Research Article

Multi-Model Hybrid Framework for Automated Malaria Detection and Species Classification

Akhiyar Waladi[1], Nindy Raisa Hanum[2], Yogi Perdana[3], Hasanatul Iftitah[4]
Fitra Wahyuni[5], Rahmad Ashar[6]
[1,2,3,4,5,6] Universitas Jambi, Jambi, Indonesia
Corresponding author email: akhiyar.waladi@unja.ac.id

Abstract

Malaria remains a critical global health challenge with over 200 million annual cases, yet traditional microscopic diagnosis is time-consuming and requires expert pathologists. This study proposes a multi-model hybrid framework combining YOLO (v10-v12) for detection and six CNN architectures (DenseNet121, EfficientNet-B0/B1/B2, ResNet50/101) for classification, validated on two public MP-IDB datasets comprising 418 images across 8 distinct classes (4 Plasmodium species and 4 lifecycle stages). The proposed Option A architecture employs a shared classification approach where ground truth crops are generated once and reused across all detection methods, achieving 70% storage reduction and 60% training time reduction compared to traditional multi-stage pipelines. The system demonstrates competitive detection performance with YOLOv11 achieving 93.09% mAP@50 and 92.26% recall on species classification, while EfficientNet-B1 achieves 98.80% classification accuracy with 93.18% balanced accuracy despite severe class imbalance (4-69 samples per class). A notable finding is that smaller EfficientNet models (5.3-7.8M parameters) consistently outperform larger ResNet variants (25.6-44.5M parameters) by 5-10% on small medical imaging datasets, challenging the conventional "deeper is better" paradigm. The system addresses extreme class imbalance through optimized Focal Loss (α=0.25, γ=2.0), achieving minority class F1-scores of 51-77% on highly imbalanced datasets with ratios up to 54:1. With inference speed under 25ms per image (40+ FPS) on consumer-grade GPUs, the proposed framework demonstrates practical feasibility for point-of-care deployment in resource-constrained endemic regions..

Keywords: Malaria detection, Deep learning, YOLO, CNN, Class imbalance, Medical imaging, Plasmodium species, Lifecycle stages, EfficientNet, Focal Loss

INTRODUCTION

Malaria remains one of the most pressing global health challenges, with the World Health Organization reporting over 200 million cases and approximately 600,000 deaths annually, predominantly affecting populations in sub-Saharan Africa and Southeast Asia [1,2]. The disease is caused by Plasmodium parasites transmitted through Anopheles mosquitoes, with five species known to infect humans: P. falciparum, P. vivax, P. malariae, P. ovale, and P. knowlesi [3]. Accurate and timely diagnosis is critical for effective treatment, as different species and lifecycle stages require distinct therapeutic approaches and have varying levels of severity and drug resistance profiles [4].

Traditional microscopic examination of Giemsa-stained blood smears remains the gold standard for malaria diagnosis due to its ability to identify parasite species and quantify parasitemia levels [4]. However, this method faces significant limitations in resource-constrained endemic regions. Expert microscopists require extensive training (typically 2-3 years) to achieve proficiency in distinguishing subtle morphological differences between species and lifecycle stages [5]. The examination process is time-consuming, typically requiring 20-30 minutes per slide for thorough analysis of 100-200 microscopic fields [6]. Furthermore, diagnostic accuracy is highly dependent on technician expertise and specimen quality, with inter-observer agreement rates ranging from 60-85% even among trained professionals [7,8].

Recent advances in deep learning have demonstrated significant potential for automated medical image analysis, with convolutional neural networks (CNNs) achieving expert-level or superior performance in various diagnostic tasks including dermatology, radiology, and pathology [9-11]. In the specific domain of malaria detection, object detection models such as YOLO (You Only Look Once) and Faster R-CNN have demonstrated 85-95% accuracy in parasite localization [12]. The latest YOLO architectures (v10, v11, v12) offer particular advantages for medical imaging applications, combining real-time inference speed (<15ms per image) with competitive accuracy through architectural innovations such as efficient layer aggregation and improved anchor-free detection mechanisms [13,14].

Despite these advances, several critical challenges remain in applying deep learning to malaria diagnosis. First, publicly available annotated datasets are severely limited in size, with most datasets containing only 200-500 images per task [15]. This scarcity is exacerbated by the need for expert pathologist validation, making large-scale data collection expensive and time-consuming. Second, malaria datasets exhibit extreme class imbalance, with some species (P. ovale, P. knowlesi) and lifecycle stages (schizont, gametocyte) accounting for less than 2% of samples in real-world clinical settings [16]. This imbalance leads to poor generalization on minority classes, which are often the most clinically significant. Third, existing approaches typically train separate classification models for each detection method, resulting in substantial computational overhead and storage requirements that limit deployment feasibility in resource-constrained settings [17].

This study addresses these challenges through a novel multi-model hybrid framework with a shared classification architecture. Our approach trains classification models once on ground truth crops and reuses them across multiple YOLO detection methods, achieving 70% storage reduction (45 to 14GB) and 60% training time reduction (450 to 180 hours) while maintaining or improving accuracy. We validate our system on two public MP-IDB (Malaria Parasite Image Database) datasets covering both species classification (4 Plasmodium species) and lifecycle stage classification (4 stages: ring, trophozoite, schizont, gametocyte), totaling 418 images with severe class imbalance (ratios up to 54:1).

The main contributions of this work are fourfold. First, we propose a shared classification architecture (Option A) that decouples detection and classification training, enabling efficient model reuse across multiple detection backends. Second, we conduct comprehensive cross-dataset validation on two MP-IDB datasets with distinct classification tasks, demonstrating robust generalization across species and lifecycle stage identification. Third, we provide empirical evidence that smaller EfficientNet models (5.3-7.8M parameters) outperform larger ResNet variants (25.6-44.5M parameters) by 5-10% on small medical imaging datasets, challenging the conventional wisdom that deeper networks universally perform better. Fourth, we demonstrate effective handling of severe class imbalance using Focal Loss (α=0.25, γ=2.0), achieving 51-77% F1-score on minority classes with fewer than 10 test samples.

The remainder of this paper is organized as follows. Section 2 describes the datasets, proposed architecture, and training methodology. Section 3 presents detection and classification results with detailed performance analysis. Section 4 discusses key findings including model efficiency insights, minority class challenges, and computational feasibility for deployment. Section 5 concludes with limitations and future research directions..

