Original Research Article

Comparative Evaluation of Mask R-CNN and YOLOv8 for Automated Malaria Parasite Detection and Life-Stage Classification in Microscopy Images

Abstract

Accurate and efficient detection of malaria parasites in stained blood smear images remains a critical challenge, particularly in resource-limited settings where expert microscopists may be unavailable. This study compares two deep learning instance segmentation models, YOLOv8 and Mask R-CNN, for automated detection, segmentation, and life-stage classification of malaria parasites in publicly available, anonymized Giemsa-stained microscopy images.

A total of 1,328 high-resolution images, annotated into seven clinically relevant categories (parasite stages and blood cell types), were used to fine-tune YOLOv8n and Mask R-CNN (ResNet-50-FPN backbone). YOLOv8 achieved higher bounding-box mAP (0.648 vs. 0.507) and mask mAP (0.624 vs. 0.511) at IoU 0.50, along with greater recall (0.72 vs. 0.12) and faster inference speed (9 ms vs. 93 ms per image). These results demonstrate YOLOv8's suitability for real-time parasite screening, while Mask R-CNN provides more detailed pixel-level segmentation that can support morphological analysis and quantitative research.

Performance metrics were further evaluated using bootstrap confidence intervals to assess statistical reliability, confirming YOLOv8's higher sensitivity across parasite stages. The findings suggest that lightweight, statistically validated deep learning models can provide reliable, scalable tools for automated malaria screening and parasite quantification, advancing the integration of AI-assisted microscopy in diagnostic and research workflows.

Keywords: Malaria, Deep Learning, Instance Segmentation, YOLOv8, Mask R-CNN, Diagnostic Microscopy, Medical Imaging

1 Introduction

Malaria is a serious disease caused by *Plasmodium* parasites transmitted through female *Anopheles* mosquitoes during a blood meal. Despite global eradication efforts, the World Health Organization reported in 2023 that more than 200 million cases and several hundred thousand deaths occur

---

each year (World Health Organization, 2024). Malaria remains a major public health challenge, particularly in resource-limited regions where it perpetuates cycles of illness and poverty. Timely and accurate diagnosis is essential, as delays or missed infections lead to worse outcomes and continued transmission. Microscopic examination of Giemsa-stained thick and thin blood smears remains the gold standard for malaria diagnosis and parasite quantification (Bronzan et al., 2008), but it is a labor-intensive process requiring skilled microscopists. Even experienced personnel can miss low-density infections or mixed-species cases, especially when multiple *Plasmodium* life stages coexist (Moody, 2002). Rapid diagnostic tests (RDTs) offer convenience but have notable limitations: they cannot quantify parasite load or reliably distinguish species, and some *P. falciparum* strains lacking the HRP2 antigen may yield false negatives (Koita et al., 2012). These challenges highlight the need for scalable, automated solutions that preserve microscopy's precision while enhancing accessibility.

Given these limitations, there is increasing interest in using artificial intelligence (AI) to automate malaria diagnosis. Deep learning based image analysis can support or even replace manual microscopy by rapidly identifying infected red blood cells and differentiating parasite life stages (Britton et al., 2016; Mujahid et al., 2024; Chibuta and Acar, 2020). Convolutional neural networks (CNNs) such as Mask R-CNN and You Only Look Once (YOLO) enable instance segmentation, which identifies and delineates individual objects using both bounding boxes and pixel-level masks. This distinction is clinically meaningful: object detection locates parasites, but segmentation quantifies their morphology and area, features critical for assessing parasite maturity, density, and infection severity (Davidson et al., 2021). Pixel-level segmentation also helps separate overlapping cells and distinguish leukocytes from artifacts, tasks that bounding boxes alone cannot perform accurately.

Mask R-CNN is a two-stage detector that extends Faster R-CNN by adding a mask prediction branch (He et al., 2017). It uses a ResNet-50 backbone with a Feature Pyramid Network (FPN) for multi-scale feature extraction. A Region Proposal Network (RPN) identifies potential object regions, followed by RoIAlign and parallel branches for classification, bounding box regression, and mask generation. This structure enables a detailed delineation of parasites and overlapping red blood cells (Brostow et al., 2009). YOLOv8, on the contrary, is a single-stage, anchor-free network that processes the entire image in one forward pass. It employs a CSP-Darknet backbone and a unified detection head that predicts class labels, bounding boxes, and masks concurrently. YOLO models are optimized for speed and efficiency; YOLOv8, in particular, achieves near real-time inference even on modest hardware, making it suitable for point-of-care diagnostics (Redmon et al., 2016).

