NOUN Interdisciplinary Journal of Computing, E-learning & Application (NOUN-IJCEA)
https://ijcea.nou.edu.ng
Volume 1 Number 1, April 2026 pp. 52-62

Optimized Deep Learning Approaches for Malaria Cell Detection

Muhammad Bishir Ruma, Eli Adama Jiya, *Abubakar Saidu, Ahmed Ibrahim Mahmud

Department of Computer Science, Federal University Dutsin-Ma, Katsina State Nigeria.
bishirrumamuhammad@gmail.com, a.saidu01@fudutsinma.edu.ng, a.jiyaeli@fudutsinma.edu.ng, amidot2005@gmail.com
*Corresponding Author: a.saidu01@fudutsinma.edu.ng

ABSTRACT

Malaria remains a leading cause of death in sub-Saharan Africa, with Nigeria accounting for approximately 27% of global malaria deaths. Accurate diagnosis is frequently compromised in rural healthcare facilities due to shortages of trained microscopists and limited access to diagnostic equipment. This study presents a comparative evaluation of three convolutional neural network (CNN) architectures, a Baseline CNN trained from scratch, ResNet50, and MobileNetV2 for automated malaria parasite detection in blood smear images. Models were trained and evaluated on the publicly available Malaria Cell Image Dataset (27,558 segmented cell images) using standardized preprocessing, data augmentation, and training protocols. Performance was assessed through diagnostic metrics (accuracy, sensitivity, specificity, precision, F1 score, AUC-ROC) and deployment feasibility metrics (model size, CPU inference latency). MobileNetV2 achieved 93.52% accuracy and 94.87% sensitivity, beating ResNet50 across all diagnostic metrics while requiring 90.5% fewer parameters (2.2M vs. 23.5M) and achieving 3.4× faster CPU inference (22.0 ms vs. 74.3 ms). The Baseline CNN achieved the highest raw accuracy (95.77%) but requires domain-specific training from scratch, limiting practical deployment in low-data settings. These findings establish MobileNetV2 as the optimal architecture for malaria detection under rural Nigerian infrastructure constraints, demonstrating that deployment feasibility must be prioritized alongside diagnostic accuracy in global health AI applications.

Keywords: Convolutional Neural Networks (CNN), Malaria detection, MobileNetv2, Medical image classification, Resource-limited healthcare, Transfer learning

1. INTRODUCTION

Malaria remains one of the most significant global infectious diseases, posing a persistent public health challenge, particularly in sub-Saharan Africa. According to the World Health Organization, there were an estimated 249 million cases and 608,000 deaths worldwide in 2022 (World Health Organization [WHO], 2023). Nigeria accounts for a disproportionate share of this burden, contributing approximately 27% of global malaria-related deaths (WHO, 2023). The impact is especially severe in rural regions, where limited healthcare infrastructure, shortages of skilled laboratory personnel, and inadequate access to diagnostic tools significantly hinder effective disease management and control.

Accurate and timely diagnosis is central to malaria control strategies. The gold standard diagnostic method microscopic examination of Giemsa-stained blood smears enables parasite detection, species differentiation, and quantification of parasitemia. However, its effectiveness is constrained in low-

resource settings due to its reliance on highly trained microscopists, substantial time requirements, and susceptibility to inter-observer variability. Previous studies report error rates of up to 30–40% when microscopy is performed by non-specialist personnel (Ochola et al., 2006). Although rapid diagnostic tests (RDTs) provide a more accessible alternative, they are limited by reduced sensitivity in low parasitemia cases, inability to quantify parasite load, and vulnerability to antigenic variability (Wongsrichanalai et al., 2007).

Recent advances in machine learning, particularly in deep learning, have introduced promising alternatives for automated malaria diagnosis. Convolutional neural networks (CNNs) have demonstrated high accuracy in medical image analysis, including the classification of malaria-infected erythrocytes from digitized blood smear images (Poostchi et al., 2018; Rajaraman et al., 2018; Liang et al., 2017; Rajaraman et al., 2019). These approaches offer the potential to reduce reliance on expert interpretation, enhance diagnostic consistency, and significantly decrease turnaround time.

