From Binary Labels to Stage-Aware Bounding Boxes: Evaluating Annotation Depth in Malaria Microscopy

Frances Adelakun [1], Comfort Adesina[2], Olamide Lawal[3], Olatokun Shamsudeen Akano[4], Stephen Adeyinka Adeniyi[1], and Abdulbasit Oyetunji[4]

[1]Federal University of Agriculture Abeokuta, Nigeria
[2]Machine Learning Collective
[3]Obafemi Awolowo University, Nigeria
[4]University of Ibadan, Nigeria
[5]Abiola Ajimobi Technical Univeristy, Ibadan, Nigeria
francesadelakun127@gmail.com

Abstract. Public malaria microscopy datasets use different image scales, task formulations, and annotation schemes, complicating direct comparison. We evaluate annotation depth using three label tiers derived from identical BBBC041 Malaria Bounding Boxes Dataset (BBBC041) images and bounding boxes: binary detection (L1), parasite detection (L2), and stage-aware detection (L3). YOLO11s models were trained using a matched protocol across three seeds and evaluated using native metrics, class-agnostic localization, confirmed-parasite detection, and cross-dataset transfer with the Lacuna thin-smear subset. Class-agnostic localization remained similar across tiers (AP50: 0.979–0.984), while deeper labels enabled stage-specific outputs without consistently improving confirmed-parasite detection. Transfer was substantially weaker: BBBC041 malaria dataset -to-Lacuna parasite AP50 ranged from 0.024 to 0.042, whereas Lacuna-to-BBBC041 malaria dataset reached 0.319. Annotation depth therefore determines output granularity but does not independently ensure better shared-endpoint performance or external portability.

Keywords: Malaria Diagnosis · Annotation Granularity · Cross-Dataset Generalization · Object Detection · Computer-Aided Diagnosis

1 Introduction

Malaria caused an estimated 610,000 deaths in 2024, predominantly in sub-Saharan Africa [1]. Blood-smear microscopy remains central to diagnosis but depends on specimen quality and trained personnel [2]. Computer-aided diagnosis may support microscopy workflows, although model performance can decline under changes in staining, acquisition, and clinical site [3].

Public malaria datasets represent related tasks using different image scales and annotations. NIH Malaria Cell Images Dataset provides binary labels for cropped cells, Lacuna provides parasite and white-blood-cell bounding boxes in

full-smear images, and BBBC041 provides stage-aware bounding boxes [11,12,4]. Cross-dataset comparisons cannot isolate annotation depth because the images, acquisition conditions, and task formulations also change.

We therefore construct three BBBC041 Malaria Bounding Boxes Dataset label tiers using identical images, bounding boxes, splits, and training settings. Our contributions are: (1) a controlled hierarchy for evaluating binary, parasite-level, and stage-aware supervision; (2) shared localization and confirmed-parasite endpoints for direct cross-tier comparison; and (3) NIH Malaria Cell Images Datasets and Lacuna Malaria Blood Smear stress tests assessing portability under compound domain shifts.

2 Related Work

Studies using the NIH Malaria cropped-cell dataset have reported strong held-out classification performance with convolutional, transfer-learning, and hybrid CNN–transformer models [5,6,7]. However, crop classification differs from parasite localization in full-smear images.

Full-field malaria detection has been studied using Faster R-CNN, YOLO-family detectors, and attention-based architectures [4,9,10]. Dataset quality, acquisition differences, and clinical-site shifts remain barriers to external generalization [3,8]. Our study holds BBBC041 Malaria Bounding Boxes Dataset images and bounding-box coordinates constant to examine label granularity, while treating NIH and Lacuna evaluations as external stress tests rather than controlled causal comparisons.

