On-Device Multi-Species Malaria Detection with Uncertainty-Calibrated Slide-Level Aggregation

Idaya Seidu[1], Ahmed Tahiru Issah[1], Charles B. Delahunt[2], and Carine Mukamakuza[1⋆]

[1] Carnegie Mellon University Africa, Kigali Innovation City, Kigali, Rwanda
{iseidu,aissah,cmukamak}@andrew.cmu.edu
[2] University of Washington, Seattle, Washington, USA
delahunt@uw.edu

Abstract. Malaria remains a leading cause of mortality in resource-limited settings, where expert microscopists are scarce. Automated diagnosis based on microscopy images thus has strong potential to improve care delivery. But for an algorithm to deploy, a necessary requirement is that it meet a suite of non-obvious (from a machine learning (ML) perspective) clinical constraints. Therefore, in close consultation with the Rwandan Biomedical Center, we developed a malaria diagnosis pipeline which addresses key requirements listed by the health care center but typically ignored in the ML malaria literature. In particular, it includes: (i) stopping criteria (to reduce image acquisition and time-to-result); (ii) human-in-the-loop functionality (for review and accountability); (iii) multi-species discrimination (since treatment varies by species); (iv) thick film detection (standard for microscopy); (v) computationally-efficient uncertainty calculations (to aid clinician review); and (vi) an edge device platform (since internet can be spotty in this catchment area). The mobile system performs inference on-device using YOLOv13n deployed via TensorFlow Lite. It detects four species and white blood cells from Giemsa-stained thick blood smear images, aggregating per-image detections into slide-level parasitemia with World Health Organization (WHO)-standard quantification. This paper highlights these various clinical constraints and offers methods to address them. Evaluated on 2,739 annotated images across all four species, the system achieves mAP@0.5 of 0.863, per-image parasite count correlation of $r = 0.812$, slide-level $r = 0.951$ (soft counting, 10 images/slide), and runs entirely offline with a pipeline time of 10.27 ± 1.65 s per image.

Keywords: Malaria detection · Object detection · Mobile deep learning · Uncertainty quantification · YOLO · Point-of-care diagnostics

1 Introduction

Malaria is responsible for over 600,000 deaths annually, with 95% of cases in sub-Saharan Africa [19]. The gold standard for diagnosis remains manual microscopy

⋆ Corresponding author: cmukamak@andrew.cmu.edu

of Giemsa-stained blood smears, which requires trained microscopists who are scarce in endemic regions [15]. Rapid diagnostic tests (RDTs) offer an alternative but cannot quantify parasitemia or reliably distinguish *Plasmodium* species, both critical for treatment decisions [10].

Deep learning has shown promising results for malaria detection [13, 5, 9], but most ML targeting malaria has been developed in an ML-centric, rather than clinic-centric perspective. For example, most systems require either cloud servers with reliable connectivity or high-performance compute hardware. In addition, the majority focus on single-cell classification rather than slide-level diagnosis, and few address parasitemia quantification with uncertainty estimates [11]. Mobile-optimized detectors such as YOLO [16] combined with TensorFlow Lite [6] now enable on-device deployment, but deployment raises open challenges regarding how to aggregate per-image detections into a reliable slide-level diagnosis, quantify uncertainty in the parasitemia estimate, and integrate human expert oversight efficiently.

Therefore we present a mobile system for malaria diagnosis, designed in consultation with the Rwandan Biomedical Center, a national health center with a large malaria-endemic catchment, to address several crucial clinical requirements. Contributions include:

1. **Uncertainty-calibrated slide-level aggregation** via soft counting with validation-derived calibration, reducing systematic bias by a factor of 3 (vs hard counts) at all slide sizes while improving Pearson r and RMSE across all configurations.
2. **Quantitation confidence intervals (CIs)** via a computationally-efficient bootstrap resampling strategy over per-image statistics to compute 95% CIs for slide-level parasitemia.
3. **An early stopping criterion**, keyed to estimated parasitemia and uncertainty, signaling when sufficient images have been processed, which speeds clinical workflow.
4. **Human-in-the-loop (HITL)** interface with uncertainty-guided prioritization to support clinician oversight, review, and feedback (i.e. ML in a decision support role).
5. **Lightweight ML models** to enable edge deployment in clinics without reliable internet or expensive compute resources.