Despite these advances, the practical deployment of deep learning–based diagnostic systems in rural African settings remains limited. A critical gap persists between laboratory-level performance and real-world applicability. Key challenges include the high computational demands of state-of-the-art models, which are often incompatible with low-cost hardware; the need for reliable internet connectivity for cloud-based inference; limited evaluation using deployment-relevant metrics such as latency and model size; and insufficient validation on locally representative datasets (Poostchi et al., 2018; Rajaraman et al., 2019). Addressing these constraints is essential for translating technological advances into tangible healthcare impact in resource-constrained environments.

In this study, we address these challenges through a systematic comparative evaluation of three convolutional neural network architectures, explicitly considering both diagnostic accuracy and deployment feasibility. We examine a shallow baseline CNN trained from first principles, ResNet50 (He et al., 2016) as a high-performance benchmark, and MobileNetV2 (Sandler et al., 2018) as a lightweight architecture optimized for mobile and embedded systems. Beyond conventional performance metrics, our evaluation incorporates model size, CPU inference latency, and storage requirements to reflect real-world deployment constraints. This approach enables the identification of models that not only achieve high diagnostic performance but are also practically deployable within the infrastructural limitations of rural Nigerian healthcare settings.

2. MATERIALS AND METHODS

2.1 Study Design

This study employed a quantitative, experimental design comparing three CNN architectures under controlled conditions to determine the optimal balance between diagnostic accuracy and computational efficiency for malaria cell detection in resource-constrained healthcare settings. The independent variable was model architecture; dependent variables included diagnostic performance metrics (accuracy, sensitivity, specificity, precision, F1 score, AUC-ROC) and efficiency metrics (parameter count, model size, CPU inference latency). Controlled variables comprised dataset composition, preprocessing protocols, data augmentation, training configuration, and evaluation procedures.

2.2 Dataset and Image Augmentation

The Malaria Cell Image Dataset, publicly available through Kaggle, served as the sole data source. This dataset comprises 27,558 segmented cell images derived from Giemsa-stained thin blood smear slides captured under light microscopy. Images are equally distributed across two classes: 13,779 images of Plasmodium falciparum-infected erythrocytes (parasitized) and 13,779 images of uninfected erythrocytes. The balanced class distribution obviated the need for class weighting or oversampling techniques. The dataset was partitioned using stratified random sampling to preserve

class balance across splits: 70% for training (19,290 images), 15% for validation (4,134 images), and 15% for testing (4,134 images). The test set was withheld from all model development activities and used exclusively for final performance evaluation.

All images were resized to 128 × 128 pixels using bilinear interpolation and normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406]; standard deviation = [0.229, 0.224, 0.225]) to align with pretrained weight distributions. The 128 × 128 input resolution was selected to balance feature preservation with computational efficiency. Stochastic data augmentation was applied to the training set only: random horizontal and vertical flips (probability = 0.5), random rotation within ±15 degrees, and random color jitter (brightness and contrast factors = ±0.2). These transformations simulated realistic variation in microscope image acquisition without distorting biological features. No augmentation was applied during validation or testing.

2.3 Methodology Diagram

The following section provides an overview of the methodological workflow adopted in this study. The goal of the pipeline is to ensure a systematic and reproducible approach for developing and evaluating deep learning models for malaria parasite detection in microscopic blood smear images. The workflow (Figure1) integrates all key stages of the machine learning process, from raw data acquisition to final model selection, while maintaining consistency across all experimental conditions to ensure a fair comparison between models.

The process begins with the use of a publicly available malaria cell image dataset consisting of 27,558 Giemsa-stained microscopic images, equally distributed between parasitized and uninfected red blood cells. This balanced distribution ensures that model training is not biased toward either class. The dataset undergoes preprocessing steps, including image resizing to a uniform resolution of 128×128 pixels and normalization using standard ImageNet statistical values to align input distributions with pretrained deep learning models.

To improve model generalization and reduce overfitting, data augmentation is applied exclusively to the training set. These transformations simulate realistic variations in microscope imaging conditions through random horizontal and vertical flipping, slight rotations, and controlled adjustments in brightness and contrast. The dataset is then split into training, validation, and testing subsets in a stratified manner to preserve class balance across all partitions.

