vs. training from scratch, (ii) optimization stability via cosine decay scheduling, and (iii) input transformations via data augmentation and super-resolution. We report results for both the literature-aligned 70/15/15 split [Minarno et al. 2023] and stratified 5-fold cross-validation.

Experiments were run on Google Colab (cross-validation and transfer-learning comparisons) and on a local workstation (AMD Ryzen Threadripper 3960X, 128 GB RAM, RTX 3090 24 GB) for ViT training and SR/augmentation ablations. EfficientNet-B0 was implemented in *TensorFlow/Keras*, optimized with Adam and binary cross-entropy.

The ViT (google/vit-base-patch16-224-in21k) [Hugging Face 2020] was trained using the Hugging Face *Trainer* with a batch size of 32 for 20 epochs, FP16, and epoch-level evaluation. We tested the learning rate of 10^{-4}. In our current configuration, ViT is evaluated only on SR-processed inputs, as it needs an input of 224×224.

3.1. Results and Analysis

Cross-validation exposes optimization instability: cosine decay fixes it. With transfer learning but a fixed learning rate (0.01), EfficientNet-B0 shows large fold-to-fold variance with an accuracy of 84.6% $\pm$ 7.64, indicating brittle optimization and limited robustness to partition difficulty. Switching to cosine decay yields both higher and more stable performance with an accuracy of 94.4% $\pm$ 2.07, with mean precision/recall/F1 around 94% across classes.

Transfer learning is not optional under robustness evaluation. Under cosine decay, removing transfer learning reduces mean accuracy to 81.2% $\pm$ 15.93 and markedly degrades infected-class recall (mean 65.4% $\pm$ 34,72). In the worst fold, infected recall collapses to 19%, implying an unacceptable false-negative rate for screening. These results indicate that ImageNet initialization is critical not only for peak performance but also for reliability under cross-validated evaluation.

Ablations under the 70/15/15 split: high accuracy can mask clinically relevant failures. Following [Minarno et al. 2023], EfficientNet-B0 trained from scratch reaches 83% accuracy and 70% recall, which is inadequate for clinical screening despite moderate accuracy. With transfer learning, EfficientNet-B0 reaches 96% accuracy and 96% recall, substantially reducing false negatives.

Data augmentation and ESRGAN do not improve the CNN baseline in this regime. The use of the proposed data augmentation yields no gains: metrics remain 0.96 without augmentation and slightly drop to $\approx$ 95% with augmentation. ESRGAN-based SR [Wang et al. 2018] likewise produces no measurable improvement for EfficientNet-B0 across optimizers (all metrics are around 96%), suggesting that (i) the native NIH resolution already preserves sufficient information for this CNN, or (ii) ESRGAN's reconstructed textures are not aligned with the discriminative cues needed for parasite detection in this dataset.

ViT achieves a state-of-the-art-level performance with improved recall. ViT trained on SR inputs achieves 97% accuracy and 98% recall at 10^{-4}. Compared to [Minarno et al. 2023], ViT matches accuracy (97%) while improving recall (98% vs. 97%), indicating fewer false negatives. EfficientNet-B0 remains competitive (96%) at a lower computational cost, making it an attractive option for constrained deployment. Compared to [Reddy and Juliet 2019], our models exceed the reported accuracy (95%), though that comparison is limited by missing precision, recall, and F1.

Results analysis: Overall, the results support two practical takeaways: (i) optimization and initialization choices (cosine decay, transfer learning) dominate performance and robustness for EfficientNet-B0, and (ii) ViT can improve infected-case sensitivity, but it comes with higher computational requirements and a need for careful input preprocessing.

4. Conclusion

We investigate automated malaria RBC classification on the NIH Malaria Dataset, focusing on deployable accuracy and clinically meaningful sensitivity. Our study analyzes the impact of architectural choice (EfficientNet-B0 vs. Vision Transformer), transfer learning, optimization strategies, and super-resolution preprocessing. Experiments show that optimization and initialization are critical for robustness: cosine decay stabilizes training across folds, and transfer learning significantly reduces false negatives. Under the standard 70/15/15 evaluation protocol, EfficientNet-B0 with ImageNet initialization achieves 96% strong and computationally efficient baseline for deployment in constrained environments.

We also evaluate ESRGAN-based super-resolution and standard data augmentation. Neither improves EfficientNet-B0 performance, suggesting that the native dataset

resolution already preserves the relevant discriminative cues for CNN-based classification. In contrast, a ViT trained on super-resolved inputs achieves 97% accuracy and 98% recall, matching state-of-the-art results while further reducing false negatives, albeit at higher computational cost. Future work will explore domain-adapted super-resolution for microscopy, robustness under acquisition variability, and more efficient Transformer or hybrid CNN–Transformer architectures for practical deployment in endemic regions.

Acknowledgments

The authors acknowledge the support of the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, project APQ-01768-24), the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), the C*oordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), the *Universidade Federal de Ouro Preto* (PROPPI/UFOP), and its Graduate Program in Computer Science (PPGCC/UFOP).

References

Çinar, A. and Yildirim, M. (2020). Classification of malaria cell images with deep learning architectures. *Ingénierie des Systèmes d'Information*, 25(1):35.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Harahap, M., Jefferson, J., Barti, S., Samosir, S., and Turnip, C. A. (2021). Implementation of convolutional neural network in the classification of red blood cells have affected of malaria. *Sinkron: jurnal dan penelitian teknik informatika*, 5(2):199–207.

Hugging Face (2020). vit-base-patch16-224-in21k. Accessed on: March, 2026.

Minarno, A. E., Aripa, L., Azhar, Y., and Munarko, Y. (2023). Classification of malaria cell image using inception-v3 architecture. *JOIV: International Journal on Informatics Visualization*, 7(2):273–278.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568.

Reddy, A. S. B. and Juliet, D. S. (2019). Transfer learning with resnet-50 for malaria cell-image classification. In *2019 International conference on communication and signal processing (ICCSP)*, pages 0945–0949. IEEE.

Secretaria de Estado da Saúde do Pará (2023). O que é a malária? `http://www.saude.pa.gov.br/a-secretaria/diretorias/dvs/malaria/o-que-e-malaria/`. Accessed on: January, 2026.

Shekar, G., Revathy, S., and Goud, E. K. (2020). Malaria detection using deep learning. In *2020 4th international conference on trends in electronics and informatics (ICOEI)(48184)*, pages 746–750. IEEE.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Waseem Sabir, M., Farhan, M., Almalki, N. S., Alnfiai, M. M., and Sampedro, G. A. (2023). Fibrovit—vision transformer-based framework for detection and classification of pulmonary fibrosis from chest ct images. *Frontiers in medicine*, 10:1282200.