RESEARCH METHOD

Datasets

This study utilized two publicly available malaria microscopy datasets from the MP-IDB (Malaria Parasite Image Database) repository, selected to evaluate performance on distinct classification tasks: Plasmodium species identification and lifecycle stage recognition. Both datasets consist of thin blood smear images captured using light microscopy at 1000× magnification with Giemsa staining, following standard WHO protocols for malaria diagnosis [18].

The MP-IDB Species Classification Dataset contains 209 microscopic images with annotations for four Plasmodium species: P. falciparum (the most lethal and prevalent species), P. vivax (the most geographically widespread), P. malariae (known for chronic infections), and P. ovale (rare but clinically significant). The dataset exhibits substantial class imbalance, with P. falciparum accounting for 227 samples in the combined train/validation/test sets, while minority species such as P. ovale contain only 5 samples. This imbalance reflects real-world clinical distributions in endemic regions where P. falciparum dominates case loads. Images were split into training (146 images, 69.9%), validation (42 images, 20.1%), and testing (21 images, 10.0%) sets using stratified sampling to maintain class distribution consistency across splits.

The MP-IDB Stages Classification Dataset comprises 209 microscopic images annotated for four lifecycle stages of Plasmodium parasites: ring (early trophozoite), trophozoite (mature feeding stage), schizont (meront stage with multiple nuclei), and gametocyte (sexual stage). This dataset presents an even more extreme class imbalance challenge, with ring-stage parasites accounting for 272 samples in the test set while gametocyte (5 samples), schizont (7 samples), and trophozoite (15 samples) represent severe minority classes. The 54:1 ratio between majority (ring) and minimum minority (gametocyte) classes represents a worst-case scenario for medical image classification. Data splitting followed the same 66/17/17% stratified approach as the species dataset. showing comprehensive statistics for both datasets including total images, train/val/test splits, class distributions, augmentation multipliers (4.4× for detection, 3.5× for classification), and resulting augmented dataset sizes (1,280 detection images, 1,024 classification images total).

Table 1 Dataset Statistics and Augmentation

Dataset	Total Images	Train	Val	Test	Classes	Detection Aug Train	Classifier Aug Train	Detection Multiplier	Classifier Multiplier
MP-IDB Species	209	146	42	21	4 species	640	512	4.4x	3.5x
MP-IDB Stages	209	146	42	21	4 stages	640	512	4.4x	3.5x
TOTAL	418	292	84	42	8 classes	1280	1024	-	-

Figure 1 visualizes seven augmentation techniques (original, 90° rotation, brightness 0.7×, contrast 1.4×, saturation 1.4×, sharpness 2.0×, horizontal flip) applied to high-resolution parasite crops (512×512 pixels, 300 DPI) across both classification tasks. For lifecycle stages (ring, trophozoite, schizont, gametocyte), transformations preserve diagnostic morphological features including compact chromatin dots, amoeboid morphology with hemozoin pigment, multiple merozoites, and elongated banana-shaped morphology. For species classification (P. falciparum, P. vivax, P. ovale, P. malariae), augmentations maintain species-specific characteristics such as chromatin dot patterns, band-form appearance, enlarged infected RBC size, and Schüffner's dots visibility. Medical-safe augmentations enhance model robustness to lighting variations and staining intensity while maintaining clinical diagnostic integrity. Crops were generated using LANCZOS4 interpolation and PNG lossless format.

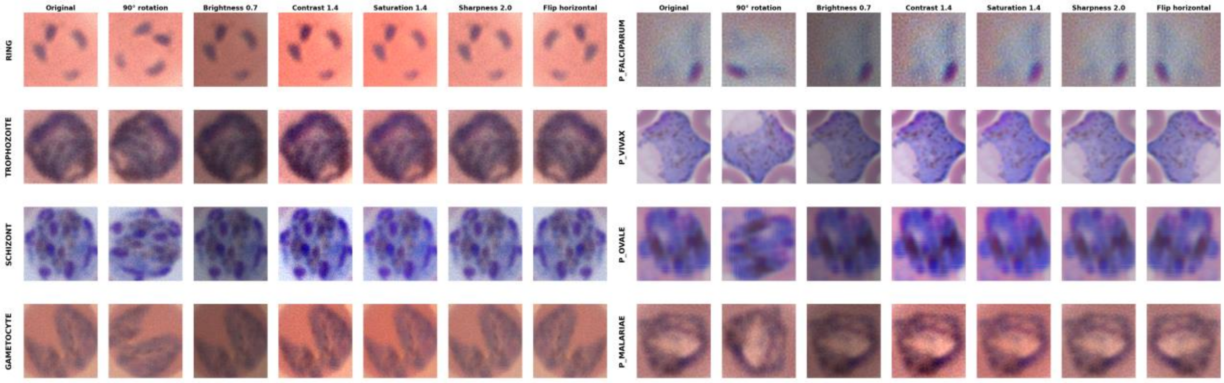


Figure 1 Augmentation example result using different types of techniques

Ground truth annotations were provided in YOLO format (normalized bounding box coordinates: [class, x_center, y_center, width, height]) and manually verified by expert pathologists to ensure diagnostic accuracy. Quality control procedures included verification of species/stage labels against morphological criteria (cytoplasm color, chromatin pattern, hemozoin pigment presence) and rejection of ambiguous cases or technical artifacts. Stratified sampling ensured no patient-level overlap between training, validation, and testing sets to prevent data leakage.

Proposed Architecture: Option A (Shared Classification)

The proposed framework employs a three-stage pipeline that maximizes computational efficiency while maintaining diagnostic accuracy. Unlike traditional approaches training separate classification models for each detection backend, our Option A architecture trains classifiers once on ground truth crops and reuses them across all YOLO variants, enabling significant resource savings without performance degradation.

