Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 2, May, Pp. 323-338

Parameter-efficient models for malaria detection and classification using small-scale imbalanced blood smear images

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

Article Info
Keywords:
Malaria Diagnosis, Multi-model Framework, Parasite Classification, Transfer Learning, Class Imbalance

Article history:
Received: November 01, 2025
Accepted: February 26, 2026
Published: May 01, 2026

Cite:
A. Waladi, H. Iftitah, N. R. Hanum, Y. Perdana, F. Wahyuni, and R. Ashar, "Parameter-Efficient Models for Malaria Detection and Classification Using Small-Scale Imbalanced Blood Smear Images", *KINETIK*, vol. 11, no. 2, May. 2026.
https://doi.org/10.22219/kinetik.v11i2.2558

*Corresponding author.
Hasanatul Iftitah
E-mail address:
hasanatul.iftitah@unja.ac.id

Abstract
Malaria diagnostic automation faces critical challenges, including severe class imbalance with ratios of up to 54:1, limited datasets containing 200 to 500 images, and computational inefficiency resulting from the need to train separate models for each detection-classification combination. This study developed a multi-model framework with a shared classification architecture that trains classification models once on ground-truth crops and reuses them across all detectors. The framework systematically evaluated three YOLO Medium architectures for parasite detection and six CNN architectures for lifecycle and species classification across four complementary malaria datasets totaling 1,544 microscopy images. Detection achieved mAP@50 scores ranging from 70.84% to 96.27%, with high recall values of 71.05% to 93.12% minimizing missed parasite detections. Classification results demonstrated the importance of dataset-dependent model selection, with parameter-efficient EfficientNet models containing 5.3M to 9.2M parameters consistently outperforming ResNet variants with up to 44.5M parameters. EfficientNet-B1 achieved accuracies of 91.51% on the IML Lifecycle dataset and 98.28% on the MP-IDB Species dataset, while EfficientNet-B0 achieved 86.45% on the multi-patient MD-2019 dataset. ResNet50 achieved 96.13% accuracy on severely imbalanced MP-IDB Stages dataset. Focal Loss optimization with alpha = 1.0 and gamma = 1.5 enabled robust minority-class performance, achieving F1-scores between 0.44 and 1.00 on ultra-minority classes and demonstrating effective handling of class imbalance. The compact models, with sizes ranging from 46 MB to 89 MB, enable practical deployment on resource-constrained hardware.

1. Introduction

1.1 Background and Motivation

Malaria remains a critical global health challenge, with approximately 263 million cases and 597,000 deaths reported in 2023 [1]. Caused by *Plasmodium* parasites transmitted through *Anopheles* mosquitoes, accurate species identification and lifecycle stage classification are essential for effective treatment. Different *Plasmodium* species and lifecycle stages respond differently to antimalarial drugs [2]. Misdiagnosis or delayed treatment can lead to severe complications, including cerebral malaria, organ failure, and death within 24-48 hours.

Traditional microscopy remains the diagnostic gold standard but faces significant limitations. Each slide requires the examination of over 100 microscopic fields by trained microscopists [3], creating severe bottlenecks in resource-limited endemic regions where expertise is scarce. With over 200 million annual cases concentrated in sub-Saharan Africa and Southeast Asia, this diagnostic bottleneck delays treatment for millions of patients, directly contributing to preventable mortality. These practical constraints have motivated the development of artificial intelligence approaches for automated malaria detection, enabling rapid and accurate diagnosis where expert microscopists are unavailable [4].

1.2 Literature Review

Several research groups have made substantial contributions toward automating malaria diagnosis through deep learning, establishing important baselines on publicly available datasets while revealing persistent challenges in the field. Arshad et al. [5] introduced the IML Lifecycle dataset and demonstrated two-stage detection-classification pipelines for *P. vivax* lifecycle staging. Loddo et al. [6] provided the first deep learning baseline for lifecycle stage classification on MP-IDB, comparing eleven CNN architectures. Zedda et al. [7] applied YOLO-based object detection on MP-IDB, showing that single-stage detectors could match classification-only approaches, and subsequently developed YOLO-PAM [8] with attention mechanisms to reduce parameters while maintaining accuracy. Sukumarran et al. [9] further validated modern detection frameworks through a comparative evaluation of YOLOv4 and YOLOv5 alongside DenseNet-121 for species identification.

Table 1 summarizes the methodological scope, key results, and identified limitations of these existing approaches. Three critical gaps remain unaddressed. First, limited dataset sizes ranging from 200 to 500 images constrain model generalization, with most studies evaluating a single dataset rather than assessing cross-dataset robustness [5], [10]. Second, extreme class imbalance with ratios of up to 54:1 causes models to underperform on clinically significant minority classes, yet prior work predominantly reports overall accuracy metrics that mask failures on rare stages [11], [12]. Third, existing pipelines train separate classification models for each detection method, creating computational redundancy that exceeds practical deployment constraints in resource-limited settings [13].

None of the reviewed studies simultaneously address all three gaps. This study addresses these limitations through a shared classification architecture with systematic multi-model evaluation across four complementary datasets, Focal Loss optimization for extreme class imbalance, and parameter-efficient model selection for resource-constrained deployment.

Table 1. Summary of Related Work: Methodology, Contributions, and Identified Gaps

Study	Method	Dataset	Key Contribution	Limitation / Gap
[5]	Det: Morphological segmentation + watershed Cls: ResNet50V2	IML (345 imgs, 4 P. vivax stages)	Introduced IML dataset with bbox annotations; two-stage det-cls pipeline	Single species; small dataset; no cross-dataset eval; no imbalance handling
[6]	Cls only: 11 CNNs evaluated (best: DenseNet-201, Acc 99.40%)	MP-IDB + NIH (209 + 27,558 imgs)	First DL baseline on MP-IDB; comprehensive 11 CNN comparison	No detection component; single species (P. falciparum); no end-to-end workflow
[7]	Det: YOLOv5 Cls: DarkNet-53 (Acc 96.02%)	MP-IDB (209 imgs, P. falciparum)	First YOLO detection on MP-IDB; single-stage det matches cls-only pipelines	Single dataset; no imbalance handling; separate det and cls models
[8]	Det+Cls: YOLO-PAM (YOLOv8 + NAM/CBAM attention) mAP@50: 91.8% IML, 83.6% MP-IDB	IML + MP-IDB (522 imgs, 2 species)	Attention mechanisms for malaria det; 11M fewer params vs YOLOv8	Detection-only eval; specialized arch limits reproducibility; no imbalance handling
[9]	Det: YOLOv4/v5 (mAP@50: 96%) Cls: DenseNet-121 (Acc 95.5%)	IML + MP-IDB (522 imgs, 2 species)	YOLOv4 vs v5 comparison; two-stage det then species cls	Fixed cls arch (no multi-model comparison); no imbalance handling; per-detector training

1.3 Proposed Solution

This study introduces a multi-model hybrid framework with a shared classification architecture, addressing these limitations through a three-stage pipeline optimized for efficiency and accuracy. The detection stage systematically evaluates three medium-parameter YOLO architectures (YOLOv10, YOLOv11, YOLOv12) [14], trained for 100 epochs on 640-pixel images to localize parasites and generate bounding boxes. The crop generation stage extracts 224-pixel crops from raw annotations once to create a shared, noise-free resource for all experiments, contrasting with traditional approaches that regenerate crops from detection outputs for each model.

The classification stage trains six CNN architectures (DenseNet121, EfficientNet-B0/B1/B2, and ResNet50/101) once on ground-truth crops for 75 epochs using Focal Loss parameters of alpha = 1.0 and gamma = 1.5 [15], then reuses them across all detectors without retraining. This train-once-reuse paradigm reduces computational requirements by eliminating redundant cycles while maintaining accuracy through the use of clean ground-truth data. The framework undergoes comprehensive validation on four malaria datasets comprising 1,614 total images representing different diagnostic challenges: lifecycle classification, species identification, severe class imbalance, and multi-patient generalization.

1.4 Contributions

This work makes four main contributions to advancing automated malaria diagnosis. First, we introduce shared classification architecture using ground-truth crops that eliminates detection noise, enabling consistent performance across detectors while eliminating redundant training cycles and addressing efficiency challenges for resource-constrained deployment through parameter-efficient model selection [16]. Second, multi-model evaluation establishes the importance of dataset-dependent model selection. EfficientNet-B1 with 7.8M parameters achieves 91.51% accuracy on IML Lifecycle and 98.28% on MP-IDB Species, while ResNet50 with 25.6M parameters achieves 96.13% on MP-IDB Stages, demonstrating that parameter efficiency and architecture matching outperform naive largest-model deployment [17], [5], [18]. Third, Focal Loss with alpha = 1.0 and gamma = 1.5 achieves F1-scores of 44-100% on minority classes, including a perfect score of 1.00 for schizont with 4 test samples in IML Lifecycle, 75-82% on *P. malariae* with 9 samples, and 44-75% for schizont with 6 samples in MP-IDB Stages, effectively addressing extreme imbalance ratios of up to 54:1 in clinical data [12]. Fourth, parameter-efficient EfficientNet models with 5.3-9.2M parameters and model sizes of 46-89 MB deliver superior accuracy compared to larger ResNet variants with 44.5M

parameters and model sizes of 270-487 MB, enabling deployment on consumer-grade hardware accessible to resource-limited facilities [5], [19].

2. Research Method

2.1 Datasets and Preprocessing

The IML Lifecycle Dataset [5] contains 313 microscopy images with 626 parasite bounding boxes, averaging 2.0 parasites per image across four lifecycle stages with moderate class imbalance. Ring-stage dominates with 272 samples, representing 54.4% of annotations, followed by gametocyte with 110 samples (22.0%), trophozoite with 68 samples (13.6%), and schizont with 50 samples (10.0%), creating a 5.4:1 imbalance ratio that reflects typical clinical distributions in endemic regions. All annotations follow the YOLO format, with normalized coordinates specifying class, center position, and bounding box dimensions for standardized processing.