2 Related Work

Deep Learning for Malaria Detection. Early CNN approaches classified individual cells [13, 5] without providing slide-level counts. Object detection frameworks (Faster R-CNN [14], YOLO [1, 8]) improved field-of-view analysis, but most work evaluates only per-image performance without slide-level aggregation for clinical diagnosis [20, 9]. Current SOTA systems perform well, but require expensive scanning microscopes and compute power that are beyond the budgets of most front-line clinics [7, 2]. The review by [11] describes further common problems with ML malaria efforts.

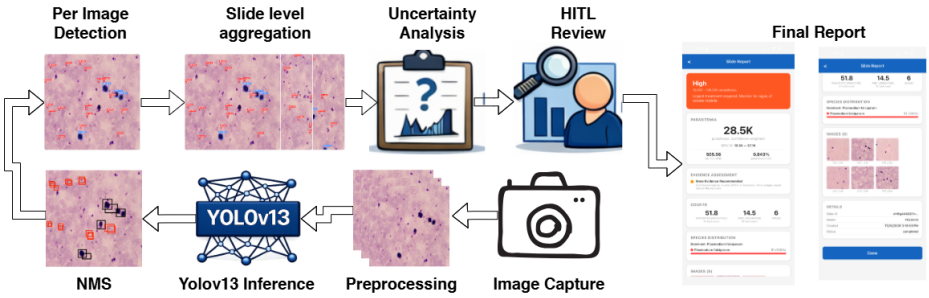

Fig. 1. End-to-end pipeline from image capture to uncertainty-calibrated slide report.

Mobile Deployment and Uncertainty. Quinn et al. [12] deployed thick-smear analysis on tablets but required server processing. Newer generations of lightweight models, e.g. TensorFlow Lite [6], enable on-device inference, yet no prior system (to our knowledge) combines multi-species detection with uncertainty-calibrated slide-level quantification on consumer hardware.

3 Methods

3.1 System Architecture

Our system is implemented as a cross-platform mobile application using React Native with custom native modules for TensorFlow Lite inference (Fig. 1). The architecture comprises four stages. (1) Image acquisition via camera or gallery import. (2) On-device inference using YOLOv13n with native TFLite modules. (3) Uncertainty-calibrated slide-level aggregation. (4) Human-in-the-loop review with uncertainty-guided prioritization. Images are stored locally; all processing occurs without network connectivity, with inference delegated to platform-specific TFLite interpreters (Kotlin on Android, Swift on iOS). Local image storage simplifies patient privacy logistics.

3.2 Clinical Requirements and Stakeholder Engagement

Requirements were informed through three iterative workshops with approximately ten clinicians at the Rwanda Biomedical Centre: five microscopists, some with over 20 years of thick-smear diagnostic experience; two physicians; and three laboratory technicians. Their feedback guided the HITL interface and clinical workflow integration, while the development team determined model architecture and performance criteria. Clinicians prioritised reliable species identification and parasitemia quantification in practice over model complexity.

3.3 Detection Model

We employ YOLOv13n [16] trained on Giemsa-stained thick blood smear images to detect five classes: *P. falciparum* (PF), *P. vivax* (PV), *P. malariae* (PM),

P. ovale (PO), and white blood cells (WBC). The model is exported to TensorFlow Lite format at 2048×2048 input resolution. Input images undergo letterbox resizing with gray padding to preserve aspect ratio. Post-processing implements DFL decoding (softmax over regression bins + dist2bbox), sigmoid on class logits, and per-class NMS (IoU 0.7, confidence threshold 0.25).

3.4 Uncertainty-Calibrated Slide-Level Aggregation

To provide a patient-level recommendation, object-level detections must be aggregated across multiple images from the patient.

