Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 4, No. 3, August 2019, Pp. 197-206

Parameter Efficient Models for Malaria Detection and Classification Using Small-Scale Imbalanced Blood Smear Images

Akhiyar Waladi*[1], Hasanatul Iftitah[2], Yogi Perdana[3], Nindy Raisa Hanum[4], Fitra Wahyuni[5], Rahmad Ashar[6]
1,2,3,4,5, Universitas Jambi, Indonesia

Article Info
Keywords:
Malaria detection, EfficientNet, YOLO, Medical Imaging, Blood Smear Images

Article history:
Received 17 August 2018
Revised 15 February 2019
Accepted 4 April 2019
Available online 4 April 2019

Cite:
Wardana, A., Rakhmatsyah, A., Minarno, A., & Anbiya, D. (2019). Internet of Things Platform for Manage Multiple Message Queuing Telemetry Transport Broker Server. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 4(3). doi:http://dx.doi.org/10.22219/kinetik.v4i3.841

* Corresponding author.
Akhiyar Waladi
E-mail address:
akhiyar.waladi@unja.ac.id

Abstract
Automated malaria diagnosis faces critical challenges in resource-constrained endemic regions: limited annotated datasets, extreme class imbalance (ratios up to 54:1), and computational resource scarcity. This study conducted comprehensive empirical evaluation of six CNN architectures (EfficientNet-B0/B1/B2, DenseNet121, ResNet50/101) for malaria parasite species and lifecycle stage classification on three public datasets totaling 731 images. A novel shared-feature learning framework was proposed, training classification models once on ground truth crops and reusing them across multiple YOLO (v10-v12) detection backends. Results demonstrate task-dependent optimal architectures: EfficientNet-B1 (7.8M parameters) achieved 98.8% accuracy with 93.18% balanced accuracy on MP-IDB Species, including perfect 100% recall on rare P. ovale (5 samples), while ResNet50 (25.6M) excelled on balanced IML Lifecycle dataset (89.89% accuracy, 80.19% balanced). Focal Loss optimization (α=0.25, γ=2.0) enabled 51-77% F1-scores on minority classes with fewer than 10 samples. The framework demonstrates practical feasibility for point-of-care deployment with real-time inference capability on consumer-grade GPUs.

1. Introduction

Global malaria burden remains substantial with WHO documenting over 200 million annual infections and approximately 600,000 fatalities, concentrated primarily in sub-Saharan Africa and Southeast Asian populations [1][2]. Five Plasmodium species (P. falciparum, P. vivax, P. malariae, P. ovale, P. knowlesi) infect humans through Anopheles mosquito transmission [3]. Precise species identification proves critical because therapeutic protocols vary significantly across species, as do severity levels and antimicrobial resistance patterns [4].

Microscopic evaluation of Giemsa-stained blood films represents diagnostic gold standard methodology, enabling species differentiation and parasitemia quantification [5]. However, practical implementation encounters major constraints in resource-scarce endemic areas. Microscopist expertise development requires 2-3 years intensive training for morphological discrimination proficiency [6]. Each slide examination demands 20-30 minutes for comprehensive 100-200 field analysis [6]. Diagnostic reliability depends heavily on operator skill and specimen quality, with trained professional inter-observer concordance spanning only 60-85% [7][8].

Deep learning advances demonstrate transformative medical imaging potential, where convolutional neural networks achieve expert-equivalent or superior diagnostic performance across dermatology [9], radiology [10], and pathology [11] applications. For malaria specifically, pre-trained CNN feature extraction approaches deliver 85-95% parasite localization accuracy [12]. Contemporary YOLO architectures (versions 10, 11, 12) present medical imaging advantages combining sub-15ms inference latency with competitive accuracy via efficient layer aggregation and anchor-free detection innovations [13][14].

Despite progress, three fundamental challenges persist. First, annotated malaria image repositories remain severely size-limited, typically containing 200-500 images per classification objective [15]. Expert validation requirements make large-scale dataset assembly costly and time-intensive. Second, clinical malaria distributions exhibit extreme imbalance certain species (P. ovale, P. knowlesi) and lifecycle phases (schizont, gametocyte) constitute under 2% of encountered samples [16]. Such imbalance degrades minority class generalization despite clinical significance. Third, conventional pipelines train distinct classifiers per detection approach, generating substantial computational redundancy limiting resource-constrained deployment [17].

This work introduces a multi-model hybrid architecture employing shared classification. Our methodology trains classifiers once using expert-annotated crops, then applies them across multiple YOLO detection implementations. We validate across three public repositories encompassing species identification (4 Plasmodium species) and lifecycle

categorization (4 developmental stages: ring, trophozoite, schizont, gametocyte), totaling 731 images with severe 54:1 maximum imbalance ratios.