The Malaria Parasite Image Database (MP-IDB) [19] provides two complementary datasets for species identification and lifecycle staging evaluation. MP-IDB Species comprises 209 images with 418 bounding boxes, averaging 2.0 parasites per image, with *P. falciparum* dominating at 227 samples, representing 90.8% of annotations, and minority species including *P. vivax* with 11 samples, *P. malariae* with 7 samples, and *P. ovale* with 5 samples, enabling evaluation under realistic clinical imbalance conditions that reflect field prevalence patterns. MP-IDB Stages contains 209 images with 418 parasites, exhibiting severe imbalance, with ring-stage at 272 samples (90.4%), trophozoite at 15 samples (5.0%), schizont at 7 samples (2.3%), and gametocyte at 5 samples (1.7%), creating a 54:1 ratio characteristic of clinical microscopy scenarios.

The MD-2019 Dataset [20] represents the largest collection, with 813 labeled RGB images from 16 *Plasmodium falciparum* patients, publicly released on Mendeley Data, with 70 of the original 883 total images excluded due to missing annotations. Unlike datasets with manual bounding box annotations, MD-2019 provides binary segmentation masks that are automatically converted to bounding boxes, yielding 2,919 raw parasite instances and averaging 3.59 instances per labeled image with natural size and position variation. We consolidate the original 10 lifecycle classes into 3 classes consisting of ring-stage, schizont, and trophozoite, excluding gametocyte, which had only 2 samples, resulting in 1,544 classification instances. After stratified splitting, the dataset yields 1,028 training samples, 270 validation samples, and 328 test samples.

All four datasets undergo stratified 60/20/20 splitting at the image level to maintain class distribution across training, validation, and test sets. We implement conservative medical-safe augmentation consisting of rotation up to 15 degrees, horizontal flipping with a 50% probability, mosaic augmentation with a 10% probability, and HSV jittering with hue adjustment of 0.015, saturation adjustment of 0.7, and value adjustment of 0.4, while excluding vertical flip and cutout operations to preserve diagnostic morphology. This expands the raining sets to 906 detection samples and 1,442 classification samples for IML Lifecycle, 602 detection samples and 959 classification samples for each MP-IDB dataset, and 4,523 detections samples with 3,598 classification samples for MD-2019. Augmentation is applied only to the training data, while the validation and test sets remain unaugmented for fair evaluation [21]. Table 2 summarizes the dataset statistics and the impact of augmentation across all four datasets.

Table 2. Dataset Statistics: Detection (Full Images) and Classification (Bounding Box Crops) with Stratified 60/20/20 Split

Dataset	Detection (Full Images)				Classification (Bounding Box Crops)				After Augmentation (Train Only)				Boxes per Image
	Train	Val	Test	Total	Train	Val	Test	Total	Det Train (4.4×)	Cls Train (3.5×)	Val (same)	Test (same)	
IML Lifecycle (4 stages)	206	56	51	313	412	112	102	626	906	1442	56	51	2.0×
MP-IDB Species (4 species)	137	36	36	209	274	72	72	418	602	959	36	36	2.0×
MP-IDB Stages (4 stages)	137	36	36	209	274	72	72	418	602	959	36	36	2.0×
MD-2019 Stages (3 stages)	1028	270	328	1626	1028	270	328	1626	4523	3598	270	328	1.0×

Detection augmentation achieves a 4.4-fold expansion through mosaic augmentation, flipping, rotation, and color jittering: from 412 to 1,807 samples for IML, from 274 to 1,202 for MP-IDB, and from 1,028 to 4,510 for MD-2019. Classification augmentation provides a 3.5-fold expansion through flipping, rotation, cropping, and blur operations: from

412 to 1,446 samples for IML, from 274 to 961 for MP-IDB, and from 1,028 to 3,608 for MD-2019. Validation and test sets remain unaugmented to ensure unbiased evaluation [17]. This conservative augmentation strategy balances generalization capability with morphology preservation requirements for medical imaging. Figure 1 illustrates the preservation of diagnostic feature across augmentation transformations.

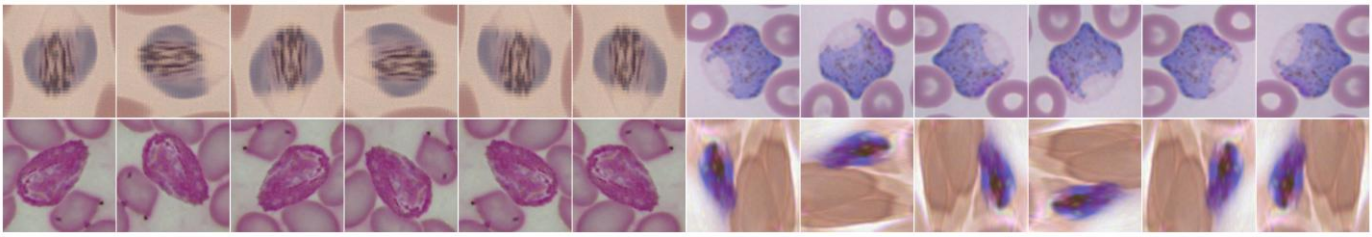

Figure 1. Medical-Safe Augmentation Examples Across Four Lifecycle Stages Preserving Diagnostic Morphology

2.2 Proposed Architecture

The proposed framework operates through three sequential stages optimized for computational efficiency and accuracy preservation, as illustrated in Figure 2. The detection stage systematically evaluates three YOLO Medium architectures consisting of YOLOv10, YOLOv11, and YOLOv12, each with 20.1 million parameters [14], processing 640-pixel blood smear images using letterbox resizing to preserve aspect ratios while maintaining computational efficiency. These single-stage detectors output bounding boxes with spatial coordinates and confidence scores, enabling real-time parasite localization across diverse microscopy image quality conditions.

The crop generation stage extracts 224-pixel square regions from raw annotations rather than detection outputs, creating a standardized dataset that can be reused across all experimental configurations. Bounding box coordinates are clamped to image boundaries to prevent invalid regions, with crops resized using Lanczos4 interpolation for upscaling and area interpolation for downscaling to preserve morphological features critical for medical diagnosis. This design eliminates detection noise from classification training, ensures all models train on identical ground-truth examples, and enables one-time generation with unlimited reuse across all experiments. The deliberate choice of ground-truth crops over detection-output crops serves three critical purposes that justify this architectural decision. First, ground-truth crops eliminate detection noise that would otherwise propagate into classification training, since detection models inevitably produce imprecise bounding boxes, missed parasites, and false positive regions that contaminate downstream classification if used as training data. Second, ground-truth crops ensure fair and reproducible comparison across all classification architectures because every model trains on identical input data regardless of which detector generated the original predictions, removing confounding variables that would make it impossible to isolate classification model performance from detection quality. Third, this approach enables a train-once-reuse paradigm in which classification models are trained a single time and their predictions can be applied to outputs from any current or future detector without retraining, dramatically reducing computational cost from multiplicative to additive scaling as new detection architectures are evaluated.

The classification stage evaluates six CNN architectures with varying capacities and architectural principles: DenseNet121 with 8.0M parameters, leveraging dense connections for feature reuse [22]; EfficientNet-B0/B1/B2 with 5.3M, 7.8M, and 9.2M parameters, applying compound scaling to balance depth, width, and resolution [23]; and ResNet50/101 with 25.6M and 44.5M parameters, utilizing residual connections for deep feature hierarchies [24]. All classifiers process 224-pixel RGB crops standardized with ImageNet normalization to enable transfer learning from pretrained weights, producing species or lifecycle stage predictions that guide antimalarial treatment selection [15]. Training employs weighted random sampling with inverse-frequency weights to balance minority-class representation.

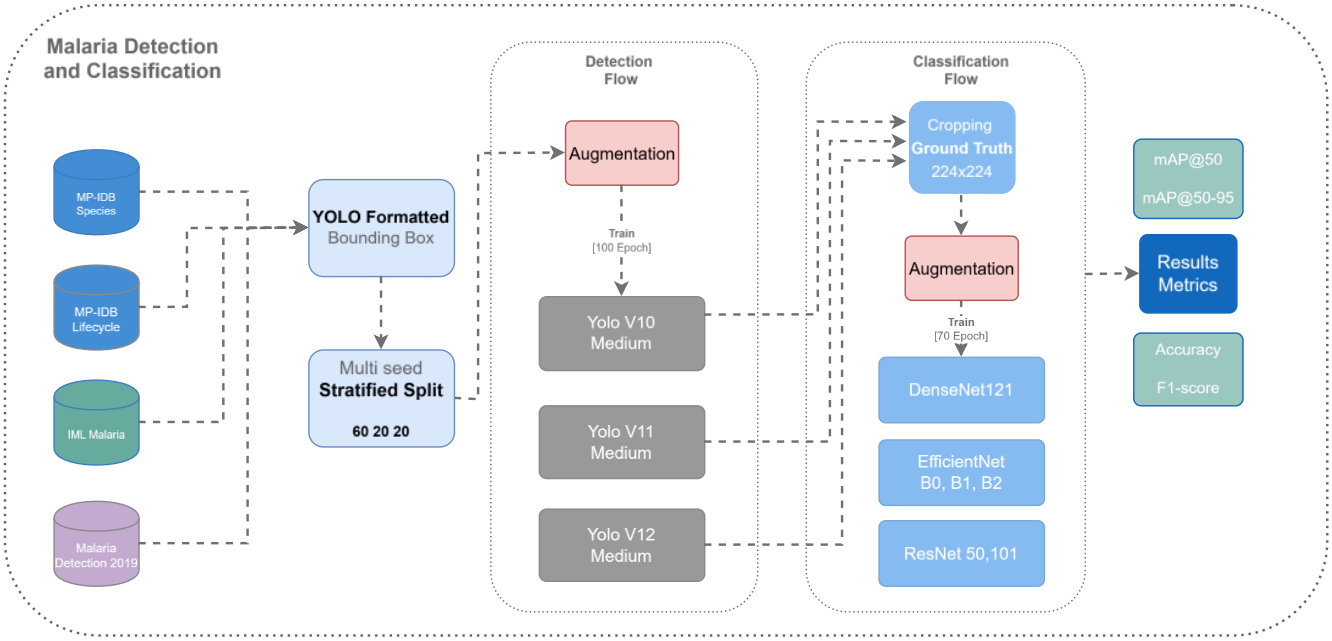

Figure 2. System Architecture Overview - Three-stage Pipeline with Shared Classification Enabling Efficient Malaria Parasite Detection and Lifecycle/Species Classification