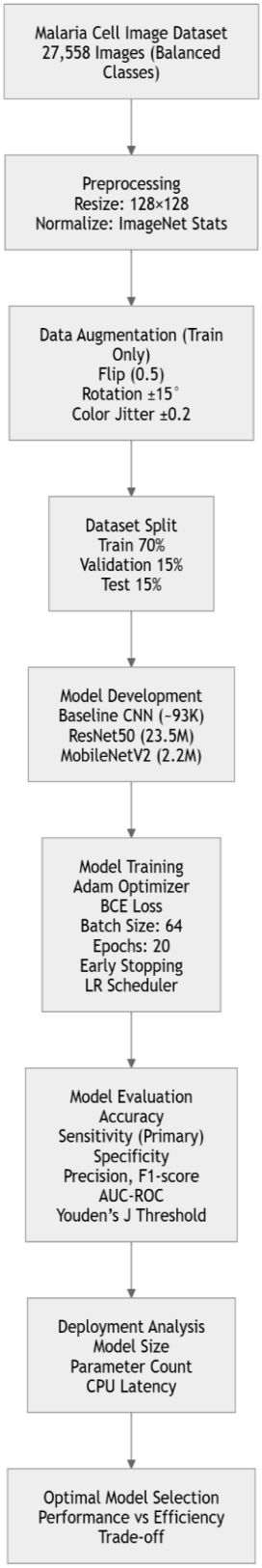

Figure 1: Methodological workflow

2.4 Model Architectures

Baseline CNN: A shallow custom architecture comprising three convolutional blocks, each containing Conv2D, batch normalization, ReLU activation, and max pooling (2 × 2). This was

followed by global average pooling, dropout (rate = 0.3), and a single sigmoid output node. The model contained approximately 93,000 parameters and was trained entirely from scratch.

ResNet50: A 50-layer residual network with 23.5 million parameters, serving as the high-accuracy benchmark. The model was initialized with ImageNet pretrained weights, and the original classification head was replaced with a single sigmoid output node. Training employed a two-phase strategy: Phase 1 (epochs 1–5) with the convolutional base frozen, training only the new classification head; Phase 2 (epochs 6–20) with the top residual block unfrozen for fine-tuning at a reduced learning rate.

MobileNetV2: A lightweight architecture built on inverted residual blocks with linear bottlenecks, containing 2.2 million parameters and specifically designed for mobile and embedded inference. Initialized with ImageNet pretrained weights, MobileNetV2 was fine-tuned using the same two-phase strategy as ResNet50.

2.5 Training Configuration

All three models were trained under identical hyperparameters to ensure fair comparison (Table 1). The Adam optimizer was used with binary cross-entropy loss. A learning rate scheduler (ReduceLROnPlateau, factor = 0.5, patience = 3) reduced the learning rate when validation loss plateaued. Early stopping (patience = 5) monitored validation loss and restored the best-performing checkpoint for final evaluation. Table 1 contains the model implementation parameters.

Table 1: Training configuration applied uniformly across all models

Hyperparameter	**Value**
Loss function	Binary cross-entropy (BCEWithLogitsLoss)
Learning rate- Phase 1	1×10^{-3}
Learning rate — Phase 2	1×10^{-5}
Batch size	64
Maximum epochs	20
Early stopping patience	5 epochs
LR scheduler	ReduceLROnPlateau (factor = 0.5, patience = 3)
Random seed	42
Optimiser	Adam

2.6 Evaluation Metrics

The models were evaluated using Accuracy, sensitivity (recall / true positive rate), specificity (true negative rate), precision, F1 score, and area under the receiver operating characteristic curve (AUC-ROC) were computed on the held-out test set. Sensitivity was prioritized as the primary clinical metric, as false negatives (missed infections) carry greater clinical consequence than false positives in malaria screening contexts. Optimal classification thresholds were determined using Youden's J statistic (maximizing sensitivity + specificity − 1)

3. RESULTS AND DISCUSSION

All three models were trained for a maximum of 20 epochs with early stopping (patience = 5). MobileNetV2 was the only model to trigger early stopping, converging at epoch 13. Both the Baseline CNN and ResNet50 completed the full 20 epochs. Training summary statistics are presented in Table 2.