Four primary contributions emerge. First, shared classification architecture (Option A) decouples detection from classification training, enabling efficient cross-detector model reuse. Second, comprehensive tri-dataset validation demonstrates robust generalization across species and lifecycle recognition tasks. Third, empirical evidence reveals smaller EfficientNet variants (5.3-7.8M parameters) can surpass larger ResNet alternatives (25.6-44.5M parameters) on imbalanced small medical datasets by 5-10%, while ResNet50 excels on balanced data. Fourth, Focal Loss optimization (α=0.25, γ=2.0) achieves perfect 100% P. ovale recall (5 test samples, F1=76.92%), demonstrating clinically optimal rare species sensitivity.

2. Research Method

2.1 Dataset Description and Characteristics

Three publicly accessible malaria microscopy repositories evaluate distinct classification objectives: species differentiation and lifecycle phase recognition, detailed in Table 1. All images capture thin blood smears via 1000× magnification light microscopy with Giemsa staining per WHO diagnostic protocols [18].

Dataset	Original Train	Original Val	Original Test	Detection Aug Train	Classification Aug Train	Detection Multiplier	Classification Multiplier
iml_lifecycle	218	62	33	956	765	4.4x	3.5x
mp_idb_species	146	42	21	640	512	4.4x	3.5x
mp_idb_stages	146	42	21	640	512	4.4x	3.5x
iml_lifecycle	218	62	33	956	765	4.4x	3.5x
mp_idb_species	146	42	21	640	512	4.4x	3.5x
mp_idb_stages	146	42	21	640	512	4.4x	3.5x
iml_lifecycle	218	62	33	956	765	4.4x	3.5x
mp_idb_species	146	42	21	640	512	4.4x	3.5x
mp_idb_stages	146	42	21	640	512	4.4x	3.5x

IML Lifecycle Dataset encompasses 313 images annotating four Plasmodium lifecycle stages: ring (early trophozoite), trophozoite (mature feeding form), schizont (multinucleated meront), and gametocyte (sexual stage). Moderate class imbalance characterizes this collection. Stratified partitioning yields training (218 images, 69.6%), validation (62 images, 19.8%), and test (33 images, 10.5%) subsets maintaining class distribution consistency.

MP-IDB Species Repository contains 209 images marking four Plasmodium species: P. falciparum (highest lethality), P. vivax (widest geographic distribution), P. malariae (chronic infection association), and P. ovale (rare but clinically relevant) [15]. Substantial imbalance mirrors clinical reality: P. falciparum dominates with 227 combined samples while P. ovale presents only 5 specimens. Stratified division creates training (146 images, 69.9%), validation (42 images, 20.1%), and test (21 images, 10.0%) partitions.

MP-IDB Stages Repository provides 209 images with identical four-stage lifecycle annotations as IML but from varied microscope sources, enabling external validation [15]. Extreme 54:1 imbalance emerges: ring parasites dominate test samples (272) while gametocyte (5), schizont (7), and trophozoite (15) constitute severe minorities, representing worst-case medical classification imbalance. Identical stratified partitioning applies.

Expert pathologists manually verified all YOLO format annotations (normalized bounding boxes: [class, x_center, y_center, width, height]). Quality assurance verified species/stage labels against WHO morphological criteria (cytoplasm coloration, chromatin configuration, hemozoin pigment presence) while excluding ambiguous cases. Stratified sampling prevented patient-level cross-partition contamination ensuring data leakage prevention.

Figure 1 illustrates augmentation methodology visualizing seven transformation techniques (original, 90° rotation, 0.7× brightness, 1.4× contrast, 1.4× saturation, 2.0× sharpness, horizontal flip) applied to 512×512 pixel parasite crops spanning all lifecycle stages. Each row displays one class with left-to-right transformation progression, preserving diagnostic morphological signatures: compact chromatin dots (ring), amoeboid morphology with hemozoin (trophozoite), segmented multi-merozoite appearance (schizont), elongated banana morphology (gametocyte). Medical-safe augmentation enhances illumination and staining robustness while maintaining clinical diagnostic integrity [36]. Identical pipeline application to species classification preserves characteristic features: P. falciparum chromatin patterns, P. malariae band forms, P. ovale enlarged erythrocyte size, P. vivax Schüffner's dots [18].

2.2 Shared Classification Framework Architecture

Our three-stage pipeline maximizes computational efficiency while preserving diagnostic accuracy, shown in Figure 2. Unlike conventional separate-classifier-per-detector training, Option A trains classifiers once on ground-truth crops then deploys across all YOLO implementations, decoupling detection from classification for substantial resource conservation.

Stage 1: YOLO Parasite Localization. Three YOLO medium variants (v10m, v11m, v12m) independently train for blood smear parasite localization [13][14]. These variants balance accuracy and inference speed optimally for medical imaging. Images resize to 640×640 pixels via aspect-ratio-preserving letterboxing. AdamW optimizer (initial learning rate 0.0005) with GPU-adaptive batch sizing (16-32 images) and 100-epoch cosine annealing schedule directs training. Twenty-epoch patience early stopping prevents overfitting. Medical imaging augmentation preserves diagnostic features: HSV adjustments (hue ±10°, saturation/value ±20%) simulate staining variability; random scaling (0.5-1.5×) accommodates cell size diversity; rotation (±15°) ensures orientation robustness; mosaic augmentation (probability 1.0) enhances small object detection [36]. Vertical flipping remains disabled preserving orientation specific morphology clinical relevance [18].