The objective of this study was to evaluate two deep learning instance segmentation models, Mask R-CNN and YOLOv8, for automated detection, segmentation and life-stage classification of malaria parasites in Giemsa-stained thin blood smear images. We benchmarked their performance using standard metrics from the Common Objects in Context (COCO) framework. While YOLOv8 prioritizes speed and recall for rapid screening, Mask R-CNN emphasizes detailed mask prediction suitable for morphological and quantitative analyses. Integrating these complementary models into microscopy workflows could improve diagnostic speed, reliability, and accessibility. By combining detection accuracy with morphological segmentation, this work contributes to the development of automated, scalable tools for parasite quantification and malaria surveillance.

Earlier studies used traditional image processing and classical machine learning to identify infected cells. Var et al. pioneered computer vision approaches using hand-crafted features (Var and Tek, 2018). Quinn et al. reviewed methods based on feature extraction and support vector machines (SVMs), noting that while they achieved moderate accuracy, they were highly sensitive to staining variations (Quinn et al., 2018). These techniques required expert preprocessing and struggled to generalize across datasets.

With advances in deep learning, CNNs have shown superior performance in malaria detection. Yang et al. developed a two-stage CNN that ran on smartphones and achieved 93.5% accuracy on thick smears (Yang et al., 2020). Fuhad et al. proposed a lightweight CNN optimized for mobile devices, reporting 99.23% accuracy with only about 4,600 floating point operations (Fuhad et al., 2020).

Maqsood et al. benchmarked several pre-trained CNNs on the NIH thin-smear dataset, while Razin et al. integrated CNNs with YOLOv5 for parasite localization, demonstrating the potential of modern object detectors (Maqsood et al., 2021; Razin et al., 2022). These studies confirm that deep learning outperforms classical methods under controlled conditions.

Recent applications have extended CNNs to segmentation. Rajaraman et al. (Rajaraman et al., 2018) evaluated five pre-trained CNNs (AlexNet, VGG-16, ResNet-50, Xception, and DenseNet-121) for classifying parasitized versus uninfected cells, achieving accuracies of 95–96%. Narayanan et al. (Narayanan et al., 2019) compared multiple deep learning models, confirming CNN superiority over classical ML classifiers on malaria images. Siłka et al. (Siłka et al., 2023) introduced an encoder–decoder CNN with U-Net skip connections that achieved 99.68% accuracy, while Kumar et al. Kumar et al. (2024) tested transfer learning models on large multi-parasite datasets, achieving 99.96% accuracy. Delgado-Ortet et al. Delgado-Ortet et al. (2020) proposed a three-step pipeline combining RBC segmentation and classification, though it did not estimate parasite density. Molina et al. Molina et al. (2021) developed a CNN to distinguish parasitized from normal RBCs, focusing on thin smears but excluding leukocytes.

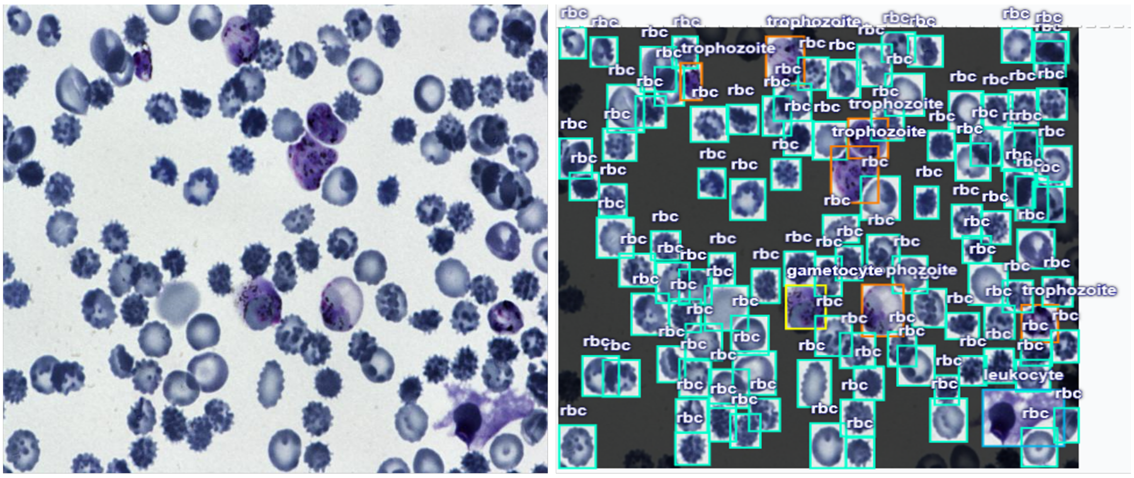
Although many studies report high accuracy, most focus on thin smears or cropped single-cell images, and few address parasite staging or parasitemia estimation. This study bridges that gap by applying Mask R-CNN and YOLOv8 to full-field Giemsa-stained images annotated for both parasite and leukocyte classes. Using data augmentation to improve generalization, we evaluated both models on their ability to detect, segment, quantify, and classify malaria parasites by life stage.