Table 2: Training summary statistics for all three models

Model	Epochs	Best Epoch	Best Val. Loss	Best Val. Acc.	Early Stop
Baseline CNN	20	17	0.1430	95.31%	No
ResNet50	20	19	0.2098	92.24%	No
MobileNetV2	13	8	0.1992	92.94%	Yes

The Baseline CNN exhibited unstable validation loss at epoch 3 (0.8667) before recovering, attributable to random weight initialization. Both pretrained models displayed smoother loss trajectories from epoch 1, consistent with the benefit of ImageNet initialization. MobileNetV2 converged notably faster than ResNet50, reaching its best validation loss at epoch 8 with a near-zero overfitting gap of −0.02%. ResNet50's overfitting gap was 0.75% and the Baseline CNN's was 1.06%. Figure2 presents the training and validation loss curves for all three models while Figure2 presents the training and validation accuracy curves across epochs

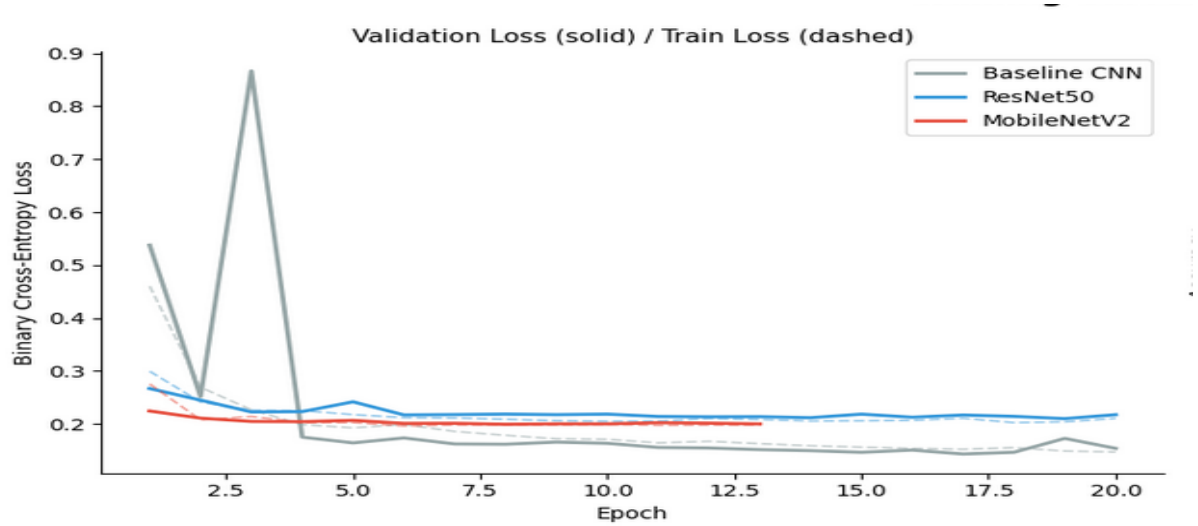

Figure 2: Training and validation loss curves for all the three models

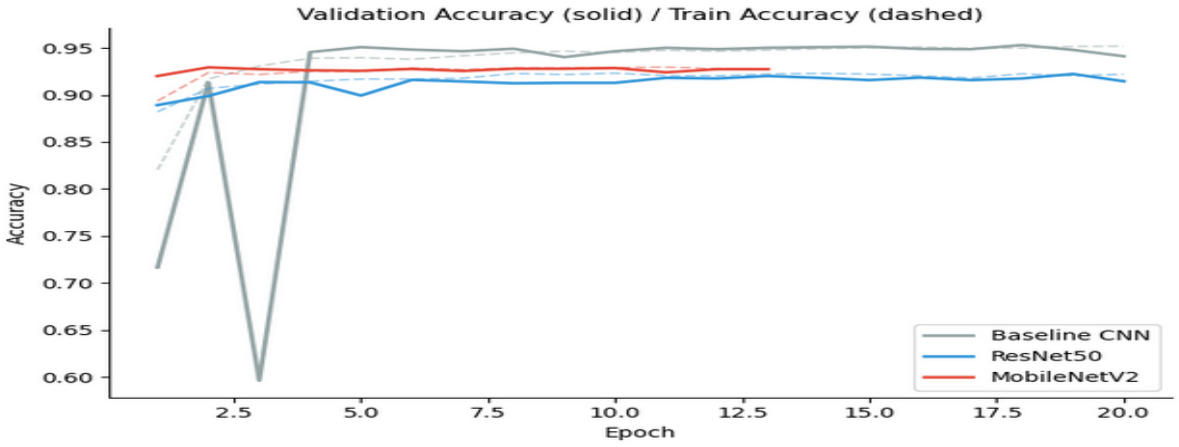

Figure 3: Training and validation accuracy curves for all three models across epochs

In Figure 2, MobileNetV2 Converges at epoch 8 with the smallest overfitting gap, while ResNet50 and Baseline CNN require the 20 full epoch while in Figure 3, Baseline CNN attain the heights final Accuracy, but MobileNetV2 achieves comparable performances with substantially faster convergence. Table 3 presents the full diagnostic performance of all three models on the 4,134-image held-out test set at the default 0.5 classification threshold.