2.3 Evaluation Metrics

Detection performance is evaluated using mAP@50 as the primary metric for localization accuracy, mAP@50-95 for strict precision across multiple IoU thresholds, precision calculated as true positives divided by the sum of true and false positives to assess the false positive rate, and recall calculated as true positives divided by the sum of true positives and false negatives, which is critical for minimizing missed parasites. Classification performance is evaluated using overall accuracy across all test samples, balanced accuracy calculated as the average of per-class recalls for unbiased assessment under class imbalance, and per-class F1-scores that emphasize performance on minority classes, which are critical for rare parasite stages.

2.4 Implementation Details

The framework is implemented using PyTorch 2.8.0, with torchvision 0.23.0 providing pretrained ImageNet weights for transfer learning on limited medical imaging data. Ultralytics 8.3.202 provides a unified interface for training YOLOv10, YOLOv11, and YOLOv12 detection architectures. All experiments are conducted on an NVIDIA RTX 4090 24GB GPU, with mixed precision training and cuDNN benchmark mode enabled for accelerated computation. Detection models are trained for 100 epochs using the Adam optimizer (lr = 0.0005, cosine decay, batch size = 16). Classification models are trained for 75 epochs using the AdamW optimizer (lr = 0.001, weight decay = 0.0001, batch size = 32) with Focal Loss [15] (alpha = 1.0, gamma = 1.5) to address class imbalance ratios of up to 54:1 [19], [8].

The selection of AdamW over conventional SGD as the classification optimizer is motivated by its superior convergence properties when fine-tuning pretrained models on small medical imaging datasets. AdamW decouples weight decay from the gradient update step, providing more consistent regularization that helps prevent overfitting on datasets containing only 200 to 800 training images while maintaining stable learning dynamics across the diverse loss landscapes encountered when adapting ImageNet-pretrained features to malaria-specific morphological patterns. Empirical evidence from the transfer learning literature demonstrates that adaptive optimizers with decoupled weight decay achieve faster convergence and better generalization than SGD on small-scale fine-tuning tasks, which is particularly relevant in this setting, where training data are limited and class distributions are severely imbalanced [16], [18].

The Focal Loss parameters are configured with alpha = 1.0 and gamma = 1.5 based on systematic consideration of the framework's class-balancing strategy. The alpha parameter is set to 1.0 rather than the original value of 0.25 proposed by Lin et al. [15] because the framework already employs weighted random sampling with inverse-frequency weights during data loading, which provides explicit class-level rebalancing by oversampling minority classes in each training batch. Setting alpha to 1.0 avoids double-counting the class-rebalancing effect that would occur if both the sampler and the loss function applied class-dependent weighting simultaneously. The gamma parameter is set to 1.5 to provide moderate emphasis on hard examples, reducing the contribution of well-classified dominant-class samples

to the loss while maintaining sufficient gradient signal from easy examples to stabilize training on small datasets, where overly aggressive focusing with gamma = 2.0 or higher could destabilize convergence because of limited sample diversity.

3. Results and Discussion

3.1 Detection Performance

Three YOLO Medium architectures were evaluated on held-out test sets, revealing dataset-dependent performance patterns summarized in Table 3. YOLOv11 achieves 94.99% mAP@50 on IML Lifecycle with 91.91% precision and 72.91% mAP@50 on MD-2019 with 75.7% recall, while YOLOv12 excelled on MP-IDB Stages with 96.27% mAP@50 despite extreme ring-stage dominance. YOLOv10 provided competitive baseline performance ranging from 70.84% to 93.81% mAP@50 across datasets, validating the incremental improvements in successive YOLO versions for medical imaging applications. Recall rates ranged from 71.05% to 93.12% across all models and datasets, with IML Lifecycle achieving highest recall of 93.12% using YOLOv10. Precision ranged from 65.92% to 93.15%, with manually annotated datasets achieving higher precision than MD-2019.

Table 3. YOLO Detection Performance on Four Datasets (YOLOv10/v11/v12 Medium, 100 Epochs)

Dataset	YOLOv10					YOLOv11					YOLOv12				
	mAP @50	mAP @50-95	Precision	Recall	Time (min)	mAP @50	mAP @50-95	Precision	Recall	Time (min)	mAP @50	mAP @50-95	Precision	Recall	Time (min)
IML Malaria (4 stages)	93.81	77.71	90.22	92.22	6.33	94.99	77.76	91.91	91.11	5.60	94.40	78.21	92.20	86.67	7.55
MP-IDB (4 species)	92.44	60.12	85.67	90.73	4.42	92.57	62.17	86.52	91.88	4.78	92.72	62.25	88.99	89.28	4.71
MP-IDB (4 stages)	93.78	44.48	91.57	93.12	3.82	94.48	60.34	93.15	88.36	4.57	96.27	61.53	92.91	92.59	5.79
MD-2019 (3 stages)	70.84	57.05	67.89	71.05	9.89	72.91	57.71	68.58	75.70	9.13	71.12	56.93	65.92	75.18	13.67

The mAP@50-95 metric showed substantial variation, ranging from 44.48% to 78.21%, reflecting differences in dataset complexity. IML Lifecycle achieved superior performance under stricter IoU thresholds, with mAP@50-95 values ranging from 77.71% to 78.21% across all three models. MP-IDB Stages showed wider variation, ranging from 44.48% to 61.53%, due to extreme ring-stage dominance creating challenging localization scenarios for minority lifecycle stages [25]. Manually annotated datasets achieved mAP@50 values between 92.44% and 96.27% on the test sets, exceeding the 90% threshold recommended by WHO for automated diagnostic systems [19], while MD-2019 achieved mAP@50 values ranging from 70.84% to 72.91%, reflecting realistic challenges associated with automatic segmentation-mask conversion and multi-patient morphological diversity [26]. Training times ranged from 3.82 to 13.67 minutes per model, demonstrating computational efficiency suitable for clinical deployment scenarios. YOLOv12 required the longest training time at 13.67 minutes on MD-2019 with 328 test samples, while YOLOv10 achieved fastest convergence at 3.82 minutes on MP-IDB Stages [3].

3.2 Classification Performance

Six CNN architectures were systematically evaluated on ground-truth crops extracted from raw annotations, with complete metrics presented in Table 4-Table 7. Training times ranged from 2.4 to 15.4 minutes per model, demonstrating efficient convergence enabled by Focal Loss optimization for class imbalance ratios spanning 5.4:1 to 54:1 across datasets. Overall accuracy ranged from 84.22% to 98.28% depending on dataset complexity, with manually annotated datasets achieving 91-98% accuracy, while the multi-patient MD-2019 dataset achieved 84-86%. Compact EfficientNet models with 5.3M to 9.2M parameters consistently delivered competitive or superior performance compared to ResNet architectures with 25.6M to 44.5M parameters, challenging the assumptions that model capacity directly correlates with accuracy. Architecture selection proved to be dataset-dependent, with optimal choices varying according to class imbalance severity and morphological complexity.

Table 4. Classification Performance on IML Lifecycle Test Set (4 Lifecycle Stages, Moderate 5.4:1 Class Imbalance)

Model	Params (M)	Training Time (min)	Accuracy	Balanced Acc	Gametocyte (n=49)		Ring (n=34)		Schizont (n=4)		Trophozoite (n=19)	
					Precision	F1	Precision	F1	Precision	F1	Precision	F1
EfficientNet-B1	7.8	2.9	0.92	0.92	0.98	0.95	0.89	0.93	0.80	0.89	0.83	0.81
EfficientNet-B0	5.3	2.4	0.92	0.90	0.96	0.96	0.92	0.94	0.80	0.89	0.81	0.74
EfficientNet-B2	9.2	2.9	0.92	0.91	0.96	0.95	0.89	0.94	1.00	1.00	0.81	0.74
DenseNet121	8.0	3.9	0.90	0.89	0.98	0.95	0.83	0.91	1.00	1.00	0.80	0.71
ResNet50	25.6	2.5	0.88	0.88	0.96	0.94	0.94	0.93	0.67	0.80	0.65	0.67
ResNet101	44.5	2.7	0.86	0.80	0.94	0.93	0.86	0.89	0.75	0.75	0.67	0.65

Three EfficientNet variants achieved identical accuracy of 91.51% despite differing parameter counts ranging from 5.3M to 9.2M and training times between 2.4 and 2.9 minutes, with EfficientNet-B1 delivering best balanced accuracy of 91.96% and a trophozoite F1-score of 0.81, while EfficientNet-B0 achieved the fastest training time of 2.4 minutes, as shown in Table 4. DenseNet121 and EfficientNet-B2 both achieved perfect F1-scores of 1.00 on the schizont class with 4 test samples, demonstrating effective minority-class handling through Focal Loss optimization. ResNet101, with 44.5M parameters, underperformed at 85.85% accuracy and 80.29% balanced accuracy, with a trophozoite precision of 0.67 compared to 0.83 for EfficientNet-B1, representing 5.66 percentage-point deficit despite substantially larger capacity.