3 Method

3.1 Supervision Regime Definition

We evaluate three annotation regimes. NIH provides binary labels for pre-cropped red blood cells without localization. Lacuna provides parasite and white-blood-cell bounding boxes in full-smear images. BBBC041 Malaria Bounding Boxes Dataset provides stage-aware bounding boxes. Using identical BBBC041 images and coordinates, we construct three tiers: L1 distinguishes infected from uninfected objects; L2 distinguishes infected cells, uninfected red blood cells, and leukocytes; and L3 retains all seven native classes. Only comparisons among these BBBC041 tiers isolate annotation granularity because the images, object locations, splits, architecture, and training settings are held constant.

3.2 Datasets and Label Harmonization

NIH Malaria Cell Images Dataset. NIH contains 27,558 Giemsa-stained thin-smear RBC crops from 200 patients [11]. A patient-grouped 70/15/15 split assigned 140/30/30 patients and 19,637/3,987/3,934 images to training, validation, and testing, respectively, with no overlap. NIH supports binary crop classification rather than full-image detection.

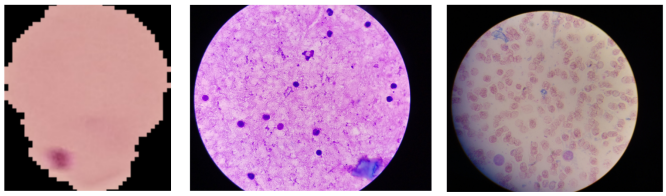

Fig. 1. The three datasets present the same biological signal at incompatible scales. (a) NIH Malaria Cell Images Dataset: a single pre-cropped cell (∼130 px). (b) BBBC041 Malaria Bounding Boxes Dataset: a full-field thin-smear image (1600×1200 px). (c) Lacuna Malaria Thin Blood Smear Dataset: a full-field smartphone-captured thin smear.

Lacuna Malaria Blood Smear Dataset. We used Lacuna Malaria Blood Smear Dataset all-thin, a curated subset containing 1,011 Ghana and 1,003 Uganda thin-smear images [12]. Gametocyte, trophozoite, ring, and other-stage annotations were merged into parasite; WBC annotations were retained, and artefacts were excluded. The resulting dataset contained 48,681 boxes—47,211 parasites and 1,470 WBCs—and nine valid background images. A source-stratified split produced 1,410 training, 302 validation, and 302 held-out test images.

Broad Bioimage Benchmark Collection image set BBBC041. BBBC041 Malaria Bounding Boxes Dataset contains 1,328 images and 86,035 bounding boxes across seven native classes [13]. The official 120-image test set was held out, and the remaining images were divided into 1,026 training and 182 validation images with rare-class stratification. Patient identifiers were unavailable.

The three BBBC041 Malaria Bounding Boxes Dataset tiers retain the same images, split, and box coordinates. L3 preserves all native classes. In L2, the parasite stages and difficult are merged into infected, while RBC and leukocyte remain separate. In L1, RBC and leukocyte are merged into uninfected, leaving infected and uninfected. Mapping leukocytes to uninfected is an operational binary mapping, not a biological classification. Difficult objects are retained as infected during L1/L2 training to keep the boxes constant but are ignored in the confirmed-parasite evaluation because their infection status is uncertain.

Table 1. BBBC041 Malaria Bounding Boxes Dataset native class distribution and mapping across the three label tiers.

Native class	Count	Proportion (%)	L3	L2	L1	Confirmed-parasite endpoint
Red blood cell	83,034	96.51	red blood cell	uninfected RBC	uninfected	Excluded
Trophozoite	1,584	1.84	trophozoite	infected	infected	Parasite
Ring	522	0.61	ring	infected	infected	Parasite
Difficult	446	0.52	difficult	infected	infected	Ignored
Schizont	190	0.22	schizont	infected	infected	Parasite
Gametocyte	156	0.18	gametocyte	infected	infected	Parasite
Leukocyte	103	0.12	leukocyte	leukocyte	uninfected	Excluded

3.3 Model Architecture and Training

NIH Malaria Cell Images Dataset classification used an ImageNet-pretrained ResNet-18 with a two-class head (224×224 inputs, ImageNet normalization, random flips, ColorJitter), optimized using cross-entropy loss and Adam (learning rate 0.001, batch size 32, up to 30 epochs, patience 5).