Confidence-Weighted Soft Counting. Traditional parasitemia estimation treats each detection as a binary count: $N_p = \sum_i \mathbb{1}[\text{class}_i \neq \text{WBC}]$. We instead weight each detection by its confidence:

$$\tilde{N}_p = \sum_{i:\text{class}_i \neq \text{WBC}} c_i, \quad \tilde{N}_w = \sum_{i:\text{class}_i = \text{WBC}} c_i \tag{1}$$

where $c_i \in [0, 1]$ is the confidence of detection i. This approach naturally down-weights borderline detections while fully counting high-confidence ones. We note that WBC quantification functionality is vital because examined blood volumes (and thus parasitemias) are determined by WBC count [17, 18]:

$$\tilde{P} = \frac{\tilde{N}_p}{\tilde{N}_w} \times 8{,}000 \quad (\text{parasites}/\mu\text{L}) \tag{2}$$

Since $c_i \in (0, 1)$, raw soft counts systematically undercount relative to binary ground-truth annotations. We correct for this by deriving calibration factors on the validation set only: $\alpha_p = N_p^{\text{GT}}/\tilde{N}_p^{\text{val}} = 1.306$ for parasites and $\alpha_w = N_w^{\text{GT}}/\tilde{N}_w^{\text{val}} = 1.128$ for WBCs (at confidence threshold 0.05). All slide-level estimates use calibrated counts $\hat{N} = \alpha \cdot \tilde{N}$. Both α values are locked after validation and applied unchanged to the test set, ensuring no data leakage.

Bootstrap Confidence Intervals. Uncertainty estimates are a valuable tool for clinical interpretation of algorithm results. To quantify quantitation reliability at low computational cost, we use bootstrap resampling [4]: given K analyzed images with per-image soft counts $(\tilde{n}_p^{(k)}, \tilde{n}_w^{(k)})$, for each of $B = 1{,}000$ iterations we draw K images *with replacement*, sum the soft counts, and compute parasitemia via Eq. 2. The 95% CI is defined by the 2.5th and 97.5th percentiles of the resulting distribution. A narrow CI indicates stable estimation; a wide CI signals that more images are needed.

Evidence Sufficiency Criterion. High patient loads require minimizing image collection time and time-to-result. To enable this on a per-patient basis, the system monitors bootstrap CI width during image acquisition and determines

when enough images have been analyzed. Two conditions must hold: (1) at least $K_{\min} = 3$ images analyzed, and (2) CI precision:

$$\tilde{P}_{\text{upper}} - \tilde{P}_{\text{lower}} < \max\Big(0.3 \times \tilde{P},\ 500\Big) \tag{3}$$

The 30% relative threshold ensures proportional precision for high-parasitemia cases, while the absolute floor of 500 parasites/µL prevents the criterion from being unreachable in low-parasitemia cases with high Poisson variability [3]. The practical time savings depend on the endemic parasitemia distribution at the deployment site: for high-parasitemia samples, the CI narrows quickly and acquisition terminates early; for negative or low-parasitemia samples, more images are required before the criterion is satisfied. Clinics with higher rates of high-parasitemia cases will therefore benefit most from early stopping.

Human-in-the-Loop Review. Clinician interaction and oversight are essential for accountability and quality control during system deployment. To enable this, the HITL interface (Fig. 1) supports soft-deletion of false positives, species re-labeling, and addition of missed detections via tap-to-place. These edits trigger live re-aggregation, allowing the clinician to correct algorithm outputs.

Uncertainty-Guided Prioritization. To focus clinician's review, each image receives a priority score to surface the most uncertain fields (i.e. those most likely to contain algorithm errors):

$$s_k = 0.4\,(1 - \bar{c}_k) + 0.3\,\hat{d}_k + 0.2\,m_k + 0.1\,(1 - \mathbb{1}[n_w^{(k)} > 0]) \tag{4}$$

where $\bar{c}_k$ is mean detection confidence, $\hat{d}_k = |n_p^{(k)} - \bar{n}_p|/(\bar{n}_p + \epsilon)$ is the normalized count deviation from the slide mean, m_k indicates multi-species detections, and the final term flags absent WBC detections. Images are grouped into high ($s_k \geq 0.5$), medium (≥ 0.25), and low priority tiers.