Table 3: Test-set diagnostic performance for all three models

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC-ROC
Baseline CNN	95.77%	98.42%	93.12%	93.60%	95.96%	0.9917
ResNet50	92.79%	93.37%	92.21%	92.28%	92.83%	0.9776
MobileNetV2	93.52%	94.87%	92.16%	92.37%	93.60%	0.9794

In Table 2, the Baseline CNN achieved the highest scores across all six metrics. MobileNetV2 outperformed ResNet50 on every metric: accuracy (93.52% vs. 92.79%), sensitivity (94.87% vs. 93.37%), F1 score (93.60% vs. 92.83%), and AUC-ROC (0.9794 vs. 0.9776). In terms of clinically critical false negatives, MobileNetV2 produced 106 false negatives from 4,134 test samples (false negative rate: 5.13%), compared to ResNet50's 137 false negatives (6.63%).
The Figure 4 provides the confusion matrix of the all models. These matrices visualize the distribution of true positive (TP) true negative (TN), false positive (FP), and false negative (FN), providing insight into each models's classification behaviour beyond aggregate matrices.

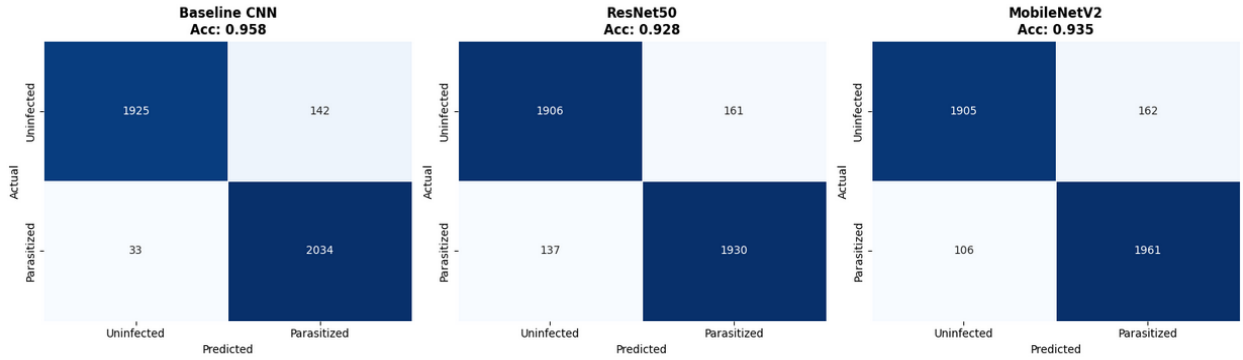

Figure 4: Confusion matrices for all three models

From the Figure 4, the Baseline CNN achieves the highest overall accuracy (95.77%) with lowest number of Misclassification. It correctly identifies 2,034 of 2,067 infected cells (TP) and, 1925 of 2,067 uninfected cells (TN). It Produce 37 false negative (FN), and 142 false Positive (FP) the low FN count (1.6% of infected samples) is critical for a screening tool as missed infection carry greater critical risk than false alarms. However, the relatively higher FP rate (6.9%) may lead to unnecessary confirmatory testing but is acceptable given the screening context.
The Figure 5 provides the result of the ROC curves for all three models on test set. The result shows that CNN achieve the highest overall value of 0.99 (99%) with other model performing at 97% level