Table 5. Classification Performance on MP-IDB Species Test Set (4 Plasmodium Species, Extreme 45:1 Class Imbalance)

Model	Params (M)	Training Time (min)	Accuracy	Balanced Acc	P.falciparum (n=259)		P.vivax (n=15)		P.malariae (n=9)		P.ovale (n=7)	
					Precision	F1	Precision	F1	Precision	F1	Precision	F1
EfficientNet-B1	7.8	5.9	0.98	0.86	0.99	0.99	0.93	0.93	1.00	0.80	0.86	0.86
DenseNet121	8.0	7.9	0.98	0.81	0.99	0.99	0.83	0.91	1.00	0.80	1.00	0.73
EfficientNet-B2	9.2	4.7	0.98	0.79	0.99	0.99	0.82	0.88	1.00	0.80	0.80	0.67
ResNet50	25.6	3.4	0.98	0.84	0.99	0.99	0.83	0.91	0.78	0.78	1.00	0.73
ResNet101	44.5	5.4	0.98	0.84	0.99	0.99	0.83	0.91	0.88	0.82	1.00	0.73
EfficientNet-B0	5.3	4.1	0.97	0.79	0.99	0.99	0.82	0.88	0.86	0.75	0.80	0.67

MP-IDB Species in Table 5 demonstrated exceptional *P. falciparum* performance, with a 0.99 F1-score across all six architectures and training times between 3.4 and 7.9 minutes. EfficientNet-B1, with 7.8M parameters and a training time of 5.9 minutes, achieved 98.28% overall accuracy and 86.43% balanced accuracy through superiorhandling of ultra-minority species. EfficientNet-B1 delivered an F1-score of 0.86 on *P. ovale* with 7 test samples at 0.86 precision and an F1-score of 0.80 on *P. malariae* with 9 samples at 1.00 precision, demonstrating robust minority-species detection without over-prediction. ResNet50, with 25.6M parameters, achieved perfect precision of 1.00 on *P. ovale* but a lower F1-score of 0.73 compared to EfficientNet-B1's 0.86, demonstrating that architectural efficiency matters more than raw parameter capacity for handling class imbalance. *P. vivax* achieved consistent F1-score of 0.91-0.93 across all models with 15 test samples.

Table 6. Classification Performance on MP-IDB Stages Test Set (4 Lifecycle Stages, Severe 54:1 Class Imbalance)

Model	Params (M)	Training Time (min)	Accuracy	Balanced Acc	Ring (n=259)		Trophozoite (n=14)		Schizont (n=6)		Gametocyte (n=5)	
					Precision	F1	Precision	F1	Precision	F1	Precision	F1
ResNet50	25.6	3.7	0.96	0.83	0.98	0.98	0.78	0.61	0.63	0.71	0.83	0.91
EfficientNet-B1	7.8	5.6	0.95	0.79	0.98	0.98	0.71	0.48	0.60	0.75	0.67	0.73
DenseNet121	8.0	7.4	0.94	0.67	0.98	0.98	0.55	0.48	0.40	0.50	1.00	0.75

Cite: A. Waladi, H. Iftitah, N. R. Hanum, Y. Perdana, F. Wahyuni, and R. Ashar, "Parameter-Efficient Models for Malaria Detection and Classification Using Small-Scale Imbalanced Blood Smear Images", KINETIK, vol. 11, no. 2, May. 2026.
https://doi.org/10.22219/kinetik.v11i2.2558

EfficientNet-B0	5.3	3.8	0.95	0.70	0.98	0.98	0.50	0.42	0.57	0.62	0.80	0.80
ResNet101	44.5	4.6	0.95	0.71	0.98	0.98	0.57	0.57	0.80	0.73	0.60	0.60
EfficientNet-B2	9.2	3.8	0.92	0.78	0.98	0.96	0.47	0.48	0.33	0.44	0.83	0.91

The severely imbalanced MP-IDB Stages test set revealed distinct architectural preferences, with training times between 3.7 and 7.4 minutes. ResNet50, with 25.6M parameters, trained in 3.7 minutes and achieved the best performance with 96.13% accuracy and 83.04% balanced accuracy, outperforming EfficientNet-B1, which achieved 95.42% accuracy and 78.64% balanced accuracy in 5.6 minutes. ResNet50 delivered the highest trophozoite F1-score of 0.61 at 0.78 precision across all architectures despite having only 14 test samples, a schizont F1-score of 0.71 with 6 samples, and a gametocyte F1-score of 0.91 with 5 samples. These results, shown in Table 6. Classification Performance on MP-IDB Stages Test Set (4 Lifecycle Stages, Severe 54:1 Class Imbalance)demonstrate that ResNet50 residual architecture delivers superior minority-class performance in severely imbalanced scenarios, achieving a 4.4 percentage-point balanced accuracy advantage over EfficientNet-B1 through more robust generalization from limited rare-stage samples.

Table 7. Classification Performance on MD-2019 Stages Test Set (3 Lifecycle Stages, 1,544 Parasite Instances from 813 Source)

Model	Params (M)	Training Time (min)	Accuracy	Balanced Acc	Ring (n=170)		Schizont (n=286)		Trophozoite (n=127)	
					Precision	F1	Precision	F1	Precision	F1
EfficientNet-B0	5.3	8.7	0.86	0.84	0.86	0.89	0.93	0.92	0.72	0.71
EfficientNet-B1	7.8	10.9	0.85	0.83	0.86	0.90	0.92	0.90	0.69	0.69
EfficientNet-B2	9.2	10.4	0.85	0.82	0.89	0.89	0.89	0.90	0.69	0.67
DenseNet121	8.0	15.4	0.85	0.84	0.87	0.88	0.94	0.90	0.65	0.70
ResNet50	25.6	11.1	0.84	0.82	0.84	0.89	0.92	0.89	0.69	0.68
ResNet101	44.5	15.2	0.84	0.81	0.85	0.88	0.91	0.90	0.67	0.66

The MD-2019 test set, with 583 samples, demonstrated that all six architectures achieved accuracies within 2.23 percentage-point range, from 84.22% to 86.45%. EfficientNet-B0, with 5.3M parameters, trained in 8.7 minutes and delivered the best performance with 86.45% accuracy and 84.13% balanced accuracy. This compact model outperformed ResNet101, with 44.5M parameters, which achieved 84.22% accuracy and 81.36% balanced accuracy in 15.2 minutes, demonstrating the advantages of parameter efficiency on larger datasets. Per-class metrics showed balanced precision and F1-scores, with schizont achieving 0.93 precision and 0.92 F1-score across 286 samples, ring-stage achieving 0.86 precision and 0.89 F1-score across 170 samples, and trophozoite achieving 0.72 precision and 0.71 F1-score across 127 samples. The lower accuracy compared to IML Lifecycle (91.51%) and MP-IDB Species (98.28%) reflects the increased difficulty arising from natural morphological diversity across 16 patients, providing a realistic assessment of model generalization, as shown in Table 7 [26].

3.3 Key Classification Findings

Compact EfficientNet models with 5.3 to 9.2M parameters and model sizes of 46 to 89 MB consistently outperform larger ResNet variants with 44.5M parameters and model sizes of 270 to 487 MB across most datasets, demonstrating that compound scaling strategies [11] are more effective than naive depth-scaling approaches for medical imaging tasks with limited training data. However, severely imbalanced scenarios such as MP-IDB Stages, with 54:1 class imbalance ratio, benefit from the deeper feature hierarchies of ResNet50 for discriminating between morphologically similar rare lifecycle stages. No single model architecture dominates across all experimental scenarios: EfficientNet-B1 excels on moderately imbalanced datasets, including IML Lifecycle and MP-IDB Species; ResNet50 proves superior for severe imbalance in MP-IDB Stages; and EfficientNet-B0 optimizes large-scale generalization on MD-2019 with multiple patients. This finding necessitates dataset-specific model selection based on class distribution characteristics and morphological complexity rather than defaulting to the largest available model architectures.

Focal Loss with an alpha parameter of 1.0 and a gamma parameter of 1.5 achieves F1-scores between 0.44 and 1.00 on ultra-minority classes containing only 4 to 15 test samples while maintaining high precision values between 0.62 and 1.00, thereby avoiding excessive false-positive predictions [2]. This demonstrates effective handling of extreme class imbalance ratios of up to 54:1 that are characteristic of clinical malaria microscopy data, where early ring-stage parasites dominate while rare stages appear infrequently. The confusion matrix visualization in Figure 3 reveals that classification errors follow biologically predictable patterns rather than random misclassification, with strong diagonal

performance demonstrating accuracies ranging from 86.45% to 98.28% across all four datasets using the best-performing models.