2 Methods

2.1 Dataset and Annotation

This study used a publicly available dataset of 1,328 high-resolution Giemsa-stained thin blood-smear images obtained from the Roboflow repository (Penelitian, 2024). The dataset is fully anonymized and publicly accessible, with no new human data collection involved. Seven object classes were defined for annotation: ring-stage parasite, trophozoite, schizont, gametocyte, “difficult” parasite (ambiguous or partially visible forms), leukocyte (white blood cell), and red blood cell (uninfected). Polygonal masks were created for each cell and parasite using the Roboflow Image Annotator, enabling both bounding-box and pixel-level representation. Annotations were exported in COCO format for Mask R-CNN and YAML format for YOLOv8. The dataset was randomly divided into training (70%), validation (20%), and testing (10%) subsets.

Sample dataset images and annotation examples are illustrated in Figure 1. Figure 1(a) shows representative Giemsa-stained thin-smear fields, while Figure 1(b) presents the corresponding ground-truth masks and bounding boxes used in training. Prior to training, all images were resized and normalized. For YOLOv8, images were resized to 640×640 pixels, whereas for Mask R-CNN (Detectron2 implementation), the shorter side was resized to 800 pixels with the longer side limited to 1333 pixels. Pixel values were normalized to the [0, 1] range and converted to RGB tensors. Basic augmentations, including random flips, rotations, and brightness/contrast adjustments, were applied to enhance model generalization.



(a) (b)

Figure 1: Annotation process for the malaria dataset. (a) Giemsa-stained thin-smear images. (b) Ground-truth masks for erythrocytes, leukocytes, and parasites used in training.

2.2 Model Training

Two deep learning instance segmentation models were trained for this study.

2.2.1 Mask R-CNN

Mask R-CNN was implemented using Detectron2 with a ResNet-50 + FPN backbone pretrained on MS COCO. The model was fine-tuned on the malaria dataset using a multi-task loss function:

$$L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}, \tag{2.1}$$

where L_{cls}, L_{box}, and L_{mask} denote classification, bounding box regression, and mask segmentation losses, respectively. Stochastic Gradient Descent (SGD) with momentum was used, with an initial learning rate of 2.5×10^{-4} and linear warm-up. A batch size of two was used due to Google colab environment constraints. Early stopping was configured after 50,000 iterations without validation loss improvement; however, training continued to the maximum of 100,000 iterations as performance continued improving.

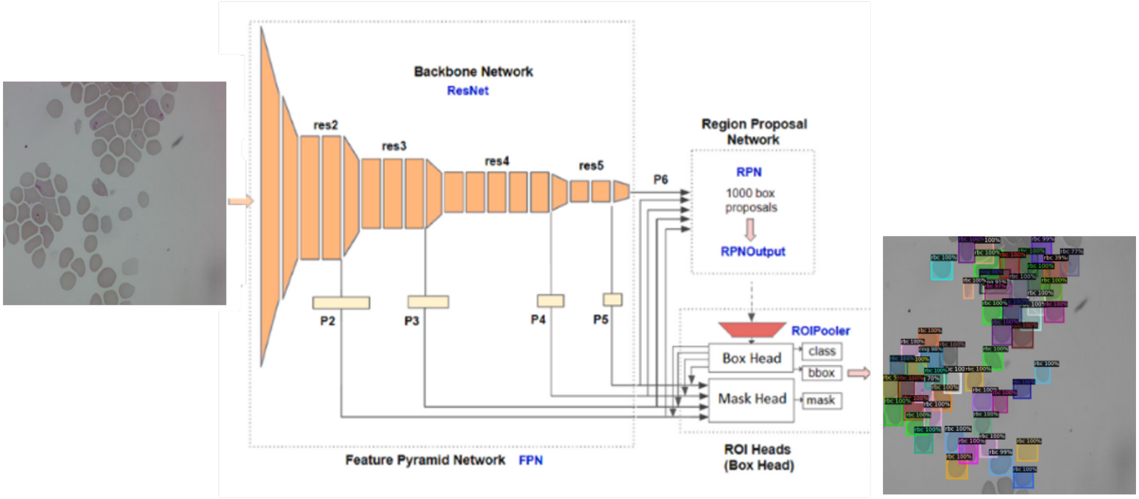


Figure 2: Mask R-CNN architecture showing the ResNet-50 + FPN backbone, region proposal network (RPN), and parallel box and mask prediction branches.