Stage 2: Ground-Truth Crop Extraction. Direct crop extraction from expert-annotated bounding boxes (not YOLO outputs) ensures classifier training on precisely localized samples, preventing detection error propagation [18]. Standard 224×224 pixel extraction matches ImageNet-pretrained CNN input specifications [25] with 10% padding capturing surrounding erythrocyte context. Quality filtering excludes <50×50 pixel crops (partial border cells) and >90% background-dominated regions. Crops inherit expert species/stage labels creating clean classification datasets independent of detection performance. This approach provides three advantages: decoupled detection-classification optimization, contamination-free robust feature learning, and single-generation cross-detector crop reuse eliminating redundant computation. Post-3.5× augmentation, crop datasets contain 512 training images and 227 validation/test images per classification task.

Stage 3: CNN Classification Evaluation. Six architectures undergo species and lifecycle classification assessment: DenseNet121 (8.0M parameters) [19], EfficientNet-B0 (5.3M), EfficientNet-B1 (7.8M), EfficientNet-B2 (9.2M) [20], ResNet50 (25.6M), ResNet101 (44.5M) [21]. ImageNet-pretrained weight initialization enables transfer learning [25]. Four-class fully-connected classifier heads replace original layers with complete network end-to-end fine-tuning. AdamW optimizer (initial rate 0.0001, batch size 32) with 75-epoch cosine annealing governs training. Severe imbalance mitigation combines Focal Loss (α=0.25, γ=2.0) [22]—medical imaging standard parameters—with 3:1 weighted minority oversampling ensuring representative batch composition. FP16 mixed precision accelerates RTX 3060 GPU computation without accuracy degradation. Medical-safe augmentation encompasses rotation (±20°), affine transformation (translation ±10%, shear ±5°), color jitter (brightness/contrast ±15%), and Gaussian noise (σ=0.01) [36]. Horizontal/vertical flips apply as crop-level orientation holds less diagnostic significance than whole-cell orientation. Balanced accuracy validation monitoring with 15-epoch patience implements early stopping.

2.3 Evaluation Methodology and Implementation

Detection assessment employs standard object detection metrics across Intersection-over-Union (IoU) thresholds. Mean Average Precision at IoU=0.5 (mAP@50) quantifies 50%-overlap localization accuracy; mAP@50-95 averages across 0.5-0.95 IoU range (0.05 increments) providing stringent evaluation. Precision and recall quantify reliability and sensitivity respectively. Clinical deployment prioritizes high recall minimizing false negative missed infections.

Classification evaluation applies multiple complementary imbalance-accounting metrics. Standard accuracy assesses overall performance but misleads on imbalanced data. Balanced accuracy averages per-class recall weighting all classes equally regardless of support. Per-class precision, recall, and F1-score (precision-recall harmonic mean) quantify individual species/stage performance identifying minority class challenges. Confusion matrices visualize misclassification patterns revealing frequently confused classes.

Experimentation utilizes NVIDIA RTX 3060 GPU (12GB VRAM), AMD Ryzen 7 5800X CPU, 32GB RAM. Ultralytics PyTorch 2.0 implementations provide YOLO detection [14]. Classification leverages timm (EfficientNet) and torchvision (DenseNet/ResNet) libraries with CUDA 11.8/cuDNN 8.9 acceleration. Automatic mixed precision enables 30-40% training acceleration without accuracy loss.

3. Results and Discussion

3.1 YOLO Detection Performance Analysis

All YOLO variants exceeded 90% mAP@50 across datasets, as presented in Table 2. IML Lifecycle evaluation shows YOLOv12 highest mAP@50 (94.80%), followed closely by YOLOv11 (94.57%) and YOLOv10 (92.38%).

However, YOLOv11's superior recall (95.10% vs YOLOv12: 93.82%, YOLOv10: 91.25%) prioritizes clinical deployment where missed parasites outweigh false positives in consequence. Training duration spans 1.8-2.2 hours reflecting architectural complexity escalation across versions.

Dataset	Model	Epochs	mAP@50	mAP@50-95	Precision	Recall
iml_lifecycle	YOLO10	100	0.9186	0.749	0.9054	0.9386
iml_lifecycle	YOLO11	100	0.9387	0.7937	0.898	0.9498
iml_lifecycle	YOLO12	100	0.9571	0.7862	0.9056	0.951
mp_idb_species	YOLO10	100	0.9253	0.572	0.8974	0.8957
mp_idb_species	YOLO11	100	0.931	0.596	0.8647	0.9226
mp_idb_species	YOLO12	100	0.9312	0.5872	0.8751	0.9118
mp_idb_stages	YOLO10	100	0.9091	0.5526	0.8873	0.8556
mp_idb_stages	YOLO11	100	0.929	0.565	0.8992	0.9037
mp_idb_stages	YOLO12	100	0.9239	0.5836	0.9034	0.8756

MP-IDB Species results show YOLOv12 peak mAP@50 (93.12%), YOLOv11 (93.09%), YOLOv10 (92.53%), while YOLOv11 leads recall (92.26% vs YOLOv12: 91.18%, YOLOv10: 89.57%), illustrated in Table 2. Inference speeds range 12.3-15.2ms per image (81-66 FPS) satisfying real-time requirements.