4 Experimental Setup

Dataset The dataset consists of 2,739 Giemsa-stained blood smear images from the Rwandan Biomedical Center, spanning four *Plasmodium* species and WBCs: PF (838 images, 7,568 instances), PM (834 images, 1,802 instances), PO (893 images, 2,353 instances), and PV (174 images, 669 instances). Images were auto-oriented, resized to 2048×2048 via letterbox padding, and split 70/15/15% for train/val/test. To counter PV class imbalance (6.4% of images), we applied targeted augmentation (rotation, hue/saturation/brightness variation) to the training set only.

Training Configuration YOLOv13n was trained on a single NVIDIA H100 GPU for 70 epochs (batch 6, SGD lr=0.01, CIoU + DFL loss, mosaic augmentation).

Table 1. YOLOv13n detection performance on the test set (410 images).

Class	Precision	Recall	mAP@50	mAP@75	mAP@50–95
All classes	0.813	0.824	0.863	0.671	0.626
PF	0.749	0.665	0.765	0.577	0.508
PM	0.818	0.840	0.871	0.649	0.631
PO	0.842	0.816	0.880	0.783	0.705
PV	0.844	0.860	0.903	0.525	0.525
WBC	0.814	0.936	0.896	0.824	0.760

Evaluation Metrics We report per-class precision, recall, mAP@50, mAP@75, and mAP@50–95. Per-image count accuracy compares model-predicted counts against ground-truth (GT) annotations across all 410 test images via Pearson r, MAE, and RMSE. GT annotations are binary integer counts, so hard counts (threshold 0.25) enable direct comparison; soft counting is evaluated at the slide level where both methods produce a comparable parasitemia estimate.

Slide-level accuracy uses simulated slides: for $K \in \{5, 10, 15\}$ images/slide, we sample 50 slides (fixed seed, without replacement from the 410 test images), compute parasitemia via the WHO formula, and measure empirical 95% CI coverage.

Mobile Performance On-device benchmarking was performed on a consumer smartphone (A15 Bionic, 6 GB RAM) running iOS, CPU backend with 2 threads, across all 410 test images.

5 Results

5.1 Detection Performance

Table 1 summarizes detection performance on the test set. The model achieves an overall mAP@50 of 0.863 and mAP@50–95 of 0.626. PF yields the lowest mAP@50 (0.765) due to its small ring-form morphology, while PV achieves the highest parasite mAP@50 (0.903) despite being the most underrepresented class, validating the targeted augmentation strategy. WBC detection achieves the highest recall (0.936) and mAP@50–95 (0.760), which is critical due to WBCs' role in the WHO parasitemia formula.

5.2 Slide-Level Parasitemia Estimation

Table 2 presents per-image count accuracy. The overall parasite count correlation is $r = 0.812$ with a 5.4% over-count (1,785 predicted vs. 1,694 GT), while WBC correlation is $r = 0.947$ with a 20.4% over-count (615 vs. 511).

Table 3 compares hard counting and calibrated soft counting across slide sizes. Soft counting consistently outperforms the hard-count baseline: Pearson r

Table 2. Per-image count accuracy: predictions vs. ground truth (410 test images).

Class	GT total	Pred total	Pearson r	MAE
PF	1,085	1,144	0.850	1.08
PM	200	215	0.934	0.12
PO	316	330	0.887	0.21
PV	93	96	0.948	0.06
WBC	511	615	0.947	0.33
All parasites	1,694	1,785	0.812	1.43

Table 3. Slide-level parasitemia estimation: soft counting vs. hard-count baseline (50 simulated slides per configuration, seed=42). MPE = $(\hat{P}_{\mathrm{GT}}-\hat{P})/\hat{P}_{\mathrm{GT}}\times 100$; positive = underprediction. Hard-count threshold (0.25) and soft-count calibration factors (α_p=1.306, α_w=1.128, threshold 0.05) were fixed on the validation set.