2.2.2 YOLOv8

YOLOv8n-seg, the nano instance segmentation variant of YOLOv8, was used with COCO pretraining. It employs a CSP-Darknet backbone and a unified detection head for simultaneous bounding box, class, and mask prediction. The model was fine-tuned using the AdamW optimizer with an effective learning rate of approximately 9×10^{-4} and a batch size of four. The total loss combined bounding box regression (L_{box}), objectness (L_{obj}), classification (L_{cls}), and mask segmentation (L_{mask}):

$$L_{\text{total}} = L_{\text{box}} + L_{\text{obj}} + L_{\text{cls}} + L_{\text{mask}}. \tag{2.2}$$

YOLOv8 training was performed for 200 epochs, while Mask R-CNN was trained for 100,000 iterations. Both models were trained on an NVIDIA A100 GPU. Mask R-CNN training required approximately 2

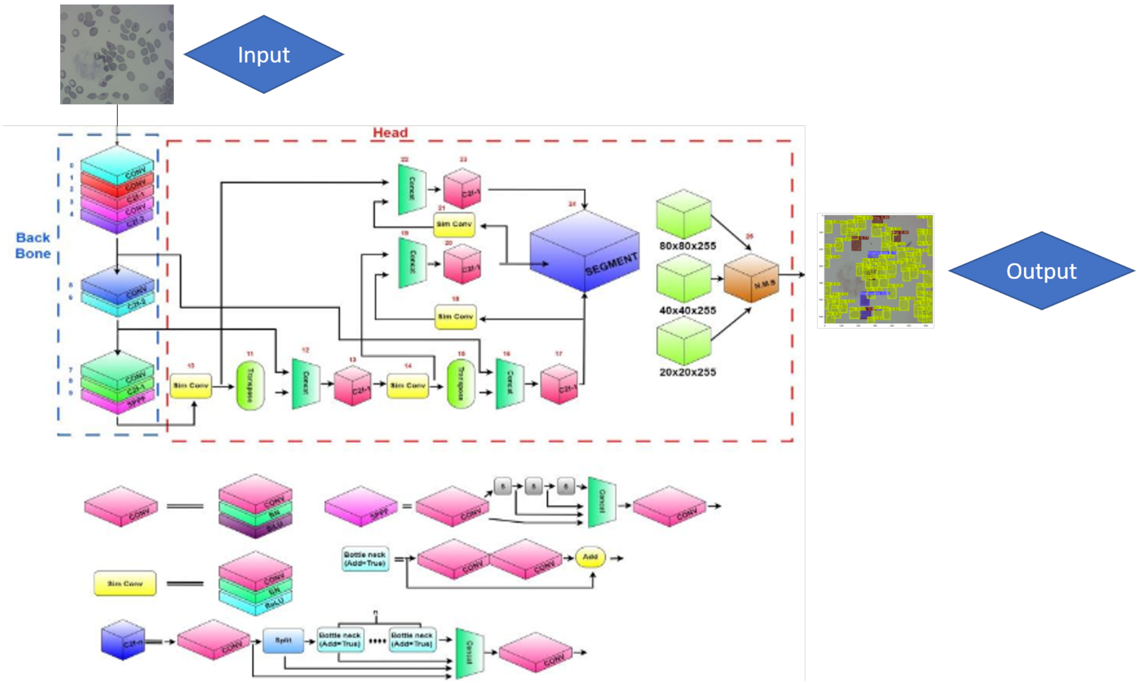


Figure 3: YOLOv8-nano segmentation architecture highlighting the backbone, detection head, and mask decoder.

hours and 45 minutes, whereas YOLOv8 completed in 1 hour and 56 minutes. For both models, the checkpoints yielding the highest validation mask AP were retained for evaluation.
Figure 4 presents the overall methodological framework used in this study.