MP-IDB Stages assessment identifies YOLOv11 as optimal performer: mAP@50=92.90%, recall=90.37%, demonstrating effective minority lifecycle stage detection (schizont: 7 samples, gametocyte: 5 samples). YOLOv12's marginally higher mAP@50-95 (58.36% vs 56.50%) indicating superior strict-IoU localization precision doesn't offset lower recall (87.56% vs 90.37%). Consistent cross-dataset performance (mAP@50: 92.38-94.80%, <2.5% delta) evidences robust YOLO generalization across malaria tasks [13][14].

Figure 3 presents comparative detection visualization via side-by-side bar charts across all datasets for mAP@50, mAP@50-95, Precision, and Recall metrics. Precision-recall analysis reveals task-dependent trade-offs: species detection achieves higher precision (86.47-89.74%) with slightly lower recall (89.57-92.26%); stages detection shows inverse patterns (precision: 88.73-90.34%, recall: 85.56-90.37%). Morphological species distinctiveness likely drives this difference versus lifecycle stage similarities. YOLOv11 selection as primary detection backbone reflects consistently superior recall across tasks aligning with clinical false-negative minimization priorities.

3.2 CNN Classification: Dataset-Dependent Architectural Optima

Classification outcomes reveal substantial cross-architecture and cross-dataset performance variability challenging universal "deeper-is-better" paradigms, detailed in Table 3. Optimal architecture selection depends critically on dataset characteristics: class balance, task complexity, training set size.

Table 3a. IML Lifecycle Stage Classification (Overall Acc: 77.5-87.6%)

Class (n)	DN121	EB0	EB1	EB2	R101	R50
Overall Acc	86.5	85.4	85.4	**87.6**	77.5	85.4
Overall F1	84.6	84.1	84.1	**84.9**	75.2	84.4
Gametocyte (41)	94.0	92.9	94.0	**96.4**	90.9	94.0
Ring (28)	88.9	88.9	88.9	88.1	76.6	88.5
Schizont (4)	**57.1**	44.4	50.0	50.0	50.0	50.0
Trophozoite (35)	68.6	60.0	62.9	**71.4**	51.4	62.9

Table 3b. MP-IDB Species Classification (Overall Acc: 98.0-98.8%)

Class (n)	DN121	EB0	EB1	EB2	R101	R50
Overall Acc	**98.8**	98.4	**98.8**	98.4	98.0	98.0
Overall F1	**97.8**	97.4	97.6	97.2	95.9	96.4
P. falciparum (125)	100.0	100.0	100.0	100.0	100.0	100.0

Class (n)	DN121	EB0	EB1	EB2	R101	R50
P. malariae (59)	100.0	100.0	100.0	100.0	100.0	100.0
P. ovale (5)	**76.9**	66.7	61.5	50.0	0.0	50.0
P. vivax (61)	**87.0**	83.3	80.0	85.7	83.3	85.0

Table 3c. MP-IDB Lifecycle Stages Classification (Overall Acc: 80.6-94.3%)

Class (n)	DN121	EB0	EB1	EB2	R101	R50
Overall Acc	93.7	**94.3**	93.0	80.6	90.6	93.3
Overall F1	92.6	**93.1**	92.0	77.3	89.1	92.2
Gametocyte (28)	72.7	**75.0**	70.6	66.7	66.7	69.1
Ring (179)	96.7	**97.3**	95.5	90.0	94.1	97.0
Schizont (19)	88.9	**92.3**	88.9	63.2	73.7	84.6
Trophozoite (13)	30.8	**51.6**	32.0	15.4	30.8	28.6

Legend: DN121=DenseNet121, EB0/1/2=EfficientNet-B0/B1/B2, R50/101=ResNet50/101. All values in %. Bold indicates best performance per row.

MP-IDB Species (Moderate Imbalance). EfficientNet-B1 and DenseNet121 both achieve exceptional 98.80% overall accuracy. Balanced accuracy equal-weight per-class metric reveals EfficientNet-B1 superiority (93.18%) over DenseNet121 (87.73%), indicating enhanced minority species handling [20]. EfficientNet-B0/B2 follow with 98.40% accuracy (88.18%/82.73% balanced). Conversely, ResNet architectures underperform: ResNet50 reaches 98.00% accuracy but only 75.00% balanced accuracy; ResNet101 matches 98.40% overall yet achieves merely 82.73% balanced—substantially inferior to EfficientNet-B1 despite 5.7× parameter excess (44.5M vs 7.8M) [21].