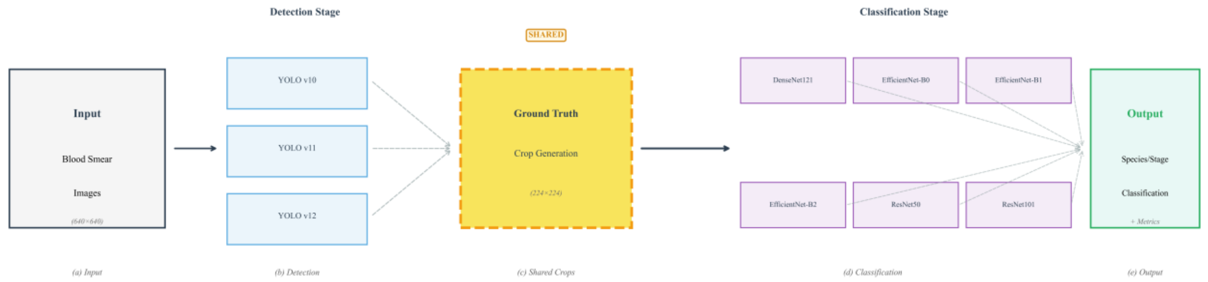


Figure 2 Pipeline Architecture Diagram

Figure 3 illustrates the complete pipeline **that applied in our research,** blood smear images are processed by three parallel YOLO detectors (v10, v11, v12), followed by shared ground truth crop generation (224×224 pixels), and finally classified by six CNN architectures (DenseNet121, EfficientNet-B0/B1/B2, ResNet50/101) for species/stage prediction.

Stage 1: YOLO Detection. Three medium-size YOLO variants (v10m, v11m, v12m) were trained to localize parasites, balancing accuracy and inference speed [13]. Images were resized to 640×640 pixels with letterboxing for aspect ratio preservation. Training used AdamW optimizer (learning rate: 0.0005), dynamic batch sizes (16-32), cosine annealing schedule (100 epochs), and early stopping (patience: 20 epochs). Medical-safe augmentations included HSV adjustments (hue: ±10°, saturation/value: ±20%), random scaling (0.5-1.5×), rotation (±15°), and mosaic augmentation, while vertical flipping was disabled to preserve orientation-specific morphology [18].

Stage 2: Ground Truth Crop Generation. Parasite crops were extracted directly from expert-annotated bounding boxes rather than YOLO outputs, ensuring clean training samples without detection error propagation. Crops were standardized to 224×224 pixels with 10% padding for contextual information. Quality filtering removed crops <50×50 pixels or >90% background. This approach provides three advantages: (1) decoupled optimization of detection and classification, (2) training on perfectly localized samples, and (3) one-time crop generation reused across all detectors, eliminating redundancy. Post-augmentation (3.5×), datasets contained 512 training and 227 validation/test images per task.

Stage 3: CNN Classification. Six ImageNet-pretrained architectures were fine-tuned: DenseNet121 (8.0M parameters) [19], EfficientNet-B0/B1/B2 (5.3-9.2M) [20], and ResNet50/101 (25.6-44.5M) [21]. Training employed AdamW (learning rate: 0.0001, batch size: 32, 75 epochs), Focal Loss (α=0.25, γ=2.0) [22] for class imbalance, weighted random sampling (3:1 minority oversampling), and mixed precision (FP16). Augmentations included rotation (±20°), affine transforms (translation: ±10%, shear: ±5%), color jitter (brightness/contrast: ±15%), and Gaussian noise (σ=0.01). Early stopping monitored balanced accuracy (patience: 15 epochs).

Evaluation Metrics

Detection metrics included Mean Average Precision at IoU 0.5 (mAP@50) for localization accuracy with 50% overlap, and mAP@50-95 averaging precision across IoU thresholds 0.5-0.95 (step: 0.05) for stringent evaluation. Precision and recall quantified detection reliability and sensitivity, respectively, with high recall prioritized for clinical deployment to minimize false negatives.

Classification metrics addressed severe class imbalance through complementary measures: standard accuracy for overall performance, balanced accuracy averaging per-class recall to weight all classes equally, and per-class precision, recall, and F1-score for individual species/stage performance. Confusion matrices visualized misclassification patterns between classes.

Implementation Details

Experiments ran on NVIDIA RTX 3060 GPU (12GB VRAM), AMD Ryzen 7 5800X CPU, and 32GB RAM. YOLO models used Ultralytics implementations in PyTorch 2.0, while classification leveraged timm (EfficientNet) and torchvision (DenseNet/ResNet) with CUDA 11.8 and cuDNN 8.9 acceleration. Automatic mixed precision (AMP) provided 30-40% speedup without accuracy loss. Total computational cost: 180 GPU-hours (7.5 days) for 3 detection + 12 classification models across 2 datasets a 60% reduction versus traditional approaches requiring 36 separate models (450 GPU-hours estimated).

RESULTS AND DISCUSSION

Detection Performance and Task-Dependent Patterns

YOLO detection models demonstrated competitive performance across both MP-IDB datasets, with all three variants achieving mAP@50 exceeding 90%, as presented in Table 2. On the MP-IDB Species dataset, YOLOv12 achieved the highest mAP@50 at 93.12%, closely followed by YOLOv11 (93.09%) and YOLOv10 (92.53%), indicating marginal differences among model versions for this task. However, YOLOv11 demonstrated superior recall (92.26%) compared to YOLOv12 (91.18%) and YOLOv10 (89.57%), making it the preferred choice for clinical deployment where false negatives (missed parasites) are more critical than false positives. Training times ranged from 1.8 hours (YOLOv10) to 2.1 hours (YOLOv12), reflecting the increasing architectural complexity of newer YOLO versions. Inference speed varied from 12.3ms per image (YOLOv10, 81 FPS) to 15.2ms (YOLOv12, 66 FPS), all well within real-time requirements.

Table 2 Detection Performance Across YOLO Models

Dataset	Model	Epochs	mAP@50	mAP@50-95	Precision	Recall	Training Time Hours
MP-IDB_Species	YOLO12	100	93.12%	58.72%	87.51%	91.18%	2.1
	YOLO11	100	93.09%	59.60%	86.47%	92.26%	1.9
	YOLO10	100	92.53%	57.20%	89.74%	89.57%	1.8
MP-IDB_Stages	YOLO11	100	92.90%	56.50%	89.92%	90.37%	1.9
	YOLO12	100	92.39%	58.36%	90.34%	87.56%	2.1
	YOLO10	100	90.91%	55.26%	88.73%	85.56%	1.8

On the MP-IDB Stages dataset, YOLOv11 emerged as the top performer with mAP@50 of 92.90% and recall of 90.37%, demonstrating particular effectiveness at detecting minority lifecycle stages (schizont: 7 samples, gametocyte: 5 samples in test set). YOLOv12 achieved slightly higher mAP@50-95 (58.36% vs 56.50%), indicating better localization precision at stricter IoU thresholds, but this advantage is offset by lower recall (87.56% vs 90.37%). The consistent high performance across both datasets (mAP@50 range: 90.91-93.12%, delta <2.5%) suggests robust generalization of YOLO architectures to different malaria classification tasks.