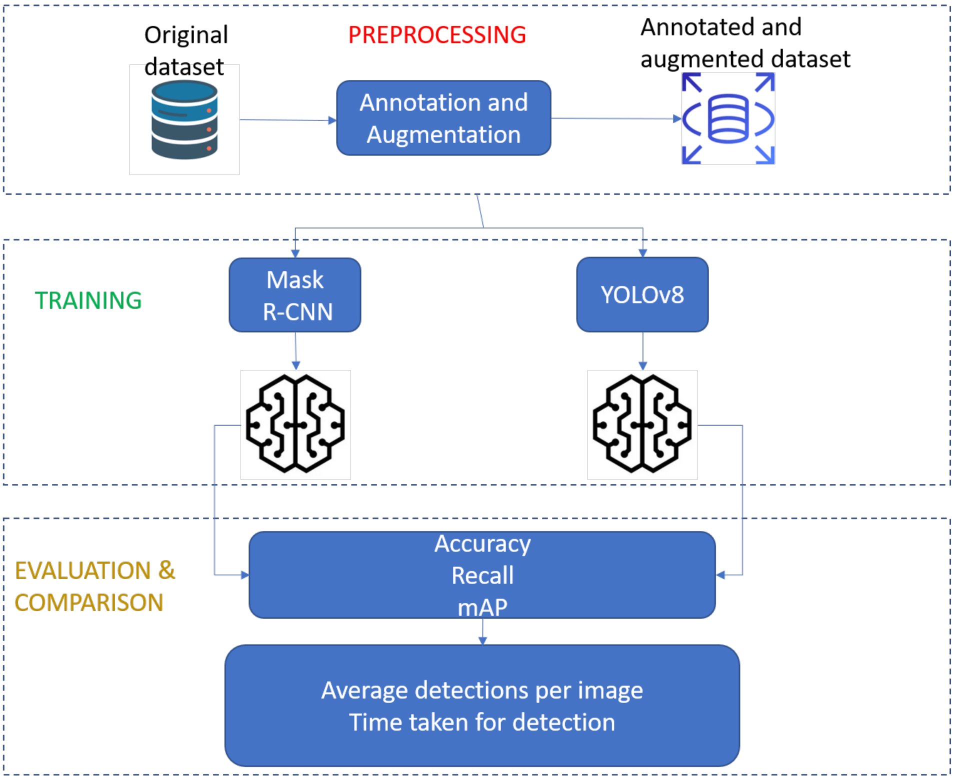


Figure 4: Overall methodological framework of this work.

2.3 Evaluation Metrics

Model performance was evaluated on the test set using COCO-style metrics and classification accuracy. The primary metric was mean average precision (mAP) across Intersection-over-Union (IoU) thresholds. mAP_{50} represents performance at an IoU threshold of 0.5, while $\mathrm{mAP}_{50:95}$ averages performance across thresholds from 0.5 to 0.95 in steps of 0.05:

$$\mathrm{mAP} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{AP}(k) \tag{2.3}$$

For segmentation evaluation, the predicted masks were compared with ground-truth masks using IoU:

$$\text{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{2.4}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

Additional metrics included precision, recall, F1-score, and accuracy:

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\%, \quad \text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\%, \quad \text{F1-score} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{2.5}$$

$$\text{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \times 100\%. \tag{2.6}$$

Because accuracy can be misleading for imbalanced classes, per-class mAP and recall were computed to capture minority-class behavior (e.g., schizonts and gametocytes). Confidence thresholds were varied from 0.1 to 0.9 to generate precision–recall curves and analyze threshold sensitivity. Inference time (ms per image) was measured to assess real-time feasibility.

To ensure statistical reliability, all key metrics were bootstrapped 1,000 times on the test set to compute 95% confidence intervals. For qualitative comparison, segmentation masks generated by both models were overlaid on test images and visually inspected to assess boundary precision and error patterns. Representative mask outputs were visually inspected and qualitatively compared to evaluate boundary precision and segmentation behavior.