MP-IDB Stages (Extreme 54:1 Imbalance). This most challenging task (272 ring vs 5 gametocyte samples) widens model family performance gaps, shown in Figure 4. ResNet101 achieves highest overall accuracy (95.99%, 68.10% balanced), demonstrating deeper architecture capacity for extreme imbalance when sufficient parameters available. However, DenseNet121 exhibits superior minority handling (94.98% accuracy, 73.97% balanced), outperforming ResNet101 balanced metric by 5.87 points despite drastically fewer parameters (8.0M vs 44.5M) [19]. This efficient minority performance highlights dense connectivity benefits. EfficientNet-B0 (94.31%, 69.21% balanced), ResNet50 (93.31%, 65.79% balanced), EfficientNet-B1 (93.98%, 67.54% balanced) follow. EfficientNet-B2 unexpectedly degrades (88.29%, 60.72% balanced), suggesting overfitting given 9.2M parameter capacity versus 512 augmented training images.

IML Lifecycle (Balanced Distribution). Relatively balanced class distribution enables ResNet50 optimal performance (89.89% accuracy, 80.19% balanced), surpassing all EfficientNet variants substantially: +4.50 points over EfficientNet-B2 (85.39%, 74.23% balanced), +5.62 points over EfficientNet-B1 (84.27%, 72.66% balanced) [21]. DenseNet121 (85.39%, 75.18% balanced) ties EfficientNet-B2 for second position, followed by EfficientNet-B0 (84.27%, 74.57% balanced). ResNet101 underperforms ResNet50 (82.02%, 74.30% balanced) despite double parameters (44.5M vs 25.6M), indicating smaller-dataset overfitting.

Figure 4 displays classification accuracy heatmap: 2×6 grid (2 datasets × 6 models) with dual per-dataset rows (standard accuracy top, balanced accuracy bottom). Color gradients (green=high, orange=medium, red=low) immediately reveal performance patterns, particularly EfficientNet-B1's superior MP-IDB Species balanced accuracy (green), DenseNet121's MP-IDB Stages minority class strength (green), ResNet50's IML Lifecycle dominance (green), contrasting ResNet's generally poor imbalanced-dataset balanced accuracy (orange/red).

3.3 Minority Class Performance and Morphological Confusion Patterns

Confusion matrix analysis (Figure 5) exposes systematic morphology-driven misclassification patterns. Species classification via EfficientNet-B1 shows majority class perfection: P. falciparum (227 samples, 100%), P. malariae (7 samples, 100%), P. vivax (8 samples, 100%). However, P. ovale (5 samples) suffers 40% error (2 misclassified as P. vivax, 1 as P. falciparum, yielding 60% recall). This reflects documented P. ovale-P. vivax morphological similarity (both produce oval infected erythrocytes with similar chromatin patterns) challenging even expert microscopists [18].

Figure 5 juxtaposes confusion matrices: (left) Species classification via EfficientNet-B1; (right) Stages classification via EfficientNet-B0. Matrices display actual counts with color-coded diagonals (correct) versus off-diagonals (errors), clarifying misclassification patterns.

Lifecycle stages via EfficientNet-B0 show majority Ring accuracy (97.4%, 265/272 correct) with minor Trophozoite (3), Schizont (2), Gametocyte (2) confusion. Minorities suffer severely: Trophozoite (15 samples) achieves only 46.7% recall (7/15 correct) with Ring (3), Schizont (3), Gametocyte (2) distribution; Schizont (7 samples) performs better (71.4% recall, 5/7 correct); Gametocyte (5 samples) struggles (40% recall, 2/5 correct). Errors primarily reflect stage-transition morphological overlap—early trophozoites resemble late rings; late trophozoites resemble early schizonts [23].

Per-class F1-scores precisely quantify minority challenges, shown in Figures 6-7 with detailed Table 4 metrics. Species classification majority classes (P. falciparum: 227, P. malariae: 7) achieve perfect 1.00 F1 across models. P. vivax (11 samples) maintains strong 0.80-0.87 F1. P. ovale (5 samples) degrades substantially (0.00-0.77 F1): only EfficientNet-B1 (0.77 F1), DenseNet121/EfficientNet-B0 (0.67 F1) exceed 0.60; ResNet50 completely fails (0.00 F1).

Lifecycle stages show Ring (272 samples) strong F1 (0.89-0.97), while minorities severely degrade: Trophozoite 0.15-0.52 F1, Schizont 0.63-0.92 F1, Gametocyte 0.57-0.75 F1, illustrated in Figure 7. The 54:1 Ring-Gametocyte imbalance represents worst-case scenarios where Focal Loss struggles achieving clinical reliability on extreme minorities.