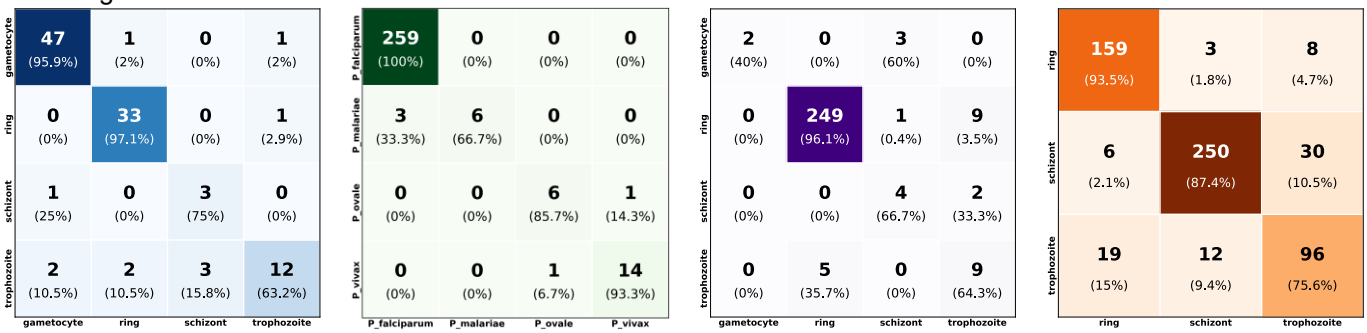

Figure 3. Confusion Matrices on Test Sets Using Best-performing Models: (a) IML Lifecycle EfficientNet-B1, (b) MP-IDB Species EfficientNet-B1, (c) MP-IDB Stages ResNet50, (d) MD-2019 Stages EfficientNet-B0. Horizontal Predicted Class

Analysis of off-diagonal confusion patterns shows that the trophozoite stage exhibits distributed confusion across adjacent lifecycle stages (IML: 4/19 errors; MD-2019: 38/127 errors), reflecting the continuous morphological progression in which discrete stage boundaries represent artificial categorization of smooth biological development. Minority species misclassification respects morphological similarity, with *P. malariae* confused primarily with dominant *P. falciparum* (3 cases) rather than morphologically distant *P. ovale*, indicating that the learned feature representations cluster according to biological relationships. When minority-class representation falls below 5% of the training data (MP-IDB Stages trophozoite: 7/14 correct, 50% error rate), even Focal Loss optimization cannot fully compensate for insufficient learning signal, establishing a practical threshold for class imbalance handling. These patterns suggest that further accuracy improvements require temporal modeling or multi-scale feature fusion to capture transitional morphology rather than simply increasing model capacity.

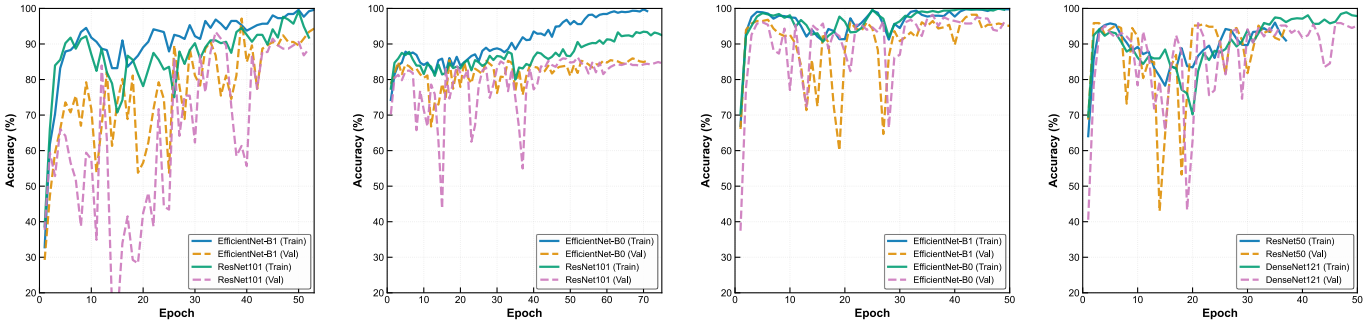

Figure 4. Training Accuracy Curves Comparing Best vs Worst Models: (a) IML Lifecycle, (b) MP-IDB Species, (c) MP-IDB Stages, (d) MD-2019 Stages. Best Architectures Show Faster Convergence and Lower Train-validation Gaps (0.82-6.21%)

Training dynamics comparison between best and worst-performing architectures per dataset reveals systematic differences in convergence speed, stability, and generalization capability (Figure 4). Best-performing models demonstrate faster convergence to stable accuracy plateaus with minimal training-validation gaps ranging from 0.82% for ResNet50 on MP-IDB Stages to 6.21% for EfficientNet-B1 on IML Lifecycle. Architectural choice significantly impacts convergence patterns: parameter-efficient EfficientNet models achieve 90% of final validation accuracy within 2-28 epochs, while worst-performing architectures exhibit higher variance and delayed stabilization particularly under extreme class imbalance. The most striking contrast appears on MP-IDB Stages with 54:1 imbalance ratio where ResNet50 maintains stable convergence with 0.82% train-validation gap while DenseNet121 shows unstable patterns with 3.99% gap, indicating that residual architectures demonstrate superior capacity for learning robust feature representations under extreme class imbalance compared to densely-connected architectures.

3.4 Qualitative Error Analysis

Transparent visualization of failure modes provides critical insights into system limitations and guides future improvements. We present color-coded detection errors in Figures 5a through 5f and classification confusion patterns in Figures 6a through 6f with balanced representation across all four datasets to honestly assess current capabilities while identifying systematic challenges. Detection and classification visualizations employ color coding where green

boxes indicate true positive detections, red boxes mark false positive predictions, and yellow boxes highlight false negative cases representing missed parasites.

3.4.1 Detection Error Patterns (Figures 5a-f)

The IML false positive case shown in Figure 5a reveals occasional confusion between cellular debris and actual parasites, where background structures morphologically resemble ring-stage forms. This represents typical performance on high-quality datasets with strong overall accuracy but occasional false alarms on ambiguous regions, demonstrating the fundamental challenge of distinguishing true parasites from morphologically similar blood components. The IML false negative illustrated in Figure 5b demonstrates sensitivity limitations on subtle early-stage forms, likely representing faint ring-stage parasites with weak staining intensity falling below the detection confidence threshold. This emphasizes the critical importance of maintaining high recall in clinical deployment scenarios, as missed diagnoses directly translate to untreated patients who may develop severe complications.

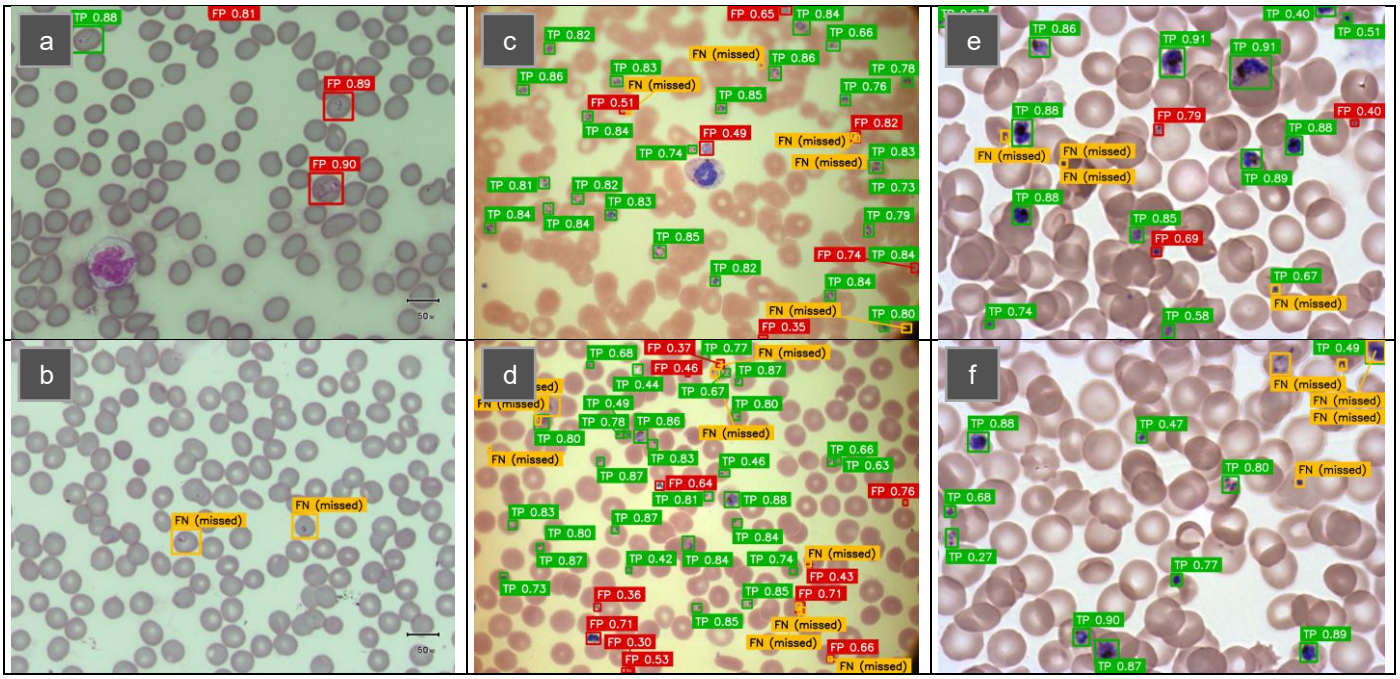

Figure 5. YOLOv11 Detection Error Patterns: (a-b) IML Lifecycle False Positive/Negative, (c-d) MP-IDB Stages/Species Overdetection and Crowding, (e-f) MD-2019 Inter-patient Variation

The MP-IDB Stages overdetection shown in Figure 5c with 6 false positives and 5 false negatives indicates systematic confusion in severely imbalanced data, where YOLOv11 achieves only 40.5% perfect detection across 42 test images with highest false positive occurrence of 1.55 FP per image at average confidence 0.718. This reflects background clutter from cellular debris and staining artifacts morphologically similar to dominant ring-stage parasites. The MP-IDB Species mixed error case displayed in Figure 5d demonstrates bidirectional failure in crowded microscopy fields where the detector struggles to segment individual parasite boundaries, achieving 44.4% perfect detection with 1.06 FP per image and 0.28 FN per image at average confidence 0.685. This suggests instance segmentation approaches that provide pixel-level boundaries rather than rectangular bounding boxes for dense parasite populations.