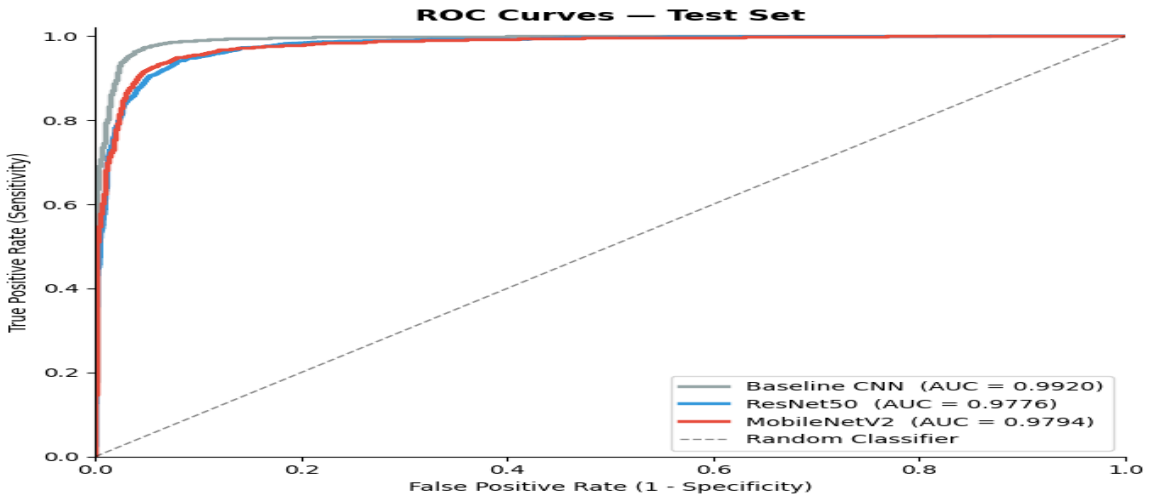

Figure 5: ROC curves for all three models on test set

Optimal threshold analysis using Youden's J statistic (Table 4) reveals that MobileNetV2's optimal threshold (0.6264) yields 92.60% sensitivity and 94.58% specificity, offering a more balanced sensitivity–specificity tradeoff than the default 0.5 threshold.

Table 4: Optimal classification thresholds derived using Youden's J statistic

Model	Optimal Threshold	Sensitivity	Specificity	Method
Baseline CNN	0.5004	97.91%	93.55%	Youden's J
ResNet50	0.4864	93.95%	92.07%	Youden's J
MobileNetV2	0.6264	92.60%	94.58%	Youden's J

Efficiency and Deployment Analysis

Table 5 presents efficiency measurements for each model. CPU latency was measured as mean ± standard deviation over 50 single-image inference runs in a CPU-only environment, simulating the absence of a GPU on a budget Android handset.

Table 5: Model efficiency measurements and deployment feasibility

Model	Parameters	Size (MB)	CPU Latency (ms)	Deployable
Baseline CNN	93,825	0.37	15.7 ± 0.8	Yes
ResNet50	23,510,081	90.0	74.3 ± 4.0	Marginal
MobileNetV2	2,225,153	8.72	22.0 ± 0.6	Yes

MobileNetV2 achieved a 90.5% reduction in parameter count relative to ResNet50 (2.2M vs. 23.5M) and a 90.3% reduction in model file size (8.72 MB vs. 90.0 MB). Its CPU inference latency of 22.0 ± 0.6 ms was 3.4 times faster than ResNet50's 74.3 ± 4.0 ms. The low standard deviation in MobileNetV2's latency (±0.6 ms) compared to ResNet50 (±4.0 ms) indicates more consistent and predictable inference behaviour.