3 Results And Discussion

3.1 Mask R-CNN Performance

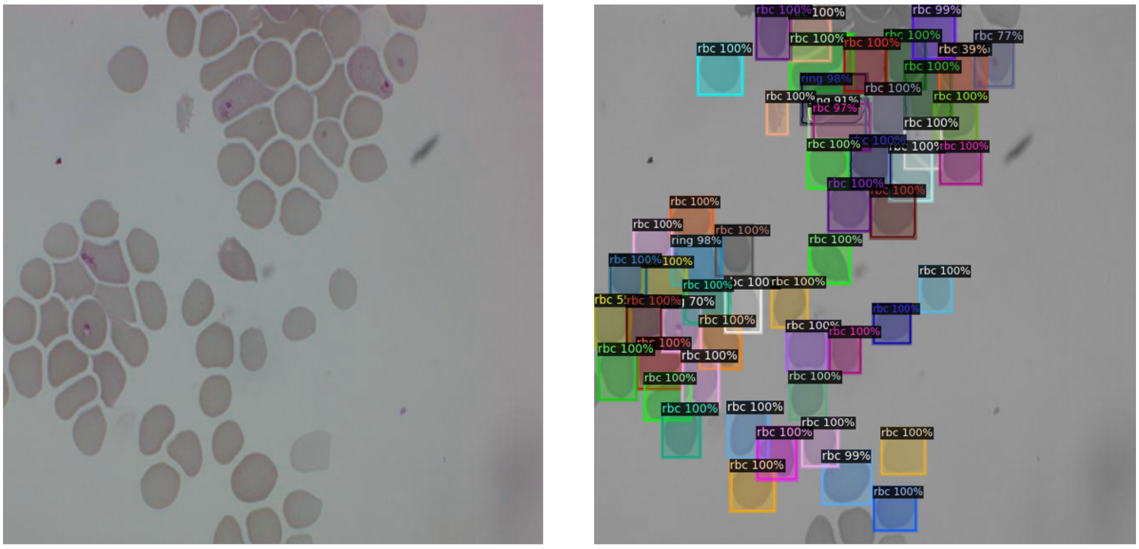
For Mask R-CNN, the overall mask AP$_{50}$ was approximately 0.511 (51.1%), with bounding box AP around 0.398 and segmentation AP about 0.396. Early in training, the model exhibited very high precision (≈ 0.997) but extremely low recall (≈ 0.12), indicating that while it produced few false positives, it missed many parasites. Table 1 summarizes the model's performance across COCO-style metrics. Although the AP$_{50}$ values appear reasonable, overall average precision, especially for small objects, remained low. This suggests that while Mask R-CNN was generally accurate on the objects it detected, it struggled with smaller or subtler parasite stages.

Table 1: Performance of Mask R-CNN using COCO-style evaluation metrics. Values are reported for both bounding box and segmentation tasks.

Task Type	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AP$_{S}$**	**AP$_{M}$**
Bounding Box	39.822	51.073	46.182	70.00	39.822
Segmentation/Mask	39.673	51.073	46.194	65.00	39.673

Qualitatively, Mask R-CNN produced accurate binary masks for red blood cells (RBCs) and some visible parasites. Figure 5 shows an example test set image (a) and the model's predictions (b). While many RBCs (green boxes) were correctly identified and some parasites were segmented by life stage, smaller parasites were frequently missed.

As illustrated in Figure 5, the model produced high-quality masks for detected objects but failed to capture a substantial portion of the parasite population. It performed best on larger and more common targets, achieving AP$_{50}$ ≈ 0.717 for RBCs and ≈ 0.502 for ring-stage parasites. Performance was notably lower for rare classes such as schizonts (AP$_{50}$ ≈ 0.270). Mask R-CNN showed precise localization (high precision) but poor sensitivity (low recall). It was also slower than YOLOv8, averaging 33.3 ms for inference and 58.4 ms for post-processing per image.



(a)

(b)

Figure 5: Mask R-CNN segmentation results: (a) example test set image; (b) predicted output showing detected cells and parasites.

3.2 YOLOv8 Performance

On the same test set, YOLOv8 achieved a bounding box mAP_{50} of approximately 0.648 and a mask mAP_{50} of around 0.624, both higher than Mask R-CNN's 0.511. The model reached bounding box precision of roughly 0.557 and recall of 0.724, with mask precision of 0.570 and recall of 0.660. Table 2 shows that YOLOv8 achieved stronger performance across both bounding box and mask tasks.

Table 2: Performance of YOLOv8 using mAP evaluation metrics. All values are scaled by 100 to match COCO format.