Detection experiments evaluated COCO-pretrained single-stage YOLO architectures (YOLO11s across all BBBC041 Malaria Bounding Boxes Dataset tiers and YOLO12s on L3; Ultralytics 8.3.100) across seeds 42, 43, and 44 using YOLO11s as the common controlled architecture. Detection runs used AdamW (learning rate 0.001667, weight decay 0.0005, batch size 16, 640×640 inputs, up to 150 epochs, patience 20) with HSV jitter, translation, scaling, flips, mosaic, and MixUp augmentations. Lacuna Malaria Blood Smear Dataset models used the identical protocol.

Because RBCs comprise 96.51% of BBBC041 Malaria Bounding Boxes Dataset objects, native metrics are supplemented with shared localization and confirmed-parasite endpoints. Primary tier comparisons maintained an unweighted loss across L1–L3 to isolate annotation depth. Primary tier comparisons maintained an unweighted loss across L1–L3 to isolate annotation depth.

3.4 Cross-Dataset Evaluation Protocol

Native multiclass AP is reported within each BBBC041 Malaria Bounding Boxes Dataset tier but is not compared directly across tiers because their class spaces differ. Cross-tier evaluation instead uses two shared endpoints: class-agnostic localization AP, for which all classes are merged, and confirmed-parasite AP, for which ring, trophozoite, schizont, and gametocyte are positives and difficult objects are ignored.

For NIH Malaria Cell Images Dataset evaluation, BBBC041 Malaria Bounding Boxes Dataset L1 detections are converted to crop-level predictions using the class of the highest-confidence detection; images without a retained detection are assigned uninfected. This is a crop-level external stress test, not end-to-end detector transfer.

For BBBC041 Malaria Bounding Boxes Dataset-to-Lacuna Malaria Blood Smear Dataset evaluation, parasite-stage predictions are merged into parasite, leukocyte maps to WBC, and RBC predictions are excluded because Lacuna Malaria Blood Smear Dataset does not annotate uninfected RBCs. In the reciprocal evaluation, Lacuna Malaria Blood Smear Dataset parasite and WBC predictions are scored against the corresponding BBBC041 Malaria Bounding Boxes Dataset targets, while RBC objects are ignored. These experiments assess transfer between aligned output labels, not the isolated causal effect of annotation depth across datasets.

4 Experiments and Results

4.1 Within-Dataset Performance

Unless otherwise stated, detection results are reported as mean ± standard deviation across seeds 42, 43, and 44. Model selection used the validation set, and the held-out test set was evaluated after training.

NIH Malaria Cell Images Dataset classification. The patient-grouped split contained 140 training, 30 validation and 30 test patients, corresponding to 19,637, 3,987 and 3,934 cell images, respectively, with no patient overlap. The ResNet-18 classifier selected at epoch 13 achieved 97.03% accuracy, 96.36% sensitivity, 97.63% specificity, an F1 score of 0.9686 and an AUROC of 0.9947 on the grouped test set.

Lacuna Malaria all-thin detection. Table 2 summarizes YOLO11s performance across three seeds on the 302-image test set. White-blood-cell AP exceeded parasite AP despite substantially fewer annotations, indicating that annotation frequency alone did not determine class-level performance.

Table 2. YOLO11s performance on the Lacuna Malaria Blood Smear Dataset all-thin test set across three seeds.