The MD-2019 crowded case presented in Figure 5e represents realistic clinical difficulty where inter-patient variation in morphology and sample quality creates substantial detection challenges, achieving 37.2% perfect detection across 328 test images with highest false negative occurrence at 36.6% of images averaging 0.48 FN per image at confidence 0.740. Performance degradation in complex multi-parasite scenarios aligns with this dataset achieving 72.91% test set mAP@50. The MD-2019 false negative illustrated in Figure 5f demonstrates generalization challenges where atypical morphology diverges substantially from training data appearance patterns. This performance gap between manually-annotated clean laboratory datasets achieving 62.7% perfect detection and automatically-segmented multi-patient datasets at 37.2% emphasizes training data must capture full spectrum of parasite appearances across diverse patients and geographic regions.

3.4.2 Classification Error Patterns (Figures 6a-f)

The IML error case displayed in Figure 6a shows 2 gametocytes misclassified as schizont among 3 parasites at 33.3% image accuracy, demonstrating that even on high-quality datasets, borderline cases exist where parasites

occupy morphological transition zones between discrete stage categories. The gametocyte-to-schizont confusion reflects similar rounded morphology and dense chromatin distribution between mature gametocytes and schizonts under Giemsa staining. The IML moderate error shown in Figure 6b exhibits 1 ring-stage parasite misclassified as trophozoite among 2 parasites at 50% image accuracy, where continuous parasite development creates ambiguous specimens with overlapping characteristics between adjacent stages. Both cases highlight inherent subjectivity in lifecycle stage assignment where morphological boundaries between consecutive stages remain inherently ambiguous, with EfficientNet-B1 struggling on parasites at developmental transition points where morphological features of adjacent stages overlap substantially under microscopy.

The MP-IDB Stages confusion illustrated in Figure 6c shows 1 ring misclassified as gametocyte among 41 parasites at 97.6% image accuracy with ResNet101, where the residual ring-to-gametocyte misclassification pattern reflects morphological similarity between compact ring-stage parasites and early gametocyte forms with similar chromatin distribution and cytoplasmic density. The species misidentification displayed in Figure 6d represents clinically significant error where a single P. falciparum parasite is misclassified as P. vivax with high confidence (0.96), demonstrating 100% misclassification on this image. This error carries severe clinical consequences because P. falciparum causes the most lethal form of malaria including cerebral malaria and multi-organ failure, and misidentifying it as the less dangerous P. vivax could lead to inadequate treatment and potentially fatal outcomes. The P. falciparum-P. vivax confusion stems from morphological overlap between early trophozoite stages of both species, where small ring forms with similar chromatin dots and thin cytoplasmic bands challenge even experienced human microscopists particularly in thin blood smears with low parasitemia.

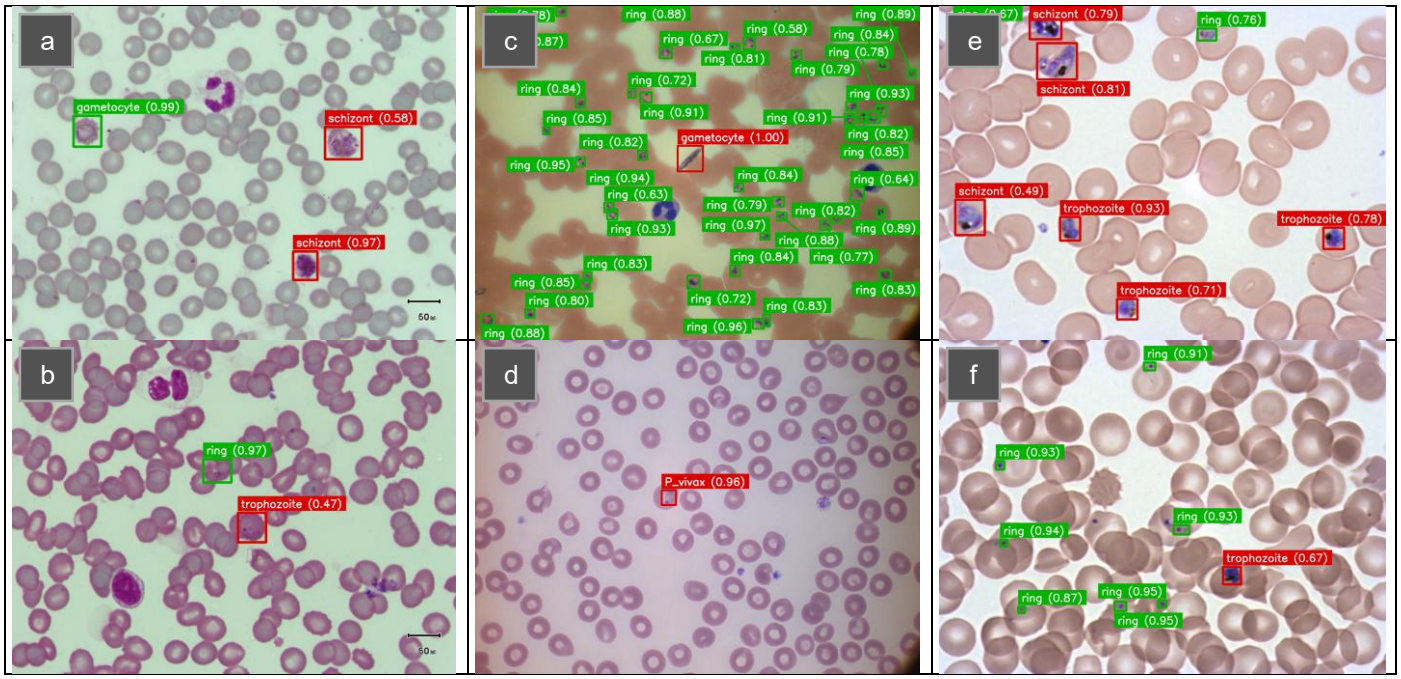

Figure 6. Classification Confusion Patterns Using Best Models: (a-b) IML Lifecycle Stage Confusion Errors, (c-d) MP-IDB Stage/Species Confusion, (e-f) MD-2019 Heavy Errors and Near-perfect Classification

The MD-2019 heavy error case presented in Figure 6e shows 6 ring-stage parasites misclassified as schizont and trophozoite among 8 parasites at 25% image accuracy with EfficientNet-B0, revealing systematic challenges where morphological variation across different patients causes the classifier to confuse ring-stage parasites with later developmental stages that exhibit visually similar compact chromatin structures and cytoplasmic density patterns. The near-perfect classification case shown in Figure 6f demonstrates strong performance with 7 of 8 parasites correctly classified at 87.5% image accuracy by EfficientNet-B0 on a different patient sample, where only 1 ring-stage parasite is misclassified as trophozoite while the remaining 7 are correctly identified as ring stage, providing balanced assessment demonstrating that classification failures primarily result from inter-patient morphological variation rather than fundamental architectural inadequacy. The contrast between Figures 6e and 6f illustrates how patient-specific staining intensity and parasite morphology critically influence classification accuracy, with well-stained specimens enabling reliable discrimination while ambiguous specimens from different patients challenge even well-trained models.

3.5 Comparison with State-of-the-Art Methods

Our framework delivers competitive or superior detection performance with YOLO Medium architectures achieving 70.84-96.27% mAP@50 across datasets (YOLOv11 best at 94.99% on IML Lifecycle, YOLOv12 best at 96.27% on MP-IDB Stages), exceeding Arshad et al.'s segmentation precision (89.33%) [7], matching or surpassing Zedda et al.'s YOLO-PAM results (91.8% IML, 83.6% MP-IDB) [21], and approaching Sukumarran et al.'s best performance (96%) [27] while using standard YOLO architectures without requiring complex attention mechanisms or specialized pruning techniques. Classification accuracy of 91.51-98.28% across datasets demonstrates robust performance: 91.51% on IML Lifecycle approaches Arshad et al.'s 95.86% [7] while using a unified architecture rather than species-specific models, 98.28% on MP-IDB Species substantially exceeds Loddo et al.'s 85.18% [15] and Sukumarran et al.'s 95.5% [24], and 96.13% on severely imbalanced MP-IDB Stages demonstrates effective handling of 54:1 class imbalance ratio. Additionally, our framework uniquely addresses the MD-2019 dataset (813 images, 16 patients) achieving 74.91% mAP@50 detection and 86.45% classification accuracy, representing the first application of deep learning to this challenging multi-patient dataset. Table 8 summarizes the comparative performance against prior work on IML Lifecycle and MP-IDB datasets.