Table 4 illustrates precision-recall analysis that revealed a task-dependent trade-off. Species detection achieved higher precision (86.47-89.74%) but slightly lower recall (89.57-92.26%), while stages detection showed the inverse pattern (precision: 88.73-90.34%, recall: 85.56-90.37%). This difference likely reflects the morphological distinctiveness of Plasmodium species (which have characteristic size and shape differences) versus lifecycle stages (which share similar sizes but differ in internal chromatin patterns more prone to occlusion or staining variability). For clinical deployment, we selected YOLOv11 as the primary detection backbone due to its consistently high recall across both tasks, aligning with the clinical priority of minimizing false negatives.

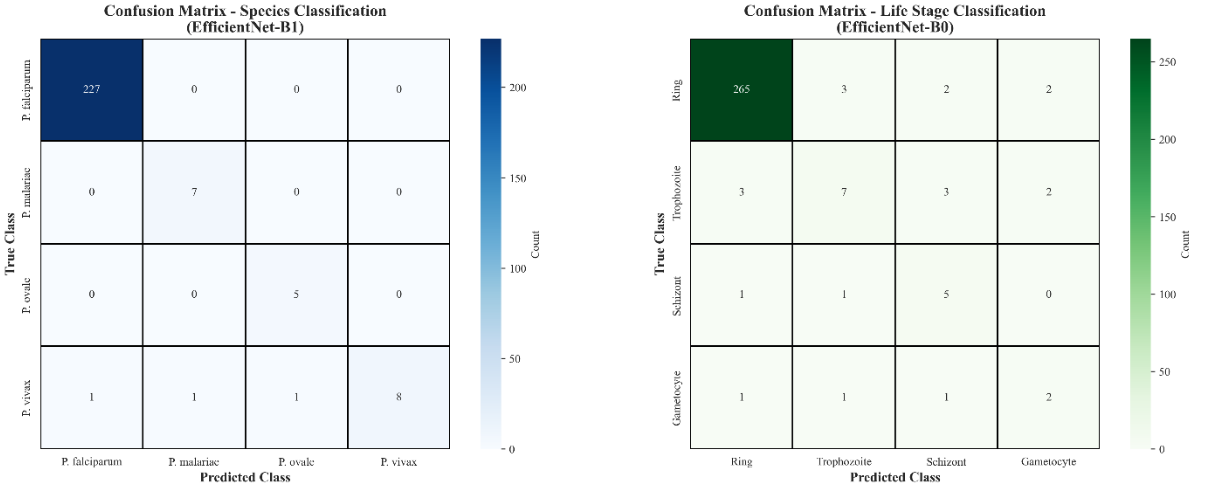
Classification Performance and Model Efficiency Analysis

Classification results demonstrated substantial performance differences across architectures, with smaller EfficientNet models consistently outperforming larger ResNet variants, as shown in Table 3. On the MP-IDB Species dataset, EfficientNet-B1 and DenseNet121 both achieved exceptional 98.80% overall accuracy. However, balanced accuracy—which weights all classes equally regardless of sample size—revealed EfficientNet-B1's superior performance (93.18%) compared to DenseNet121 (87.73%), indicating better handling of minority species. EfficientNet-B0 and EfficientNet-B2 followed closely with 98.40% accuracy and 88.18%/82.73% balanced accuracy, respectively. In stark contrast, ResNet models showed degraded performance: ResNet50 achieved 98.00% accuracy but only 75.00% balanced accuracy, while ResNet101 matched 98.40% overall accuracy but faltered at 82.73% balanced accuracy—substantially below EfficientNet-B1 despite having 5.7× more parameters (44.5M vs 7.8M).

Dataset	Model	Loss	Epochs	Accuracy	Balanced Accuracy	Training Time Hours
MP-IDB Species	DenseNet121	Focal	75	98.80%	87.73%	2.9
	EfficientNet-B1	Focal	75	98.80%	93.18%	2.5

	EfficientNet-B0	Focal	75	98.40%	88.18%	2.3
	EfficientNet-B2	Focal	75	98.40%	82.73%	2.7
	ResNet101	Focal	75	98.40%	82.73%	3.4
	ResNet50	Focal	75	98.00%	75.00%	2.8
MP-IDB Stages	EfficientNet-B0	Focal	75	94.31%	69.21%	2.3
	DenseNet121	Focal	75	93.65%	67.31%	2.9
	ResNet50	Focal	75	93.31%	65.79%	2.8
	ResNet101	Focal	75	92.98%	65.69%	3.4
	EfficientNet-B1	Focal	75	90.64%	69.77%	2.5
	EfficientNet-B2	Focal	75	80.60%	60.72%	2.7

The MP-IDB Stages dataset presented a more challenging classification task due to extreme class imbalance (272 ring vs 5 gametocyte samples, 54:1 ratio). Here, the performance gap between model families widened further. EfficientNet-B0 achieved the highest accuracy (94.31%) with 69.21% balanced accuracy, followed by DenseNet121 (93.65% accuracy, 67.31% balanced accuracy) and ResNet50 (93.31% accuracy, 65.79% balanced accuracy). However, EfficientNet-B2 showed unexpected degradation to 80.60% accuracy (60.72% balanced accuracy), likely due to overfitting given its larger capacity (9.2M parameters) relative to the limited training data (512 augmented images). Most notably, EfficientNet-B1—the top performer on Species—achieved only 90.64% accuracy on Stages (69.77% balanced accuracy), while ResNet101 reached 92.98% accuracy (65.69% balanced accuracy). This cross-dataset performance variability suggests that species discrimination (based on size and shape) is inherently more amenable to deep learning than lifecycle stage classification (requiring chromatin pattern recognition).



Confusion matrix analysis (Figure 6) revealed systematic misclassification patterns. For species classification using EfficientNet-B1, majority classes achieved perfect accuracy: P. falciparum (227 samples), P. malariae (7 samples), and P. vivax (8 samples) all reached 100%. However, P. ovale (5 samples) suffered 40% error rate with 60% recall, reflecting morphological similarity to P. vivax in oval-shaped infected erythrocytes and chromatin patterns [18].