Metric	Score	Metric	Score
mAP@0.5	0.5173 ± 0.0035	Parasite AP@0.5	0.2901 ± 0.0027
mAP@0.5:0.95	0.2464 ± 0.0012	Parasite AP@0.5:0.95	0.1178 ± 0.0014
Precision	0.5885 ± 0.0166	WBC AP@0.5	0.7445 ± 0.0056
Recall	0.5217 ± 0.0085	WBC AP@0.5:0.95	0.3750 ± 0.0026

BBBC041 Malaria Bounding Boxes Dataset tiered detection. Table 3 reports the three BBBC041 Malaria Bounding Boxes Dataset supervision tiers evaluated on the same 120-image test set. L1 distinguishes infected from uninfected cells, L2 separates infected cells, uninfected red blood cells, and leukocytes, and L3 preserves the seven original classes, including individual parasite stages. Native mAP describes performance within each label space but is not directly comparable across tiers because the number and meaning of the classes change. Cross-tier conclusions are therefore based on the class-agnostic localization and confirmed-parasite endpoints computed from the same bounding boxes.

4.2 Cross-Dataset Generalization

Cross-dataset evaluation showed substantial asymmetry. Importantly, the NIH Malaria Cell Images Dataset-to-BBBC041 Malaria Bounding Boxes Dataset experiment used ground-truth boxes to provide cell crops and therefore measures externally transferred classification under oracle localization. It is not directly equivalent to the end-to-end detector-transfer experiments.

BBBC041 Malaria Bounding Boxes Dataset L1 → NIH Malaria Cell Images Dataset. The BBBC041 Malaria Bounding Boxes Dataset L1

Table 3. BBBC041 Malaria Bounding Boxes Dataset results across three seeds. Values are mean ± standard deviation.

Tier	Model	Loc. AP50	Loc. AP50:95	Conf. Parasite AP50	Conf. Parasite AP50:95
L1	YOLO11s	0.9835 ± 0.0004	0.8318 ± 0.0032	0.5530 ± 0.0364	0.4600 ± 0.0342
L2	YOLO11s	0.9793 ± 0.0064	0.8203 ± 0.0101	0.5230 ± 0.0958	0.4220 ± 0.0935
L3	YOLO11s	0.9829 ± 0.0008	0.8262 ± 0.0101	0.4812 ± 0.0914	0.3976 ± 0.0811
L3	YOLO12s	0.9871 ± 0.0008	0.8246 ± 0.0089	0.5633 ± 0.0636	0.4547 ± 0.0497

Table 4. Cross-dataset detection performance. Values are mean ± standard deviation across seeds 42–44.

Source → Target	Model	Parasite AP50	Parasite AP50–95
BBBC041 Malaria Bounding Boxes Dataset L1 → Lacuna Malaria Blood Smear Dataset all-thin	YOLO11s	0.0419 ± 0.0156	0.0208 ± 0.0066
BBBC041 Malaria Bounding Boxes Dataset L2 → Lacuna Malaria Blood Smear Dataset all-thin	YOLO11s	0.0303 ± 0.0044	0.0122 ± 0.0009
BBBC041 Malaria Bounding Boxes Dataset L3 → Lacuna Malaria Blood Smear Dataset all-thin	YOLO11s	0.0237 ± 0.0001	0.0105 ± 0.0005
BBBC041 Malaria Bounding Boxes Dataset L3 → Lacuna Malaria Blood Smear Dataset all-thin	YOLO12s	0.0363 ± 0.0131	0.0192 ± 0.0080
Lacuna Malaria Blood Smear Dataset all-thin → BBBC041 Malaria Bounding Boxes Dataset	YOLO11s	0.3192 ± 0.0963	0.2124 ± 0.0560

detector produced zero true-positive infected predictions across the 4,134 NIH Malaria Cell Images Dataset test images, corresponding to 0% sensitivity. Because all uninfected images were classified correctly and the NIH Malaria Cell Images Dataset test set was balanced, the resulting 50% accuracy reflects prediction of only the uninfected class rather than useful transfer. No detection was produced in 41.8% of the images. The BBBC041 Malaria Bounding Boxes Dataset source images are full-field micrographs, whereas the NIH Malaria Cell Images Dataset images contain individual cropped cells. The result is consistent with acquisition, framing, and object-scale mismatch, although the experiment does not isolate scale as the sole cause.