Task Type	**mAP**$_{50}$	**mAP**$_{50:95}$
Bounding Box	63.495	54.152
Segmentation/Mask	62.037	40.254

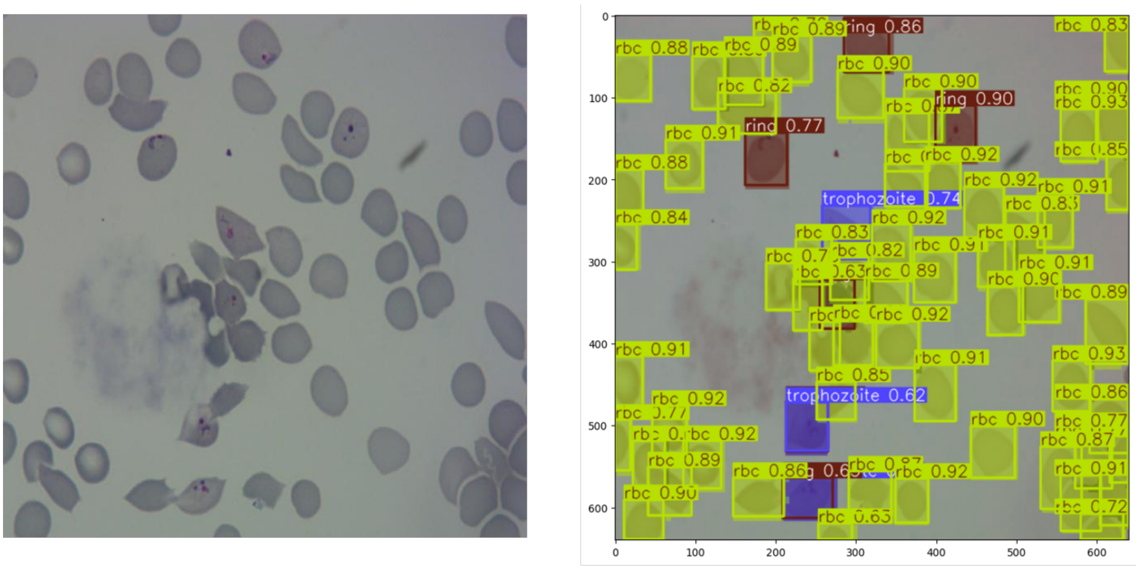
YOLOv8 successfully detected most parasite instances, with higher recall but more false positives compared to Mask R-CNN. This balance, higher recall at the cost of some precision, aligns with requirements for screening applications where minimizing missed infections is critical.

Class-wise, YOLOv8 performed best on abundant categories. It detected nearly all RBCs ($\text{AP}_{50} \approx 0.992$), most ring-stage parasites ($\text{AP}_{50} \approx 0.732$), and trophozoites ($\text{AP}_{50} \approx 0.845$). Performance on rare classes was lower: schizonts scored around 0.315 and gametocytes 0.369. YOLOv8's AP for rings (0.732) was higher than Mask R-CNN's (0.502), reflecting greater sensitivity for this common stage.

Figure 6 shows a representative test set image (a) with YOLOv8 predictions (b). While the model detected many parasites accurately, smaller ones were sometimes missed or only partially localized. Importantly, YOLOv8 processed each image in approximately 4.9 ms for inference and 4.2 ms for post-processing (Table 3), enabling near real-time performance on GPU hardware.

Table 3: Inference and post-processing times per image for Mask R-CNN and YOLOv8.

Model	**Inference**	**Processing**	**Total**
Mask R-CNN	33.3 ms	58.4 ms	93.2 ms
YOLOv8	4.9 ms	4.2 ms	9.1 ms



(a)

(b)

Figure 6: YOLOv8 detection and segmentation output: (a) example test set image; (b) predicted bounding boxes and masks.

3.3 Bootstrap Confidence Interval Analysis

To validate performance differences statistically, bootstrap confidence interval analysis (n=1000 iterations) was conducted on per-class detection confidence scores. Table 4 presents mean confidence scores with 95% confidence intervals.

The bootstrap analysis highlights a precision–recall trade-off. Mask R-CNN showed higher confidence on detected classes but conservative detection behavior, failing to detect trophozoites entirely. YOLOv8 had lower average confidence but broader coverage, including trophozoites with mean confidence of 0.620 [0.597, 0.643]. Confidence interval widths indicated stable detection for abundant classes (RBCs, rings) and higher uncertainty for rare classes (schizonts, gametocytes).

Table 4: Bootstrap confidence interval analysis comparing YOLOv8 and Mask R-CNN detection confidence scores across parasite classes.