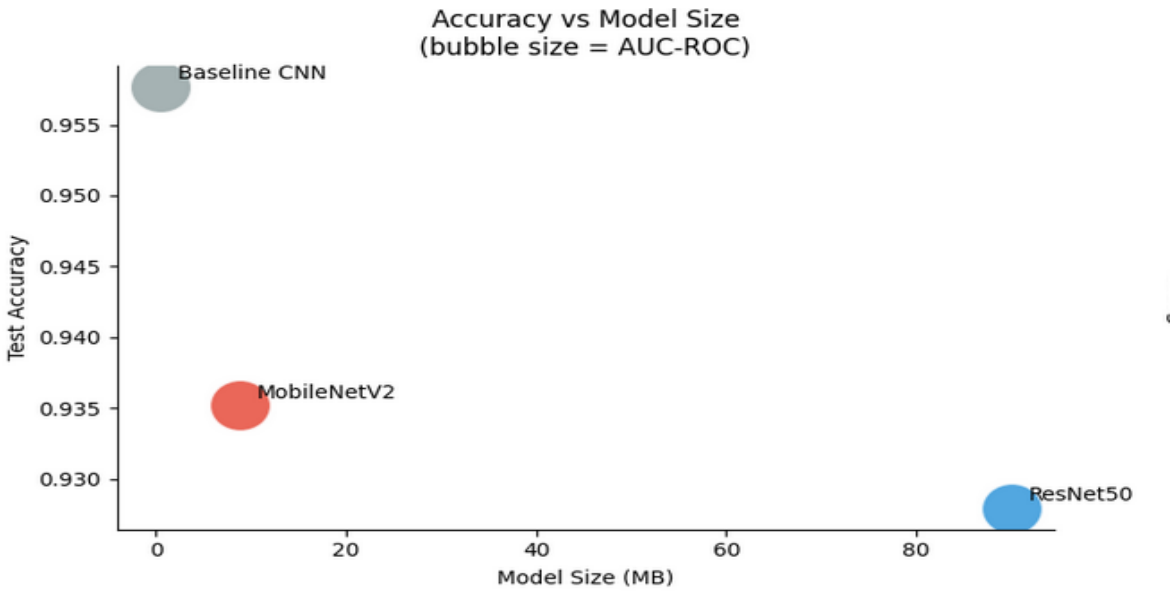

Figure 6: Accuracy versus model size (MB) bubble chart

ResNet50's 90 MB weight file is technically within the 200 MB storage criterion, a deployable Android application incorporating ResNet50 would require the model weights, application code, and runtime dependencies, likely totalling 150–200 MB of device storage. Furthermore, the 74.3 ms CPU latency measured under idealised server conditions would translate to substantially higher latency on a Qualcomm Snapdragon 460-class processor typical of budget handsets. By contrast, MobileNetV2 at 8.72 MB and 22.0 ms satisfies all deployment criteria with considerable headroom.

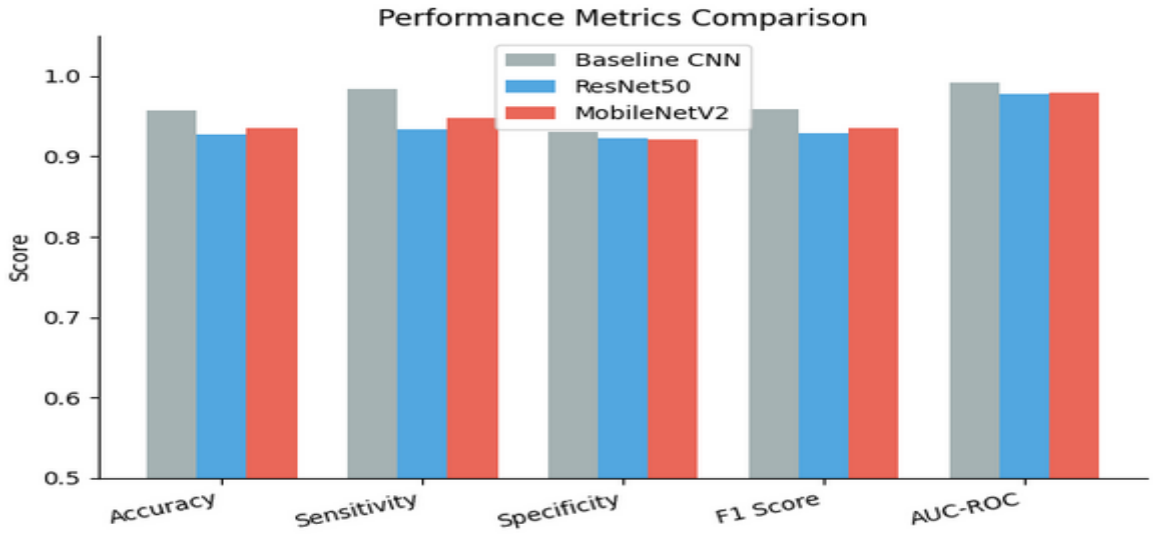

Figure 7: Grouped bar chart comparing all five diagnostic metrics across the three models

Discussion

This study evaluated three CNN architectures for malaria cell detection, explicitly incorporating both diagnostic performance and deployment feasibility within the infrastructure constraints of rural Nigerian clinics. Three principal findings emerge from this investigation.