NIH Malaria Cell Images Dataset → BBBC041 Malaria Bounding Boxes Dataset. Ground-truth bounding boxes were used to extract 5,922 cell crops from the 120 BBBC041 Malaria Bounding Boxes Dataset test images before applying the NIH Malaria Cell Images Dataset classifier. The classifier achieved 94.48% accuracy, 96.75% sensitivity, and an AUROC of 0.9885. Precision was 48.46%, reflecting the approximately 18:1 uninfected-to-infected distribution in the BBBC041 Malaria Bounding Boxes Dataset test crops. These results demonstrate strong crop-level classification transfer when localization is supplied, but they do not establish end-to-end detection generalization.

Bidirectional detector transfer. Evaluating the Lacuna Malaria Blood Smear Dataset all-thin and BBBC041 Malaria Bounding Boxes Dataset L2 models bidirectionally on harmonized parasite and leukocyte classes (excluding unannotated RBCs) revealed marked asymmetry. The BBBC041 L2 model yielded a parasite AP_{50} of only 0.0303 ± 0.0044 on Lacuna, whereas the Lacuna-trained YOLO11s achieved 0.3192 ± 0.0963 on BBBC041 Malaria Bounding Boxes Dataset(Table 4). This asymmetry demonstrates that label harmonization alone does not guarantee symmetric or reliable cross-dataset generalization.

5 Discussion

The controlled BBBC041 comparison demonstrated that class-agnostic localization remained stable across annotation tiers. While confirmed-parasite AP_{50} means varied across L1, L2, and L3, these differences were minor relative to variance across the three seeds. Consequently, deeper annotation did not systematically improve or degrade performance on the shared parasite-detection endpoint under matched experimental conditions.

Richer annotation primarily expanded the granularity of model outputs rather than core detection accuracy. L1 supported binary infected-versus-uninfected localization, L2 separated parasites from non-parasitic cells, and L3 enabled stage-specific predictions. However, severe class imbalance among rare parasite stages prevents aggregate L3 metrics from confirming dependable recognition across all life cycles. Stage-wise performance must therefore be interpreted cautiously, and L3 should be viewed as enabling stage-aware outputs rather than ensuring robust detection of every stage.

Cross-dataset generalization proved weak and asymmetric. In the BBBC041 Malaria Bounding Boxes Dataset-to-Lacuna Malaria Thin Blood Smear Dataset evaluation, increasing annotation depth did not improve transferred parasite detection. Conversely, the NIH classifier transferred successfully to BBBC041 Malaria Bounding Boxes Dataset only when supplied with ground-truth cell crops, whereas full-image transfer failed entirely. Because the datasets differ across acquisition protocols, staining conditions, task formulations, and parasite species (*P. vivax* in BBBC041 Malaria Bounding Boxes Dataset versus *P. falciparum* in others), these transfer failures reflect compound domain shifts rather than the causal effect of any single factor.

The three-seed design is insufficient to establish a definitive statistical ordering among annotation tiers. BBBC041 contains *P. vivax*, while the motivating African deployment context is dominated by *P. falciparum*. Both evaluated detector variants belong to the single-stage YOLO family, so the findings may not extend to two-stage or other detector families.

6 Impact in RCS

Annotation depth should match the intended clinical workflow: binary cell labels suffice for screening pre-segmented cells, bounding boxes are required for localization and quantification, and stage-specific labels are justified only when stage-level outputs are needed and rare classes are adequately sampled.

Future dataset curation must systematically preserve acquisition metadata including staining protocols, optical settings, and parasite species, while structuring evaluation splits grouped by patient or clinical site. Because deeper labels alone failed to overcome compound domain shifts, cross-site acquisition standards and target-domain validation should take precedence over granular annotation investments. These insights serve as dataset design principles rather than formal cost-effectiveness claims.