K	Method	Pearson r	MPE(%)	RMSE	MAE
5	Hard	0.845	+9.4	22,346	13,528
	Soft	**0.889**	**−0.5**	**18,477**	**11,073**
10	Hard	0.905	+15.0	15,032	8,667
	Soft	**0.951**	**+4.5**	**10,679**	**6,227**
15	Hard	0.804	+9.7	7,415	5,118
	Soft	**0.874**	**−1.8**	**5,391**	**3,661**

improves at every K (0.889 vs. 0.845, 0.951 vs. 0.905, 0.874 vs. 0.804), RMSE is lower at every K, and systematic bias is reduced from 9.4–15.0% (hard) to $\leq$4.5% (soft). Hard counting over-counts WBCs by 20.4% vs. 5.4% for parasites, deflating the parasites/WBC ratio and causing underprediction of parasitemia; soft counting down-weights borderline WBC detections proportionally, correcting this imbalance. The non-monotonic hard-count Pearson r across K (0.845, 0.905, 0.804) reflects sampling variance in 50 simulated slides; RMSE and MAE decrease monotonically, confirming consistent accuracy gains with more images.

5.3 Confidence Interval Calibration

Table 4 reports bootstrap CI calibration. Coverage increases monotonically from 80% (K=5) to 94% (K=15), approaching the nominal 95% level. The majority of misses at K=5 involve slides with $\leq$14 predicted WBCs, confirming WBC scarcity as the primary failure mode. Mean CI width narrows from 80,718 to 37,801 p/µL with more images.

5.4 Evidence Sufficiency and HITL Prioritization

The evidence sufficiency module monitors CI width during acquisition and alerts the microscopist when additional images are needed; its clinical motivation is to

Table 4. Bootstrap 95% CI calibration (50 simulated slides, 1000 iterations each). CI width in p/µL.

Images/slide	Coverage	CI width (mean)	CI width (median)
5	80.0% (40/50)	80,718	49,345
10	88.0% (44/50)	48,871	36,946
15	94.0% (47/50)	37,801	29,766

Table 5. On-device performance (A15 Bionic SoC, CPU, 2 threads, $n = 410$).

Metric	Value	Metric	Value
Model input size	2048×2048	Preprocessing time	2.15 ± 0.35 s
Peak memory usage	1.7 GB	Model invoke time	7.90 ± 1.29 s
Mean detections/image	5.9	Post-processing time	220 ± 32 ms
Model size (TFLite)	48 MB	Total per-image pipeline	10.27 ± 1.65 s

terminate acquisition as soon as the estimate is reliable, reducing unnecessary image capture in resource-limited settings. In our batch simulation, CI widths remain wide relative to the stopping threshold at all tested slide sizes, indicating that the criterion is designed for scenarios where parasitemia is high enough to yield narrow per-image counts quickly, prospective evaluation with real-time sequential acquisition is needed to quantify fields-of-view saved. The HITL module surfaces low-confidence, outlier-count, and WBC-absent images for priority review. The HITL interface was designed in direct response to the clinical requirement of expert oversight for deployed systems. It enables corrections to false positives, species labels, and missed detections with live re-aggregation. Quantitative user-study evaluations of both modules are important future work.

5.5 Mobile Performance

Table 5 summarizes on-device performance. The per-image pipeline averages 10.27 ± 1.65 s, dominated by model invocation (7.90 ± 1.29 s). Post-processing (DFL decode + NMS) takes 220 ± 32 ms; bootstrap CI is negligible (< 50 ms).

6 Discussion

We presented an on-device mobile system for multi-species malaria detection designed to specifically address key use case constraints described by our healthcare collaborators. The primary contribution lies in identifying and systematically addressing clinical deployment requirements that the ML community typically overlooks, providing ML methods to meet constraints articulated by our clinical partners. The system achieves $r = 0.812$ per-image and $r = 0.951$ slide-level parasite count correlation, operating entirely offline on consumer smartphones for deployment in resource-limited settings.

Limitations. The 2048×2048 input results in 1.7 GB peak memory and $\sim$10 s pipeline time, limiting deployment to higher-end devices. The CPU-only backend was necessitated by hardware delegate incompatibilities (CoreML crashes, Metal hangs); resolving these would substantially reduce inference time. Simulated slides sample images without patient grouping, as patient identifiers were unavailable under the clinical partners' data governance policy, making slide-level correlations a lower bound on performance with true per-patient slides. Our evaluation is limited to thick smears from a single reference laboratory; extension to thin smears and cross-site evaluation remains future work, along with prospective clinical validation against expert microscopy across diverse settings.