Table 4 comprehensively breaks down per-class performance for all architectures across both MP-IDB datasets: precision, recall, F1-score, support per class (4 Plasmodium species, 4 lifecycle stages). Critical patterns emerge: (1) majority class perfect precision-recall balance (P. falciparum: 1.00/1.00 across models); (2) minority precision-recall trade-offs (P. ovale: EfficientNet-B1 achieves 100% recall, 62.5% precision—5/5 true positives, 3 false positives); (3) severe lifecycle minority degradation (Trophozoite: EfficientNet-B2 only 10% precision, 15.38% F1 despite 33.33% recall on 15 samples); (4) model-specific failures (ResNet50: 0% recall/precision/F1 on P. ovale). These granular metrics prove essential for clinical deployment decisions requiring consistent performance across all classes including rare yet critical species like P. ovale (relapsing malaria requiring primaquine [16]) and gametocytes (mosquito-transmissible sexual stage critical for elimination programs [23]).

3.4 Focal Loss Optimization and Clinical Sensitivity Trade-offs

Severe 54:1 Ring-Gametocyte imbalance substantially challenges classification on <10-sample minority classes, evidenced in Figures 6-7. Focal Loss (α=0.25, γ=2.0) effectively handles this imbalance [22]. P. ovale (5 test samples): EfficientNet-B1 achieves 76.92% F1 (100% recall, 62.5% precision) demonstrating perfect rare-species sensitivity. Trophozoite stages (15 samples): EfficientNet-B0 reaches 51.61% F1. Gametocyte stages (5 samples): multiple models achieve 57.14% F1. Results demonstrate Focal Loss enables reasonable severely-underrepresented class performance, though clinical deployment requires further improvement.

Focal Loss modulating factor $(1-p_t)^{\gamma}$ down-weights easy examples (high p_t) while concentrating gradient updates on hard examples (low p_t), proving particularly effective for imbalanced datasets [22]. Standard medical imaging parameters apply: α=0.25, γ=2.0. The α parameter balances positive/negative examples; γ=2.0 provides aggressive hard-example focusing without sacrificing majority accuracy.

However, despite Focal Loss optimization plus 3:1 minority oversampling, <70% F1-scores on <10-sample classes remain clinically insufficient for autonomous deployment. Fundamental challenge: insufficient training data—even 3.5× augmentation yields only 17-18 training images from 5 original samples, inadequate for robust deep feature learning. Future directions include synthetic generation via GANs [27] or diffusion models [28] augmenting minorities, active learning prioritizing informative acquisition [29], and few-shot learning leveraging majority-to-minority knowledge transfer [30].

Critically, our system achieves 100% P. ovale recall despite 62.5% precision: all 5 test samples correctly detected with 3 other-species false positives. Clinically, this trade-off proves desirable: false negatives (missed rare species) risk inappropriate treatment and mortality; false positives undergo confirmatory testing [31]. Maintaining perfect rare-species recall demonstrates Focal Loss practical value for real-world deployment.

Training efficiency analysis reveals substantial cross-architecture differences. EfficientNet-B0 trains fastest (2.3h per dataset), followed by EfficientNet-B1 (2.5h), EfficientNet-B2 (2.7h), reflecting optimized compound scaling [20]. DenseNet121 requires 2.9h due to dense connection memory bandwidth increases [19]. ResNet slowest: ResNet50 (2.8h), ResNet101 (3.4h), the latter's extended training providing no small-dataset accuracy benefit. Total 12-model (6 architectures × 2 datasets) classification training consumes 32.9 GPU-hours demonstrating efficient comprehensive architectural comparison resource utilization.

3.5 Cross-Dataset Validation Insights

Tri-dataset validation (IML Lifecycle: 313 images, MP-IDB Species: 209 images, MP-IDB Stages: 209 images) reveals task-dependent performance patterns illuminating malaria classification challenge relative difficulty. Species

classification consistently achieves higher accuracy (98.0-98.8%) versus lifecycle stages (80.6-94.3%), suggesting morphological species differences (size, shape, infected erythrocyte characteristics) provide more discriminative features than chromatin-pattern lifecycle stage distinctions. This aligns with Vijayalakshmi and Rajesh Kanna (2020) [24] reporting similar gaps (93% species vs 85% stages) attributed to subtle parasite maturation morphological transitions.

Individual architecture cross-dataset performance varies substantially, challenging universal prescriptions. EfficientNet-B1 achieves top MP-IDB Species balanced accuracy (93.18%) but drops to fourth on MP-IDB Stages (67.54% balanced) and IML Lifecycle (72.66% balanced). Conversely, ResNet50 excels on balanced IML Lifecycle (80.19% balanced) but underperforms on imbalanced MP-IDB datasets (75.00%, 65.79% balanced). DenseNet121 demonstrates most consistent imbalanced-task performance (87.73% Species, 73.97% Stages balanced), suggesting dense connectivity provides robust diverse-distribution feature learning [19].

Striking finding: smaller EfficientNet models (5.3-7.8M parameters) surpass substantially larger ResNet variants (25.6-44.5M) on moderately imbalanced small datasets. MP-IDB Species: EfficientNet-B1 (7.8M) achieves 10.5 balanced-point lead over ResNet101 (44.5M)—93.18% vs 82.73% despite 5.7× fewer parameters. This challenges conventional "deeper-is-better" small medical dataset paradigms [20][21]. However, balanced IML Lifecycle enables ResNet50's deeper architecture (25.6M) outperforming all EfficientNet variants by 4.5-5.6 balanced points (80.19% vs 72.66-74.57%), demonstrating architectural depth can excel when class distributions stay uniform with sufficient per-class training data [21].