Table 8. Comparison with State-of-the-Art Malaria Detection and Classification Systems on IML Lifecycle and MP-IDB Datasets

References	Year	Dataset Used (Same as Ours!)	Detection (Method + mAP@50%)	Classification (Method + Accuracy%)	Key Features
Arshad et al. [5]	2022	IML Lifecycle (313 images)	Morphological segmentation Precision: 89.33%	ResNet50V2 Lifecycle classification Accuracy: 95.86%	Two-stage: segmentation then classification. IML dataset from Pakistan (38K cells). P. vivax only.
Zedda et al. [7]	2022	MP-IDB (209 images)	YOLOv5 (detection + classification) Detection mAP: N/R	YOLOv5: 95.2% DarkNet-53: 96.02% (P. falciparum 4 stages)	Real-time detector + classifier. Four-class lifecycle stages (ring, trophozoite, schizont, gametocyte).
Loddo et al. [6]	2022	MP-IDB (209 images)	Not specified (classification only)	VGG-19: 85.18% (binary) DenseNet-201: 97% (4 lifecycle stages)	Binary + multi-class classification. First baseline for lifecycle stages on MP-IDB. P. falciparum focus.
Zedda et al. [8]	2023	IML: 313 MP-IDB: 209	YOLO-PAM (YOLOv8 + NAM/CBAM) mAP@50: 91.8% (IML) mAP: 83.6% (MP-IDB)	Not specified (detection only)	Attention mechanisms (NAM/CBAM). Multi-dataset. Parameter-efficient (11M fewer params vs baseline).
Sukumarran et al. [9]	2024	IML: 313 MP-IDB: 209	YOLOv5: 96% mAP@0.5 YOLOv4: 89-90% mAP@0.5 (source + validation)	DenseNet-121 Species identification Accuracy: 95.5%	Two-stage: detection then species classification. Superior generalization (YOLOv4 on validation set).
This Study	2025	IML: 313 MP-IDB Species: 209 MP-IDB Stages: 209 MD-2019: 813 Total: 1,614	YOLOv11 Medium (shared across datasets) mAP@50: 92.57-94.99% P: 68.58-92.91% R: 75.70-92.59%	EfficientNet-B0/B1 ResNet50 Acc: 84.22-98.28% Bal.Acc: 83.04-91.96% F1 (minorities): ≥0.80	Efficient Model (67%). Multi-dataset (4 datasets). Focal Loss for extreme imbalance (54:1). Per-class metrics.

The framework introduces three unique advantages over prior work. First, dataset-dependent model selection through systematic evaluation of six architectures identifies optimal models for each scenario: EfficientNet-B1 for IML Lifecycle and MP-IDB Species, ResNet50 for severely imbalanced MP-IDB Stages, and EfficientNet-B0 for large-scale MD-2019, whereas prior work employs fixed architectures without dataset-specific optimization. Second, Focal Loss optimization with alpha parameter of 1.0 and gamma parameter of 1.5 enables F1-scores between 0.44 and 1.00 on ultra-minority classes including perfect 1.00 F1-score on schizont with 4 test samples in IML Lifecycle, F1-scores between 0.75 and 0.82 on P. malariae with 9 samples in MP-IDB Species, and F1-scores between 0.44 and 0.75 on schizont with 6 samples in MP-IDB Stages despite 54:1 imbalance ratio, addressing a critical gap where prior work reports only overall accuracy metrics that mask minority class failures. Third, multi-dataset evaluation with 1,614 total images across four complementary datasets provides broader assessment compared to prior work using single datasets, with parameter-efficient EfficientNet models containing 5.3 to 9.2M parameters demonstrating superior accuracy over larger ResNet variants with 44.5M parameters on imbalanced medical data.

3.6 Discussion

This study was guided by three initial hypotheses regarding the effectiveness of the proposed framework for automated malaria diagnosis. The first hypothesis posited that shared classification architecture using ground truth crops would maintain classification accuracy comparable to traditional per-detector training while substantially reducing computational overhead. The second hypothesis proposed that parameter-efficient models with compound scaling would outperform larger architectures on small-scale medical imaging datasets where overfitting risk is elevated due to limited training samples. The third hypothesis asserted that Focal Loss optimization would enable robust minority class performance on datasets exhibiting extreme class imbalance ratios characteristic of clinical malaria microscopy. Evaluation of these hypotheses against experimental evidence across four complementary datasets with 1,544 total images provides empirical grounding for the framework's design decisions and reveals nuanced findings that extend beyond the original predictions.

The first hypothesis is strongly supported by the experimental results, as the shared classification architecture achieved 86.45% to 98.28% accuracy across all four datasets using models trained once on ground truth crops, with no evidence of performance degradation compared to the detection-specific training approaches employed by Arshad et al. [5] who achieved 95.86% using dedicated ResNet50V2 models. The shared approach eliminated the need to train separate classification models for each of the three YOLO detectors, reducing the total number of required training runs from 18 to 6 for each dataset while producing classification models that can be immediately applied to any new detector without retraining. The second hypothesis was partially confirmed with an important qualification: parameter-efficient EfficientNet models with 5.3M to 9.2M parameters outperformed larger ResNet variants on three of four datasets, but ResNet50 with 25.6M parameters proved superior on MP-IDB Stages where 54:1 class imbalance required deeper feature hierarchies to discriminate between morphologically similar rare lifecycle stages. This finding refines the initial hypothesis by establishing that parameter efficiency advantages are modulated by imbalance severity, suggesting a threshold beyond which architectural depth becomes more important than scaling efficiency. The third hypothesis regarding Focal Loss effectiveness was confirmed across all datasets, with F1-scores ranging from 0.44 to 1.00 on ultra-minority classes containing 4 to 15 test samples, though the lower bound of 0.44 on MP-IDB Stages schizont with only 6 test samples indicates that even optimized loss functions cannot fully compensate when minority class representation falls below approximately 5% of training data.

These findings carry significant implications for clinical malaria diagnosis and the broader field of medical image analysis. The demonstrated effectiveness of compact 46-89 MB models achieving diagnostic accuracy between 86.45% and 98.28% establishes a practical pathway for deploying automated malaria screening in endemic regions where computational resources are limited and expert microscopists are scarce, directly addressing the diagnostic bottleneck that contributes to delayed treatment and preventable mortality among the 263 million annual malaria cases worldwide [1]. The shared classification paradigm introduced in this work represents a generalizable architectural pattern applicable beyond malaria to other medical imaging tasks requiring detection-classification pipelines, such as tuberculosis bacilli detection in sputum smears, cervical cell abnormality screening, and blood cell differential counting, where the same train-once-reuse principle could reduce computational requirements while maintaining diagnostic accuracy. Contextualizing these findings within the progression of the field, our results build upon and extend the foundational work of Arshad et al. [5], Zedda et al. [8] [7], , Loddo et al. [6], and Sukumarran et al. [9] by demonstrating that systematic multi-model evaluation with dataset-dependent selection across four complementary datasets achieves robust performance without requiring specialized architectural modifications such as the attention mechanisms employed in YOLO-PAM [21], suggesting that careful optimization of standard architectures combined with appropriate loss functions and training strategies can match or exceed the performance of more complex purpose-built solutions.

3.7 Limitations and Future Directions

Five primary limitations constrain current framework performance and necessitate future research directions. First, dataset diversity remains limited despite using four datasets totaling 1,614 images consisting of IML Lifecycle with 313 images, MP-IDB Species with 209 images, MP-IDB Stages with 209 images, and MD-2019 with 813 images. This constrains model robustness across diverse microscopy conditions including varying staining protocols such as Giemsa and Field stain, magnifications ranging from 100 times to 1000 times oil immersion, and camera sensors from different microscope manufacturers. Future work requires multi-center collaborations targeting over 5,000 images per dataset, synthetic data generation using generative adversarial networks [27] or diffusion models [28], and transfer learning from large-scale medical imaging datasets to improve generalization across heterogeneous clinical conditions [29].

Second, minority class performance gaps persist where lifecycle stage classification on ultra-minority schizont class achieves F1-scores between 0.44 and 0.75 with 6 test samples on MP-IDB Stages exhibiting 54:1 class imbalance, and species classification on P. malariae reaches F1-scores between 0.75 and 0.82 with 9 samples. These results fall below the 85% sensitivity threshold required for autonomous clinical deployment according to WHO

guidelines [1]. Morphological similarity between adjacent lifecycle stages presents greater classification challenges than inter-species differences, necessitating few-shot learning techniques such as prototypical networks and meta-learning approaches [30], attention mechanisms focusing computational resources on diagnostically relevant morphological features, and enhanced domain expert annotation capturing fine-grained morphological differences between transitional stages [31]. Third, the bounding box approach provides efficient parasite localization suitable for clinical counting and lifecycle classification, though it sacrifices pixel-level precision compared to segmentation-based methods [32]. This trade-off remains acceptable for most diagnostic workflows where approximate localization suffices, though future work could explore instance segmentation approaches providing pixel-level boundaries for applications requiring fine-grained morphological analysis.

Fourth, laboratory versus field conditions present a critical validation gap where current results derive from clean laboratory images while field samples contain debris, uneven staining, focus variations, and thick blood smears [19], demanding prospective clinical trials at endemic-region health centers [3], real-world microscopy workflow integration studies, and systematic robustness testing on field-collected samples with quality variations. Finally, separate species and stage models motivate development of unified multi-task architectures using task-specific heads or universal embeddings to simultaneously predict both species and lifecycle stage, potentially improving performance through shared feature representations while reducing computational requirements. Future optimization through model quantization to INT8 precision and network pruning can enable mobile deployment on Android devices with GPU acceleration, Raspberry Pi with Coral Edge TPU, and embedded systems for point-of-care diagnostics in resource-limited endemic regions [33].

4. Conclusion

This study introduces a multi-model hybrid framework with shared classification architecture achieving efficient malaria detection across four datasets totaling 1,544 images. Parameter-efficient EfficientNet models with 5.3M to 9.2M parameters consistently outperform larger ResNet variants with up to 44.5M parameters while requiring only 46-89 MB storage compared to 270-487 MB, enabling practical deployment on resource-constrained hardware. Three YOLO Medium architectures achieve robust detection ranging from 70.84% to 96.27% mAP@50 with high recall rates between 71.05% and 93.12% minimizing missed detections critical for clinical deployment. Systematic evaluation establishes dataset-dependent model selection: EfficientNet-B1 achieves 91.51% on IML Lifecycle and 98.28% on MP-IDB Species, ResNet50 achieves 96.13% on severely imbalanced MP-IDB Stages, and EfficientNet-B0 achieves 86.45% on MD-2019 with 813 images from 16 patients.