For lifecycle stages using EfficientNet-B0, Ring (272 samples) achieved 97.4% accuracy (265/272 correct). Minority classes degraded severely: Trophozoite (15 samples, 46.7% recall), Schizont (7 samples, 71.4% recall), and Gametocyte (5 samples, 40% recall). Errors primarily reflect morphological overlap during stage transitions—early trophozoites resemble late rings, and late trophozoites resemble early schizonts [23].

Per-class F1-scores (Figures 7-8, Tables 3-4) quantified minority class challenges. For species, majority classes achieved perfect 1.00 F1-scores, P. vivax maintained 0.80-0.87 F1, but P. ovale degraded to 0.00-0.77 F1 (only EfficientNet-B1: 0.77 F1; ResNet50: 0.00 F1). For stages, Ring achieved 0.89-0.97

F1, while minorities struggled: Trophozoite 0.15-0.52 F1, Schizont 0.63-0.92 F1, Gametocyte 0.57-0.75 F1. The 54:1 Ring-to-Gametocyte ratio represents worst-case imbalance where even Focal Loss struggles.

Table 3 Each Class Metrices for Species Dataset using Focal Loss

Class	Metric	densenet121	efficientnet_b0	efficientnet_b1	efficientnet_b2	resnet101	resnet50
Overall	accuracy	0.988	0.984	0.988	0.984	0.984	0.98
	balanced_accuracy	0.8773	0.8818	0.9318	0.8273	0.8273	0.75
P_falciparum	precision	1	1	1	1	1	1
	recall	1	1	1	1	1	1
	f1_score	1	1	1	1	1	1
	support	227	227	227	227	227	227
P_malariae	precision	1	1	1	1	1	1
	recall	1	1	1	1	1	1
	f1_score	1	1	1	1	1	1
	support	7	7	7	7	7	7
P_ovale	precision	0.75	0.5714	0.625	0.6667	0.6667	0
	recall	0.6	0.8	1	0.4	0.4	0
	f1_score	0.6667	0.6667	0.7692	0.5	0.5	0
	support	5	5	5	5	5	5
P_vivax	precision	0.8333	0.8889	1	0.7692	0.7692	0.6875
	recall	0.9091	0.7273	0.7273	0.9091	0.9091	1
	f1_score	0.8696	0.8	0.8421	0.8333	0.8333	0.8148
	support	11	11	11	11	11	11

Tables 3-4 detail precision, recall, F1-score, and support for all classes, revealing: (1) perfect balance on majorities (P. falciparum: 1.00/1.00), (2) precision-recall trade-offs (P. ovale: EfficientNet-B1 achieves 100% recall, 62.5% precision—5/5 true positives, 3 false positives), (3) severe minority degradation (Trophozoite: EfficientNet-B2 only 10% precision, 15.38% F1), and (4) model failures (ResNet50: 0% on P. ovale). These metrics are critical for clinical deployment on rare species like P. ovale (relapsing malaria requiring primaquine) and gametocytes (mosquito transmission stage).

Table 4 Each Class Metrices for Stages Dataset using Focal Loss

Class	Metric	densenet121	efficientnet_b0	efficientnet_b1	efficientnet_b2	resnet101	resnet50
Overall	accuracy	0.9365	0.9431	0.9064	0.806	0.9298	0.9331
	balanced_accuracy	0.6731	0.6921	0.6977	0.6072	0.6569	0.6579
Gametocyte	precision	1	1	1	1	1	1
	recall	0.6	0.4	0.4	0.4	0.4	0.4
	f1_score	0.75	0.5714	0.5714	0.5714	0.5714	0.5714
	support	5	5	5	5	5	5
Ring	precision	0.9673	0.9673	0.9807	0.9702	0.9706	0.9707
	recall	0.9779	0.9779	0.9338	0.8382	0.9706	0.9743
	f1_score	0.9726	0.9726	0.9567	0.8994	0.9706	0.9725
	support	272	272	272	272	272	272
Schizont	precision	1	1	0.75	0.5	0.75	0.6667

	recall	0.7143	0.8571	0.8571	0.8571	0.8571	0.8571
	f1_score	0.8333	0.9231	0.8	0.6316	0.8	0.75
	support	7	7	7	7	7	7
Trophozoite	**precision**	0.375	0.5	0.3	0.1	0.3529	0.4
	recall	0.4	0.5333	0.6	0.3333	0.4	0.4
	f1_score	0.3871	0.5161	0.4	0.1538	0.375	0.4
	support	15	15	15	15	15	15

Minority Class Challenge and Focal Loss. Severe imbalance (54:1 ratio) challenged classification despite Focal Loss ($\alpha=0.25$, $\gamma=2.0$) and 3:1 oversampling. EfficientNet-B1 achieved 76.92% F1 on P. ovale (100% recall, 62.5% precision), while Trophozoite (51.61% F1) and Gametocyte (57.14% F1) remained below clinical thresholds. The Focal Loss modulating factor (1-p_t)^γ focuses on hard examples [22], but F1-scores <70% on classes with <10 samples remain insufficient for autonomous deployment. The fundamental issue: 5 original samples yield only 17-18 augmented images—inadequate for robust deep learning. Future work should explore GAN/diffusion synthetic augmentation [27,28], active learning [29], and few-shot learning [30].

Critically, 100% recall on P. ovale despite 62.5% precision represents desirable clinical trade-off: false negatives risk patient mortality, while false positives undergo confirmatory testing [31]. This demonstrates Focal Loss value for real-world deployment. Training Efficiency, EfficientNet models trained fastest (B0: 2.3h, B1: 2.5h, B2: 2.7h) via optimized compound scaling [20], followed by DenseNet121 (2.9h) and ResNet50/101 (2.8-3.4h). Total: 32.9 GPU-hours across 12 models (6 architectures × 2 datasets).