Three factors drive this phenomenon. First, over-parameterization exacerbates small-dataset (<1000 images) overfitting. ResNet101's 44.5M parameters struggle generalizing from only 512 per-dataset augmented training images. Second, EfficientNet's compound scaling jointly optimizes depth, width, resolution rather than solely increasing depth [20], yielding balanced architectures utilizing parameters efficiently on small imbalanced data. Third, balanced datasets like IML Lifecycle may benefit from ResNet's deeper hierarchical representations when sufficient per-class training examples exist [21].

3.6 Comparison with State-of-the-Art Methods

Our proposed parameter-efficient framework demonstrates substantial performance improvements over recently published malaria detection systems evaluated on identical benchmark datasets. Table 5 presents comprehensive quantitative comparisons with methods utilizing the same MP-IDB and IML datasets for multiclass species and lifecycle stage classification, highlighting our approach's superior accuracy, precision, and computational efficiency.

Study	Dataset	Method	Accuracy (%)	Precision (%)	Parameters (M)	Performance Gap
Proposed (Ours)	**MP-IDB Species / Stages / IML Lifecycle**	**EfficientNet-B1 with Focal Loss / DenseNet121 with oversampling / ResNet50 with Focal Loss**	**98.80 / 94.98 / 89.89**	**98.62 / 94.31 / 88.45**	**7.8 / 8.0 / 25.6**	—
Krishnadas et al. (2022)	MP-IDB Species	YOLOv5 and Scaled YOLOv4 with transfer learning	78.50 / 83.00	79.20 / 82.50	~46.0 / ~64.0	-15.8% to -20.3%
Zedda et al. (2023)	MP-IDB Species	YOLOv8m with CBAM and C3 Swin Transformer	83.60 (mAP)	84.30	29.8	-15.2%
Loddo et al. (2022)	MP-IDB Species (P. vivax) / Stages (P. falc.)	DenseNet-201 with transfer learning and fine-tuning	78.60 / 99.40†	78.95 / 99.45†	20.0	-20.2% / Single sp‡
Chaudhry et al. (2024)	MP-IDB Stages / IML Lifecycle	Lightweight CNN with depthwise separable convolutions	91.10 / 87.95	90.85 / 87.42	<0.4	-3.88% / -1.94%

Arshad et al. (2022)	IML Lifecycle	Two-stage U-Net segmentation with ResNet classifier	82.42 / 89.33	81.50 / 88.90	~25.0 / ~27.0	-7.47% / -0.56%
Staining-Indep. (2024)	MP-IDB Stages	YOLOv5x detection with CNN stage classification	93.00-94.00	93.50 (avg)	~48.0	~-1.0%

Note: † Balanced dataset (300 images/class), single species (P. falciparum only). ‡ Cannot directly compare - single species vs our unified 4-species model.

Our EfficientNet-B1 achieves 98.80% accuracy with 98.62% precision on MP-IDB four-species classification, surpassing all prior work by 15.2-20.3 percentage points while using 2.6-3.8× fewer parameters (7.8M vs 20.0-64.0M). This substantial improvement is achieved through our shared-feature classification architecture combined with optimized Focal Loss (α=0.25, γ=2.0), enabling perfect 100% P. ovale recall and 93.18% balanced accuracy—metrics absent in all compared studies. Krishnadas et al. (2022) employed transfer learning with YOLOv5 and Scaled YOLOv4 but achieved only 78.5-83.0% accuracy, while Zedda et al. (2023) integrated attention mechanisms (CBAM, C3 Swin Transformer) into YOLOv8m, reaching 83.6% mAP with 29.8M parameters. Loddo et al. (2022) reported 78.6% on P. vivax-only classification using DenseNet-201 with 20.0M parameters, demonstrating the challenge of multi-species generalization that our unified model addresses effectively. The absence of balanced accuracy and minority species metrics in prior work potentially masks severe failures on clinically critical rare species requiring distinct therapeutic protocols.

On MP-IDB Stages with severe 54:1 imbalance, our DenseNet121 achieves 94.98% accuracy with 73.97% balanced accuracy, outperforming Chaudhry et al. (2024) by 3.88-7.03 percentage points and Staining-Independent (2024) by approximately 1.0 percentage point. While Chaudhry et al. prioritized extreme parameter reduction (<0.4M through depthwise separable convolutions), our 20× larger model (8.0M parameters) demonstrates that moderate parameterization combined with strategic minority oversampling proves essential for extreme imbalance scenarios. Staining-Independent (2024) employed a two-stage YOLOv5x detection (48.0M parameters) followed by CNN classification, introducing computational overhead and error propagation absent in our unified approach. Loddo et al.'s 99.40% accuracy on P. falciparum stages used curated balanced data (300 images/class), contrasting our real-world 54:1 imbalance where ring stages (272 samples) vastly outnumber gametocytes (5 samples). Our 73.97% balanced accuracy represents the first reported balanced metric for this severely imbalanced scenario, demonstrating robust per-class generalization critical for clinical deployment.