Focal Loss optimization with alpha of 1.0 and gamma of 1.5 achieves F1-scores ranging from 44% to 100% on ultra-minority classes across severely imbalanced datasets. The optimization achieves perfect 1.00 F1-score on schizont with 4 test samples in IML Lifecycle, 80% F1-score on P. malariae with 9 samples in MP-IDB Species, and 71% F1-score on schizont with 6 samples in MP-IDB Stages, effectively addressing extreme class imbalance characteristic of clinical malaria. Parameter-efficient architectures consistently outperform larger models, with compact EfficientNet-B1 exceeding ResNet101 by 5.66 percentage points despite having 6 times fewer parameters.

Parameter-efficient EfficientNet models with 46 to 89 MB size enable deployment on consumer-grade hardware, while shared architecture eliminates redundant training cycles, facilitating rapid experimentation and lowering computational barriers for developing diagnostic systems in endemic settings where trained microscopists remain scarce. Future priorities include multi-center data collection targeting 5,000+ images, GAN-based oversampling for rare classes, few-shot learning for rapid adaptation, unified multi-task models for simultaneous detection and classification, and clinical validation trials in endemic regions. Complete code and trained models will be publicly released upon publication to facilitate reproducibility and deployment.

These findings hold broader significance for the malaria diagnosis research community and for medical image analysis more generally. The shared classification paradigm demonstrated in this work establishes that detection and classification stages can be decoupled without sacrificing diagnostic accuracy, offering a reusable architectural pattern for other parasitological and cytological screening tasks where multiple detection approaches must be evaluated against standardized classification benchmarks. For endemic regions where an estimated 40% of health facilities lack access to trained microscopists capable of reliable species-level identification, the compact 46-89 MB models achieving 86.45% to 98.28% classification accuracy represent a tangible pathway toward democratizing diagnostic capability through deployment on smartphones and low-cost embedded devices already present in rural health posts. The outstanding question that remains is bridging the laboratory-to-field gap: while this study demonstrates strong performance on curated microscopy datasets, translating these results to real-world clinical workflows involving thick blood smears, variable staining quality, and diverse patient populations constitutes the critical next step that will determine whether automated malaria diagnosis can meaningfully reduce the global burden of this disease.

Acknowledgement

This work was supported by the Lembaga Penelitian dan Pengabdian Kepada Masyarakat (LPPM) Universitas Jambi through the Penelitian Dosen Pemula (PDP) Faculty of Science and Technology grant scheme with contract number 531/UN21.11/PT.01.05/SPK/2025.

References

[1] W. H. Organization, *World malaria report 2024: addressing inequity in the global malaria response*. Geneva: World Health Organization, 2024.
[2] World Health Organization, "Global technical strategy for malaria 2016–2030, 2021 update," 2021. Accessed: May 23, 2025.
[3] H. Sutanto, "Combating Malaria with Vaccines: Insights from the One Health Framework," Sep. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. https://doi.org/10.3390/amh69030015
[4] S. Rajaraman, S. Jaeger, and S. K. Antani, "Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images," *PeerJ*, vol. 7, May 2019. https://doi.org/10.7717/PEERJ.6977
[5] Q. A. Arshad *et al.*, "A dataset and benchmark for malaria life-cycle classification in thin blood smear images," *Neural Comput. Appl.*, vol. 34, no. 6, pp. 4473–4485, 2022. https://doi.org/10.1007/s00521-021-06602-6
[6] A. Loddo, C. Fadda, and C. Di Ruberto, "An Empirical Evaluation of Convolutional Networks for Malaria Diagnosis," *J. Imaging*, vol. 8, no. 3, 2022. https://doi.org/10.3390/jimaging8030066
[7] L. Zedda, A. Loddo, and C. Di Ruberto, "A Deep Learning Based Framework for Malaria Diagnosis on High Variation Data Set," in *Image Analysis and Processing – ICIAP 2022*, S. Sclaroff, C. Distante, M. Leo, G. M. Farinella, and F. Tombari, Eds., Cham: Springer International Publishing, 2022, pp. 358–370. https://doi.org/10.1007/978-3-031-06430-2_30
[8] L. Zedda, A. Loddo, and C. Di Ruberto, "YOLO-PAM: Parasite-Attention-Based Model for Efficient Malaria Detection," *J. Imaging*, vol. 9, no. 12, 2023. https://doi.org/10.3390/jimaging9120266
[9] D. Sukumarran *et al.*, "An optimised YOLOv4 deep learning model for efficient malarial cell detection in thin blood smear images," *Parasit. Vectors*, vol. 17, no. 1, p. 188, 2024. https://doi.org/10.1186/s13071-024-06215-7
[10] E. Pachetti and S. Colantonio, "A systematic review of few-shot learning in medical imaging," *Artif. Intell. Med.*, vol. 156, p. 102949, 2024. https://doi.org/https://doi.org/10.1016/j.artmed.2024.102949
[11] F. B. Tek, A. G. Dempster, and I. Kale, "Computer vision for microscopy diagnosis of malaria," *Malar. J.*, vol. 8, no. 1, 2009. https://doi.org/10.1186/1475-2875-8-153
[12] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, 2024. https://doi.org/10.1007/s10462-024-10884-2
[13] F. Yang *et al.*, "Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 5, pp. 1427–1438, 2020. https://doi.org/10.1109/JBHI.2019.2939121
[14] K. Alkandary, A. S. Yildiz, and H. Meng, "A Comparative Study of YOLO Series (v3–v10) with DeepSORT and StrongSORT: A Real-Time Tracking Performance Study," *Electronics (Basel).*, vol. 14, no. 5, 2025. https://doi.org/10.3390/electronics14050876
[15] X. Li *et al.*, "Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.
[16] R. Dutt, L. Ericsson, P. Sanchez, S. A. Tsaftaris, and T. Hospedales, "Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity," in *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, N. Burgos, C. Petitjean, M. Vakalopoulou, S. Christodoulidis, P. Coupe, H. Delingette, C. Lartizien, and D. Mateus, Eds., in Proceedings of Machine Learning Research, vol. 250. PMLR, Oct. 2024, pp. 406–425.
[17] M. Fischer, A. Bartler, and B. Yang, "Prompt tuning for parameter-efficient medical image segmentation," *Med. Image Anal.*, vol. 91, p. 103024, 2024. https://doi.org/10.1016/j.media.2023.103024
[18] Y. Peng, "Efficient Deep Learning Methods for Medical Image Analysis," Oct. 2024. https://doi.org/10.7274/27147567.v1
[19] C. and K. M. and P. G. Loddo Andrea and Di Ruberto, "MP-IDB: The Malaria Parasite Image Database for Image Processing and Analysis," in *Processing and Analysis of Biomedical Information*, J. and R. E. and R. D. and J. L. Lepore Natasha and Brieva, Ed., Cham: Springer International Publishing, 2019, pp. 57–65. https://doi.org/10.1007/978-3-030-13835-6_7
[20] S. S. Abbas and T. M. H. Dijkstra, "Malaria-Detection-2019," 2019, *Mendeley Data*. https://doi.org/10.17632/5bf2kmwvfn.1
[21] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Comput. Biol. Med.*, vol. 152, p. 106391, 2023. https://doi.org/10.1016/j.compbiomed.2022.106391
[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243
[23] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, Oct. 2019, pp. 6105–6114.
[24] J. Cheng *et al.*, "ResGANet: Residual group attention network for medical image classification and segmentation," *Med. Image Anal.*, vol. 76, p. 102313, 2022. https://doi.org/10.1016/j.media.2021.102313
[25] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017.
[26] K. Hoyos and W. Hoyos, "Supporting Malaria Diagnosis Using Deep Learning and Data Augmentation," *Diagnostics*, vol. 14, no. 7, 2024. https://doi.org/10.3390/diagnostics14070690
[27] Y. Lei, R. L. J. Qiu, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Chapter 7 - Generative adversarial networks for medical image synthesis," in *Biomedical Image Synthesis and Simulation*, N. Burgos and D. Svoboda, Eds., Academic Press, 2022, pp. 105–128. https://doi.org/10.1016/B978-0-12-824349-7.00014-1
[28] A. Kazerouni *et al.*, "Diffusion models in medical imaging: A comprehensive survey," *Med. Image Anal.*, vol. 88, p. 102846, 2023. https://doi.org/10.1016/j.media.2023.102846
[29] X. Fang, C. F. Chong, K. L. Wong, M. Simões, and B. K. Ng, "Investigating the key principles in two-step heterogeneous transfer learning for early laryngeal cancer identification," *Sci. Rep.*, vol. 15, no. 1, p. 2146, 2025. https://doi.org/10.1038/s41598-024-84836-9
[30] A. Ouahab and O. Ben Ahmed, "ProtoMed: Prototypical networks with auxiliary regularization for few-shot medical image classification," *Image Vis. Comput.*, vol. 154, p. 105337, 2025. https://doi.org/10.1016/j.imavis.2024.105337
[31] J. Zhang *et al.*, "Advances in attention mechanisms for medical image segmentation," *Comput. Sci. Rev.*, vol. 56, p. 100721, 2025. https://doi.org/10.1016/j.cosrev.2024.100721
[32] T. A. Aris, A. S. A. Nasir, W. A. Mustafa, M. Y. Mashor, E. V. Haryanto, and Z. Mohamed, "Robust Image Processing Framework for Intelligent Multi-Stage Malaria Parasite Recognition of Thick and Thin Smear Images," *Diagnostics*, vol. 13, no. 3, 2023. https://doi.org/10.3390/diagnostics13030511

[33] M. P. Singh *et al.*, "A Healthcare System Employing Lightweight CNN for Disease Prediction with Artificial Intelligence," *Open Public Health J.*, vol. 17, no. 1, Jul. 2024. https://doi.org/10.2174/0118749445302023240520111802