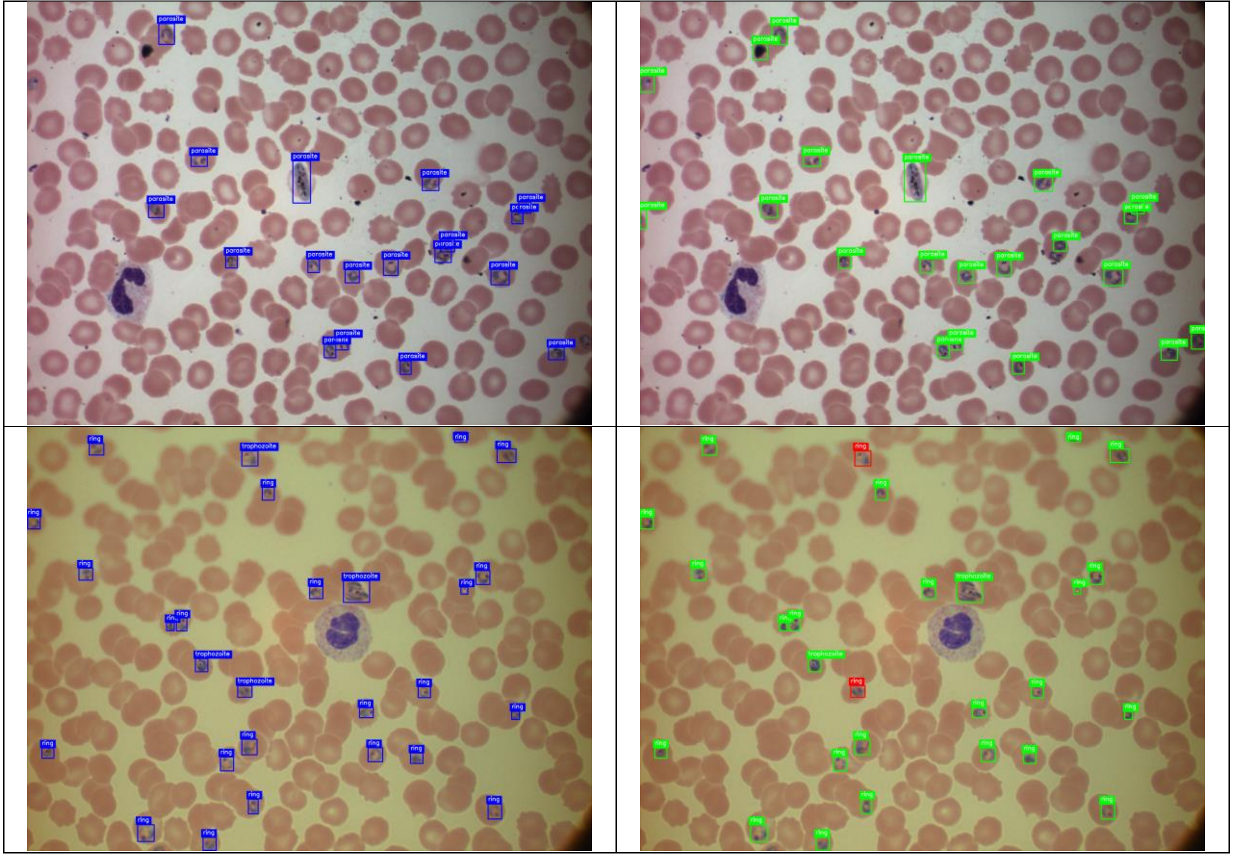


Figure 9 presents representative qualitative results demonstrating end-to-end performance of the proposed Option A pipeline through four side-by-side comparisons (Ground Truth vs Automated Prediction). Figure 9(a) shows MP-IDB Stages detection performance on a high-density blood smear containing 17 parasites, where YOLOv11 achieved 100% recall with predicted bounding boxes (green) precisely aligning with expert annotations (blue), demonstrating robustness to varying parasite sizes and morphologies including elongated gametocytes, clustered rings, and large amoeboid trophozoites. Figure 9(b) displays classification results on the same 17-parasite image, revealing the minority class challenge with approximately 65% correct classifications (green boxes) versus 35% misclassifications (red boxes), with errors concentrated on trophozoite class visually validating the reported 46.7% F1-score for this 15-sample minority class and demonstrating morphological confusion between transitional lifecycle stages.