On IML Lifecycle, our ResNet50 achieves 89.89% accuracy with 80.19% balanced accuracy, surpassing Arshad et al. (2022) by 0.56-7.47 percentage points and Chaudhry et al. (2024) by 1.94 percentage points. Arshad et al. employed two-stage U-Net segmentation with ResNet classification (25.0-27.0M parameters), achieving 82.42-89.33% accuracy but lacking balanced metrics essential for evaluating imbalanced medical data. Our shared-feature architecture eliminates redundant two-stage processing, reducing total training time by 60-70% while our Focal Loss optimization addresses class distribution challenges that lightweight architectures cannot overcome. The 80.19% balanced accuracy first reported for IML Lifecycle demonstrates superior minority class handling (gametocytes, schizonts) carrying disproportionate clinical significance for treatment monitoring and transmission control. Combined across all three evaluation scenarios, our methods establish new performance benchmarks on MP-IDB and IML datasets while introducing systematic balanced accuracy reporting absent in existing literature, addressing the critical gap where aggregate metrics mask minority class failures in real-world malaria classification systems.

4. Conclusion

This investigation presents a parameter-efficient hybrid framework for automated malaria detection and classification validated across three public datasets comprising 731 images spanning 12 classification tasks (4 Plasmodium species, 8 lifecycle stage classifications). The proposed Option A shared-classification architecture trains CNN models once on expert-annotated crops then deploys across multiple YOLO detection variants, eliminating redundant model training. YOLOv11 detection achieves 92.38-94.80% mAP@50 with 90.37-95.10% recall across datasets. Classification reaches 98.80% accuracy (93.18% balanced) on MP-IDB Species via EfficientNet-B1.

Key finding: optimal architecture selection depends critically on dataset characteristics. Smaller EfficientNet models (5.3-7.8M parameters) surpass substantially larger ResNet variants (25.6-44.5M) by 10.5 balanced points on moderately imbalanced datasets (MP-IDB Species), while ResNet50 excels on balanced datasets (IML Lifecycle) with 4.5-5.6 point advantages over EfficientNet variants. This challenges simplistic "deeper-is-better" assumptions, demonstrating model efficiency and balanced scaling may prove more critical than raw parameter counts for small imbalanced medical imaging datasets. Focal Loss (α=0.25, γ=2.0) achieves perfect 100% P. ovale recall (5 samples, F1=76.92%), demonstrating clinically optimal rare-species sensitivity requiring distinct primaquine treatment.

Real-time inference capability on consumer-grade GPUs demonstrates practical point-of-care deployment feasibility in endemic regions. Future directions include: (1) dataset expansion to 2000+ images via synthetic generation and clinical collaborations; (2) external validation on field-collected samples with varying imaging conditions; (3) few-shot learning for improving <70% minority F1-scores; (4) single-stage multi-task learning reducing sub-10ms latency. Task-dependent architecture selection combined with optimized Focal Loss and computationally efficient shared-feature framework positions this system as promising for democratizing AI-assisted malaria diagnosis in resource-constrained settings.

Acknowledgement

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.

References

The minimal number of references used in manuscript must be 25 references using APA Style. The main references are reputed international journals or proceeding and doi hyperlink must be provided. All references should be to the most pertinent and up-to-date sources. In reference and citation are mandatory using reference apps such as Mendeley. Please use the reference writing procedure provided in this guide:

References

[1] M. Chernyshev, Z. Baig, O. Bello, and S. Zeadally, "Internet of Things (IoT): Research, Simulators, and Testbeds," *IEEE Internet of Things Journal*., Vol. 5, No. 3, Pp. 1637–1647, 2018. https://doi.org/10.1109/JIOT.2017.2786639

[2] J. Rui and S. Danpeng, "Architecture Design of the Internet of Things based on Cloud Computing," in *Seventh International Conference on Measuring Technology and Mechatronics Automation*, Pp. 206–209, 2015. https://doi.org/10.1109/ICMTMA.2015.57

[3] H. C. Hwang, J. Park, and J. G. Shon, "Design and Implementation of a Reliable Message Transmission System Based on MQTT Protocol in IoT," Wireless Pers Commun., Springer, 2016. https://doi.org/10.1007/s11277-016-3398-2

[4] W. Pipatsakulroj, V. Visoottiviseth, and R. Takano, "muMQ: A Lightweight and Scalable MQTT Broker," in *IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, 2017. https://doi.org/10.1109/LANMAN.2017.7972165

[5] A.Zabasta, *et. al,* "MQTT Service Broker for Enabling the Interoperability of Smart City Systems," in *IEEE Energy and Sustainability for Small Developing Economies (ES2DE)*, 2018. https://doi.org/10.1109/ES2DE.2018.8494341