Prospects of Application. Clinics treating malaria in low-resource settings have operational requirements which are currently under-studied in the ML literature, but which carry strong implications for ML development teams. Guided by our healthcare system collaborators, this paper highlights these under-represented but crucial design constraints that gate deployment, and offers ML methods to address them.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdurahman, F., Fante, K.A., Aliy, M.: Malaria parasite detection in thick blood smear microscopic images using modified YOLOV3 and YOLOV4 models. BMC Bioinformatics **22**, 112 (2021). https://doi.org/10.1186/s12859-021-04036-4
2. Das, D., Vongpromed, R., Dhorda, M., et al.: Field evaluation of the diagnostic performance of easyscan go: a digital malaria microscopy device based on machine-learning. Malaria J (2022)
3. Delahunt C.B., M.C., M.P., H.: Reducing Poisson error can offset classification error: a technique to meet clinical performance requirements. Proc 4th ML4H Symposium, PMLR (2024)
4. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall/CRC (1993). https://doi.org/10.1201/9780429246593
5. Fuhad, K.F., Tuba, J.F., Sarker, M.R.A., Momen, S., Mohammed, N., Rahman, T.: Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. Diagnostics **10**(5), 329 (2020). https://doi.org/10.3390/diagnostics10050329
6. Google: TensorFlow Lite (litert) documentation. `https://ai.google.dev/edge/litert` (2026), accessed: 2026-02-25
7. Horning, M., Delahunt, C., Bachman, C., et al.: Performance of a fully automated system on a WHO malaria microscopy evaluation slide set. Malaria J (2021)
8. Krishnadas, P., Chadaga, K., Sampathila, N., Rao, S., S. K., R., Prabhu, S.: Classification of malaria using object detection models. Informatics **9**(4), 76 (2022). https://doi.org/10.3390/informatics9040076

9. Manescu, P., Shaw, M.J., Zewdie, M., Stell, B.M., Kiwanuka, N.: Expert-level automated malaria diagnosis on routine blood films with deep neural networks. American Journal of Hematology **95**(8), 883–891 (2020). https://doi.org/10.1002/ajh.25827
10. Murray, C.K., Gasser, Robert A., J., Magill, A.J., Miller, R.S.: Update on rapid diagnostic testing for malaria. Clinical Microbiology Reviews **21**(1), 97–110 (2008). https://doi.org/10.1128/CMR.00035-07
11. Poostchi, M., Silamut, K., Maude, R.J., Jaeger, S., Thoma, G.: Image analysis and machine learning for detecting malaria. Translational Research **194**, 36–55 (2018). https://doi.org/10.1016/j.trsl.2017.12.004
12. Quinn, J.A., Andama, A., Munabi, I., Kiwanuka, F.N.: Automated blood smear analysis for mobile malaria diagnosis. In: Karlen, W., Iniewski, K. (eds.) Mobile Point-of-Care Monitors and Diagnostic Device Design, pp. 115–132. CRC Press (2014)
13. Rajaraman, S., Antani, S.K., Poostchi, M., Silamut, K., Hossain, M.A., Maude, R.J., Jaeger, S., Thoma, G.R.: 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
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems. pp. 91–99. NIPS'15, MIT Press (2015)
15. Tangpukdee, N., Duangdee, C., Wilairatana, P., Krudsood, S.: Malaria diagnosis: a brief review. The Korean Journal of Parasitology **47**(2), 93–102 (2009). https://doi.org/10.3347/kjp.2009.47.2.93
16. Ultralytics: Ultralytics YOLO. `https://github.com/ultralytics/ultralytics` (2024), accessed: 2026-02-24
17. World Health Organization, Geneva, Switzerland: Basic Malaria Microscopy: Tutor's guide (2010)
18. World Health Organization, Geneva, Switzerland: Malaria Microscopy Standard Operating Procedure MM-SOP-09: Malaria Parasite Counting (2016)
19. World Health Organization: World malaria report 2023 (2023)
20. Yang, F., Poostchi, M., Yu, H., Zhou, Z., Silamut, K., Yu, J., Maude, R.J., Jaeger, S., Antani, S.: Deep learning for smartphone-based malaria parasite detection in thick blood smears. IEEE Journal of Biomedical and Health Informatics **24**(5), 1427–1438 (2020). https://doi.org/10.1109/JBHI.2019.2939121