7 Conclusion

The matched evaluation demonstrated that richer annotations expanded model output capabilities without establishing a consistent advantage on the shared parasite-detection endpoint. Cross-dataset performance remained weak and dataset-dependent under compound domain shifts. Annotation depth should therefore be selected according to the intended output, while portability must be established separately using grouped, target-domain evaluation.

Acknowledgements.

Acknowledgments

The authors thank the 2026 SPARK Academy instructors for technical guidance, Aondana Iorumbur and Raymond Confidence for their valuable support and contributions, the Digital Research Alliance of Canada for computational resources, and the McGill University Doctoral Internship program for knowledge translation support. This work was supported by the Lacuna Fund for Health and Equity, the RSNA R&E Foundation Derek Harwood-Nash International Education Scholar Grant, and the NSERC Discovery Launch Supplement.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. World Health Organization: World Malaria Report 2025. Geneva: World Health Organization (2025). https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024
2. Centers for Disease Control and Prevention: Evaluation and Diagnosis of Malaria. CDC Clinical Guidance (2026). https://www.cdc.gov/malaria/hcp/clinical-guidance/evaluation-diagnosis.html
3. Guillon, L., et al.: Towards Field-Ready AI-based Malaria Diagnosis: A Continual Learning Approach. arXiv preprint arXiv:2507.23648 (2025)
4. Hung, J., et al.: Applying Faster R-CNN for Object Detection on Malaria Images. In: CVPR Workshops, pp. 808–813 (2017). https://doi.org/10.1109/cvprw.2017.112
5. Hemachandran, K., et al.: Performance Analysis of Deep Learning Algorithms in Diagnosis of Malaria Disease. Diagnostics 13(3), 534 (2023). https://doi.org/10.3390/diagnostics13030534
6. Ahishakiye, E., et al.: Enhancing malaria detection and classification using convolutional neural networks–vision transformer architecture. Discover Applied Sciences 7, 612 (2025). https://doi.org/10.1007/s42452-025-06704-z
7. Laghari, A.A., Muhammad, W., Memon, M.L., Hussain, A., Kumar, A.: Malaria Parasite Cell Classification Using Transfer Learning with State-of-Art CNN Architectures. Biology 14(12), 1792 (2025). https://doi.org/10.3390/biology14121792

8. Kabore, K.K., Guel, D.: Addressing Challenges in Data Quality and Model Generalization for Malaria Detection. J. Sens. Netw. Data Commun. 4(3) (2024). https://doi.org/10.33140/JSNDC.04.03.09
9. Koirala, A., et al.: Deep Learning for Real-Time Malaria Parasite Detection and Counting Using YOLO-mp. IEEE Access 10, 102157–102172 (2022). https://doi.org/10.1109/ACCESS.2022.3208270
   Zedda, L., Loddo, A., Di Ruberto, C.: A deep architecture based on attention mechanisms for effective end-to-end detection of early and mature malaria parasites in a realistic scenario. Comput. Biol. Med. 186, 109704 (2025). https://doi.org/10.1016/j.compbiomed.2025.109704
10. A deep architecture based on attention mechanisms for effective end-to-end detection of early and mature malaria parasites in a realistic scenario. Comput. Biol. Med. 186, 109704 (2025). https://doi.org/10.1016/j.compbiomed.2025.109704
11. Rajaraman, S., et al.: Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ 6, e4568 (2018). https://doi.org/10.7717/peerj.4568
12. Makerere AI Lab: Lacuna Malaria Datasets. Harvard Dataverse (2023). https://doi.org/10.7910/DVN/VEADSE
13. Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nat. Methods 9, 637 (2012). https://doi.org/10.1038/nmeth.2083; BBBC041: https://bbbc.broadinstitute.org/BBBC041