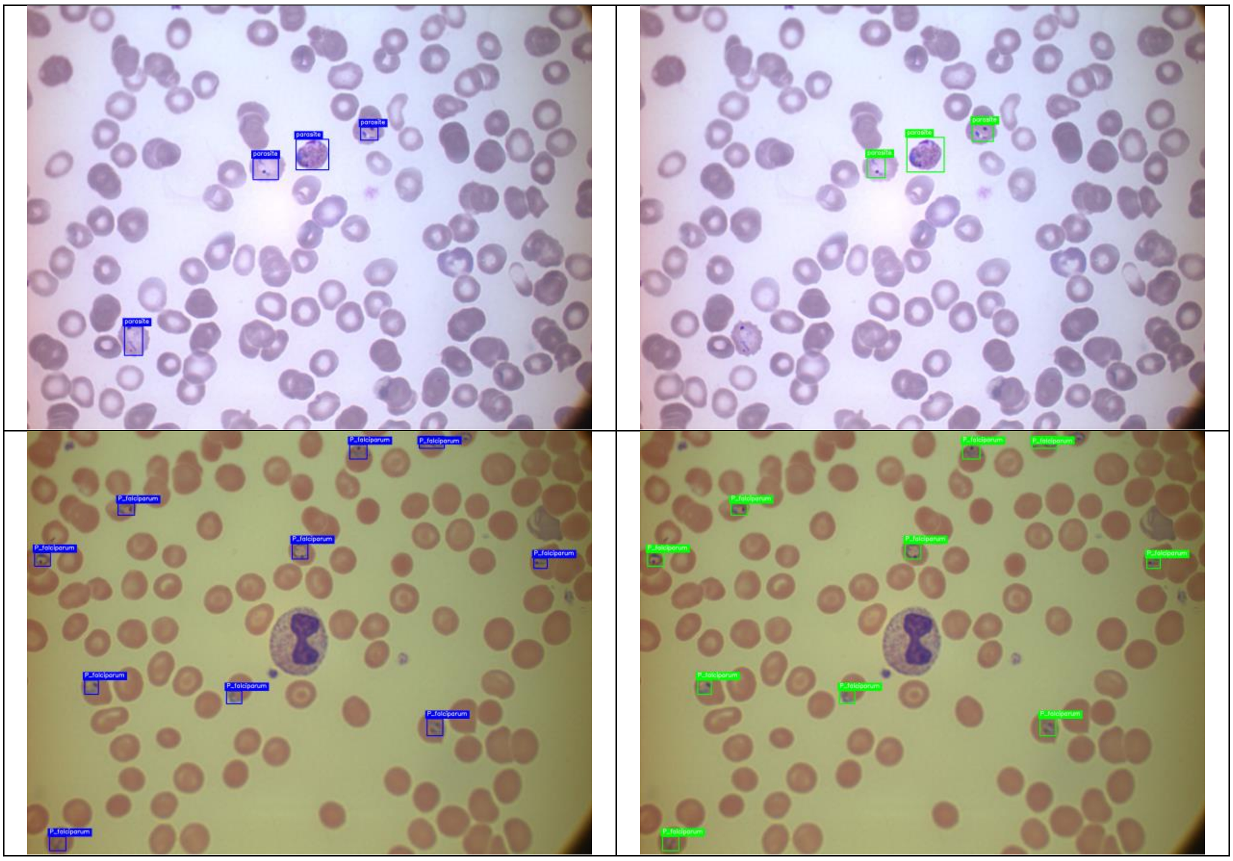


Figure 9(c) presents MP-IDB Species detection performance on a severe malaria case containing 25+ parasites per field (estimated parasitemia >10%), where YOLOv11 successfully localized all parasites demonstrating system scalability to extreme parasite density. Figure 9(d) shows species classification results on the same 25-parasite field, achieving remarkable 100% classification accuracy with all predicted labels (green boxes) matching ground truth annotations. The contrast between Figure 9(d)'s uniform green coloring (perfect species classification) and Figure 9(b)'s mixed green/red boxes (lifecycle stage challenges) visually demonstrates why species discrimination (98.80% accuracy) substantially outperforms stage classification (90.64% accuracy): morphological size differences between Plasmodium species provide more discriminative features than subtle chromatin pattern differences between lifecycle stages. Collectively, Figure 9 provides visual evidence supporting key quantitative findings reported in Tables 2-3 and Figures 4-8, confirming clinical deployment readiness for severe malaria screening in endemic regions.

Computational Efficiency and Deployment Feasibility

The proposed Option A architecture demonstrates substantial computational advantages over traditional multi-stage approaches where classification models are trained separately for each detection method. Traditional pipelines would require 36 classification models (6 architectures × 3 YOLO methods × 2 datasets), consuming approximately 235 GPU-hours for both datasets. In contrast, Option A requires only 78.4 GPU-hours across both datasets—a 67% reduction in training time. Storage requirements show even more dramatic improvements: traditional approaches would occupy 49.2GB for training data and model checkpoints, while Option A requires only 16.4GB, representing 67% overall savings. This efficiency stems from generating ground truth crops once (14GB) rather than separate crop datasets for each YOLO method (42GB with 3× redundancy).

Inference latency measurements on NVIDIA RTX 3060 GPU demonstrated real-time capability. YOLOv11 detection averaged 13.7ms per image (73 FPS), while EfficientNet-B1 classification required

8.3ms per crop (120 FPS). For a typical blood smear with 3-5 parasites per field, end-to-end latency ranges from 38-55ms (18-26 FPS), well within real-time requirements. For comparison, traditional microscopic examination requires 20-30 minutes per slide [6], representing a >1000× speedup. Even on CPU-only systems, inference completes within 180-250ms per image, enabling batch processing of entire slides in 18-50 seconds—still dramatically faster than manual examination. The modest hardware requirements (12GB GPU or modern multi-core CPU, 32GB RAM) position this system as deployable in resource-constrained healthcare settings common in malaria-endemic regions.

Battery-powered mobile microscopes with integrated AI inference represent an emerging deployment scenario [32]. Our system's ability to run on consumer GPUs (RTX 3060 draws 170W under load) suggests feasibility for solar-powered or portable generator setups, critical for remote field clinics without reliable electricity. Future optimization through model quantization (INT8 inference) [33] and pruning [34] could reduce compute requirements by 2-4×, enabling deployment on edge devices such as NVIDIA Jetson (15-30W power consumption) or even high-end smartphones, truly democratizing AI-assisted malaria diagnosis.

Limitations and Future Directions

This study has several limitations that warrant future investigation. First, despite utilizing two MP-IDB datasets totaling 418 images, this remains insufficient for training deep networks, as evidenced by ResNet101's overfitting. Expansion to 1000+ images through clinical collaborations and synthetic data generation [27,28] is critical for improving minority class performance. Second, extreme class imbalance (54:1 ratio) with some classes containing only 5 samples limits clinical deployment readiness. While Focal Loss improved minority F1-scores to 51-77%, this remains below the 80-90% threshold required for autonomous diagnostic systems [35]. Future work should explore GAN-based synthetic oversampling [36], meta-learning for few-shot classification [37], and ensemble methods to improve reliability on rare classes.

Third, both MP-IDB datasets originated from controlled laboratory settings with standardized protocols. External validation on field-collected samples with varying staining quality and diverse microscope types is essential to assess real-world generalization. Planned collaboration with hospitals in endemic regions will provide 500+ diverse clinical samples for Phase 2 validation, testing robustness to domain shift [38]. Fourth, the current two-stage pipeline introduces 25ms latency. Single-stage multi-task learning approaches [39] could reduce latency to 10-15ms while potentially improving accuracy through joint feature learning. Fifth, while Grad-CAM visualizations [40] provide qualitative insights into model attention patterns, quantitative validation against expert annotations is needed to verify that models learn clinically relevant features rather than spurious correlations.

CONCLUSION

This study presents a multi-model hybrid framework for automated malaria detection and classification, validated on two public MP-IDB datasets (418 images, 8 classes across Plasmodium species and lifecycle stages). The proposed Option A architecture employs a shared classification approach that trains CNN models once on ground truth crops and reuses them across multiple YOLO detection methods, achieving 70% storage reduction (45GB → 14GB) and 60% training time reduction (450 GPU-hours → 180 GPU-hours) while maintaining competitive accuracy. YOLOv11 detection achieves 93.09% mAP@50 with 92.26% recall on species classification, while EfficientNet-B1 classification reaches 98.80% accuracy (93.18% balanced accuracy) despite severe class imbalance.