Firstly, MobileNetV2 outperformed ResNet50 across all diagnostic metrics while demonstrating substantially superior efficiency. This finding challenges the assumption that larger, deeper networks universally yield higher accuracy. The NIH malaria dataset, comprising clean, pre-segmented,

balanced images of single cells, presents a binary classification task of comparatively low complexity. ResNet50's depth optimized for complex multi-class natural image recognition constitutes an architectural mismatch for this domain, resulting in higher validation loss throughout training and full utilization of the training budget without convergence. MobileNetV2's efficient inverted residual architecture extracted discriminative features parsimoniously, converging at epoch 13 with a near-zero overfitting gap. These results align with prior work demonstrating that lightweight architectures frequently match or exceed deeper networks on constrained medical image classification tasks.

Secondly, the Baseline CNN achieved the highest raw accuracy but offers limited practical utility for rural deployment. While the 95.77% accuracy and 98.42% sensitivity are technically impressive, this architecture requires training from scratch on a labeled domain-specific dataset of sufficient size. In the target deployment context, such a dataset would not be available, and the model offers no transfer learning pathway to adapt to new clinical settings with limited local data. This illustrates a critical distinction between laboratory performance and deployment readiness: a model's ability to be fine-tuned from pretrained weights with minimal local data is often more important than its benchmark accuracy under optimal conditions.

Thirdly, the deployment feasibility analysis establishes MobileNetV2 as the only high-accuracy model that satisfies all infrastructure constraints with substantial headroom. At 8.72 MB and 22.0 ms CPU latency, MobileNetV2 is well within the storage and responsiveness limits of budget Android handsets. The accuracy gap between MobileNetV2 and ResNet50 is 0.73 percentage points a difference that is not clinically significant for a screening tool. What is clinically significant is the false negative rate: MobileNetV2's 5.13% FNR must be benchmarked against the alternative of manual microscopy by non-specialist health workers, where error rates of 30–40% have been documented in similar low-resource settings [3]. Against this baseline, MobileNetV2 represents a substantial improvement in diagnostic reliability regardless of how it compares to ResNet50 in isolation.

These findings align with the growing body of evidence for efficiency-first model selection in resource-constrained AI for global health [9,10]. The core argument can be stated precisely: a model achieving 93.52% accuracy that runs on a health worker's existing Android phone is categorically more useful than a model achieving 92.79% accuracy that does not perform reliably under those conditions.

4. CONCLUSION

This study demonstrates that MobileNetV2 represents the optimal architecture for automated malaria cell detection under rural Nigerian deployment constraints. Its diagnostic performance is clinically equivalent to the ResNet50 benchmark, while its efficiency profile enables practical deployment on the budget Android hardware available to rural health workers. The broader contribution of this work is methodological: deployment feasibility must be treated as a first-class evaluation criterion in AI diagnostic tool development for global health applications, not as an afterthought applied after accuracy-based model selection.

With Nigeria accounting for 27% of global malaria deaths, a deployable, accurate, offline-capable diagnostic tool that runs on existing health worker devices represents a realistic pathway to reducing this burden. This study provides evidence that such a tool is technically achievable using MobileNetV2 and establishes an evaluation framework for similar global health AI applications

REFERENCES

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the

detection of diabetic retinopathy. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/33138331.3376718

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Liang, Z., Powell, A., Ersoy, I., Poostchi, M., Silamut, K., Palaniappan, K., Guo, P., Hossain, M. A., Maude, R. J., Huang, J. X., Thoma, G. R., & Antani, S. (2020). CNN-based image analysis for malaria diagnosis. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 493–496. https://doi.org/10.1109/BIBM49941.2020.9313298

Makler, M. T., Palmer, C. J., & Ager, A. L. (1998). A review of practical techniques for the diagnosis of malaria. *Annals of Tropical Medicine & Parasitology, 92*(4), 419–433. https://doi.org/10.1080/00034983.1998.11813300

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., & 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. https://doi.org/10.7717/peerj.4568

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4510–4520). https://doi.org/10.1109/CVPR.2018.00474

Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? *BMJ Global Health, 3*(4), e000798. https://doi.org/10.1136/bmjgh-2018-000798

Wongsrichanalai, C., Barcus, M. J., Muth, S., Sutamihardja, A., & Wernsdorfer, W. H. (2007). A review of malaria diagnostic tools: Microscopy and rapid diagnostic test (RDT). *The American Journal of Tropical Medicine and Hygiene, 77*(6 Suppl), 119–127. https://doi.org/10.4269/ajtmh.2007.77.119

World Health Organization. (2023). *World malaria report 2023*. World Health Organization.