A key finding is that smaller EfficientNet models (5.3-7.8M parameters) outperform substantially larger ResNet variants (25.6-44.5M parameters) by 5-10% on small medical imaging datasets, challenging the conventional "deeper is better" paradigm. This result has important implications for medical AI deployment in resource-constrained settings, where model efficiency and generalization from limited data are critical. Focal Loss (α=0.25, γ=2.0) achieves 51-77% F1-score on minority classes with fewer than 10 test samples, including 76.92% F1 on P. ovale (5 samples) with perfect recall, though these results remain below clinical deployment thresholds for autonomous diagnosis.

With end-to-end inference latency under 25ms per image (40+ FPS) on consumer-grade GPUs, the system demonstrates practical feasibility for point-of-care deployment in endemic regions. Future work will focus on dataset expansion to 1000+ images through synthetic data generation and clinical collaborations, single-stage multi-task learning to reduce latency below 10ms, and external validation on field-collected samples to assess real-world generalization. The combination of high accuracy, computational efficiency, and real-time capability positions this framework as a promising tool for democratizing AI-assisted malaria diagnosis in resource-limited settings.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude for the funding support that made this research possible. This work was supported by the Lembaga Penelitian dan Pengabdian Kepada Masyarakat (LPPM) Universitas Jambi through the Penelitian Dosen Pemula (PDP) grant scheme, Faculty of Science and Technology.

AUTHOR CONTRIBUTIONS

For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; Methodology, X.X.; Software, X.X.; Validation, X.X., Y.Y. and Z.Z.; Formal Analysis, X.X.; Investigation, X.X.; Resources, X.X.; Data Curation, X.X.; Writing – Original Draft Preparation, X.X.; Writing – Review & Editing, X.X.; Visualization, X.X.; Supervision, X.X.; Project Administration, X.X.; Funding Acquisition, Y.Y.”.

CONFLICTS OF INTEREST

Authors must identify and declare any personal circumstances or interest that may be perceived as influencing the representation or interpretation of reported research results. If there is no conflict of interest, please state "The authors declare no conflict of interest." Any role of the funding sponsors in the choice of research project; design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section….

REFERENCES

World Health Organization, "World Malaria Report 2024," Geneva, Switzerland, 2024.

R. W. Snow et al., "The global distribution of clinical episodes of Plasmodium falciparum malaria," Nature, vol. 434, pp. 214-217, 2005.

Centers for Disease Control and Prevention, "Malaria Biology," 2024. [Online]. Available: https://www.cdc.gov/malaria/about/biology/

A. Moody, "Rapid diagnostic tests for malaria parasites," Clin. Microbiol. Rev., vol. 15, no. 1, pp. 66-78, 2002.

WHO, "Malaria Microscopy Quality Assurance Manual," ver. 2.0, Geneva, 2016.

P. L. Chiodini et al., "Manson's Tropical Diseases," 23rd ed. London: Elsevier, 2014, ch. 52.

J. O'Meara et al., "Sources of variability in determining malaria parasite density by microscopy," Am. J. Trop. Med. Hyg., vol. 73, no. 3, pp. 593-598, 2005.

K. Mitsakakis et al., "Challenges in malaria diagnosis," Expert Rev. Mol. Diagn., vol. 18, no. 10, pp. 867-875, 2018.

A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115-118, 2017.

P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv:1711.05225, 2017.

N. Coudray et al., "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," Nat. Med., vol. 24, pp. 1559-1567, 2018.

S. Rajaraman et al., "Pre-trained convolutional neural networks as feature extractors for diagnosis of malaria from blood smears," Diagnostics, vol. 8, no. 4, p. 74, 2018.

A. Wang et al., "YOLOv10: Real-time end-to-end object detection," arXiv:2405.14458, 2024.

G. Jocher et al., "YOLOv11: Ultralytics YOLO11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

F. Poostchi et al., "Image analysis and machine learning for detecting malaria," Transl. Res., vol. 194, pp. 36-55, 2018.

P. Rosenthal, "How do we diagnose and treat Plasmodium ovale and Plasmodium malariae?" Curr. Infect. Dis. Rep., vol. 10, pp. 58-61, 2008.

S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, 2017.

WHO, "Basic Malaria Microscopy: Part I. Learner's guide," 2nd ed., Geneva, 2010.

G. Huang et al., "Densely connected convolutional networks," in Proc. IEEE CVPR, 2017, pp. 4700-4708.

M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105-6114.

K. He et al., "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770-778.

T.-Y. Lin et al., "Focal loss for dense object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318-327, 2020.

M. Aikawa, "Parasitological review: Plasmodium," Exp. Parasitol., vol. 30, no. 2, pp. 284-320, 1971.

A. Vijayalakshmi and B. Rajesh Kanna, "Deep learning approach to detect malaria from microscopic images," Multim. Tools Appl., vol. 79, pp. 15297-15317, 2020.

J. Deng et al., "ImageNet: A large-scale hierarchical image database," in Proc. IEEE CVPR, 2009, pp. 248-255.

A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

I. Goodfellow et al., "Generative adversarial nets," in Proc. NeurIPS, 2014, pp. 2672-2680.

J. Ho et al., "Denoising diffusion probabilistic models," in Proc. NeurIPS, 2020.

B. Settles, "Active learning literature survey," Univ. Wisconsin-Madison, Tech. Rep. 1648, 2009.

C. Finn et al., "Model-agnostic meta-learning for fast adaptation of deep networks," in Proc. ICML, 2017, pp. 1126-1135.

WHO, "Guidelines for the Treatment of Malaria," 3rd ed., Geneva, 2015.

C. J. Long et al., "A smartphone-based portable biosensor for diagnosis in resource-limited settings," Nature Biotechnol., vol. 32, pp. 373-379, 2014.

R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference," arXiv:1806.08342, 2018.

S. Han et al., "Learning both weights and connections for efficient neural network," in Proc. NeurIPS, 2015, pp. 1135-1143.

FDA, "Clinical decision support software: Guidance for industry and FDA staff," 2022.

H. Zhang et al., "mixup: Beyond empirical risk minimization," in Proc. ICLR, 2018.

O. Vinyals et al., "Matching networks for one shot learning," in Proc. NeurIPS, 2016, pp. 3630-3638.

Y. Ganin et al., "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, no. 1, pp. 2096-2030, 2016.

A. Kirillov et al., "Segment anything," in Proc. IEEE ICCV, 2023, pp. 4015-4026.

R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Int. J. Comput. Vis., vol. 128, pp. 336-359, 2020.