Class	**YOLOv8 Mean (95% CI)**	**Mask R-CNN Mean (95% CI)**	**Difference**
RBC	0.809 [0.807, 0.812]	0.993 [0.992, 0.994]	-0.184
Ring	0.675 [0.626, 0.722]	0.859 [0.822, 0.893]	-0.184
Trophozoite	0.620 [0.597, 0.643]	*Not detected*	N/A
Difficult	0.352 [0.331, 0.376]	0.749 [0.690, 0.805]	-0.397
Schizont	0.485 [0.405, 0.566]	0.806 [0.753, 0.853]	-0.321
Gametocyte	0.365 [0.307, 0.421]	0.722 [0.628, 0.813]	-0.357
Leukocyte	0.791 [0.638, 0.931]	0.934 [0.886, 0.973]	-0.143

Most class differences were supported by non-overlapping confidence intervals, except for leukocytes where minimal overlap suggested borderline significance. These findings confirm that observed performance differences reflect systematic model behavior rather than random sampling.

3.4 Comparative Evaluation

YOLOv8 achieved higher mAP$_{50}$ values and substantially greater recall compared to Mask R-CNN. For example, YOLOv8 reached a mask AP of 0.624 versus 0.511 for Mask R-CNN. The bootstrap analysis clarified this trade-off: Mask R-CNN attained very high accuracy on detected objects (99.7% vs. 96.7% for YOLOv8) but detected far fewer parasite instances. Figure 7 shows accuracy-over-training curves, with Mask R-CNN (blue) reaching near-perfect accuracy on limited detections, while YOLOv8 (orange) improved more gradually as it captured a broader range of parasite instances.

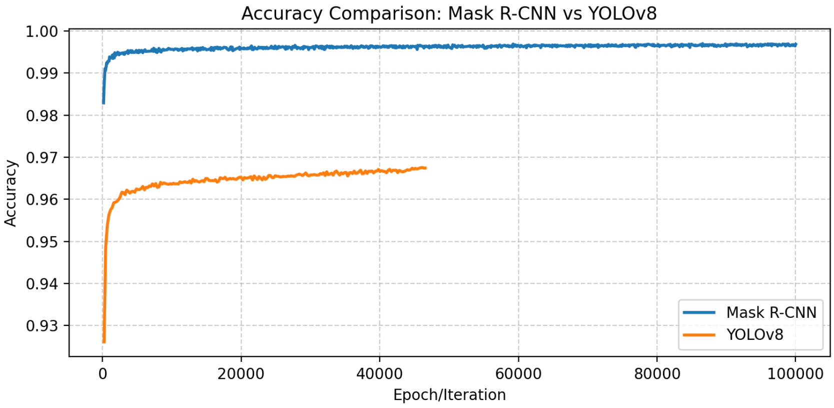


Figure 7: Comparison of Mask R-CNN and YOLOv8 accuracy curves during training.

These differences have potential clinical implications. Mask R-CNN's high precision, low recall behavior may be better suited for confirmatory diagnosis where false positives must be minimized, whereas YOLOv8's high recall behavior aligns better with screening scenarios where sensitivity is prioritized.

3.5 Segmentation Output Visualization and Interpretation

To complement the quantitative metrics, representative segmentation results from both models were qualitatively examined. Each test image was evaluated alongside its predicted masks and class overlays to assess spatial accuracy and coverage. Mask R-CNN produced spatially precise masks on detected objects, particularly red blood cells and ring-stage parasites, confirming accurate boundary delineation where detections occurred. However, many smaller or low-contrast parasites were missed, consistent with its low recall. YOLOv8 generated broader coverage with smoother but occasionally less refined masks, successfully capturing trophozoites and difficult parasite instances that Mask R-CNN omitted. Ground-truth masks were used during training and quantitative evaluation but are not displayed here for brevity; however, visual inspection confirmed that both models learned meaningful cell-level boundaries consistent with the annotated data.

These qualitative results explain the apparent discrepancy between numerical mAP values (39–65%) and statements of precision. Mask R-CNN exhibited high boundary fidelity for detected parasites, yielding accurate morphological segmentation, but its conservative region proposals limited overall recall. YOLOv8, conversely, achieved higher coverage at the expense of marginally reduced boundary sharpness. Thus, “precision” in this context refers to boundary accuracy on positive detections rather than aggregate mAP.

3.6 Clinical and Analytical Relevance of Segmentation

Pixel-level segmentation is clinically relevant because it enables parasite quantification and morphological analysis beyond simple presence detection. Accurate mask delineation allows estimation of infected-cell area fractions and stage-specific parasitemia metrics that are critical for assessing infection severity and treatment response. Segmentation also separates overlapping cells, distinguishing parasites from leukocytes and debris, which reduces counting bias in dense smears. Therefore, although overall mAP values were moderate, the generated masks provide the necessary structural information for downstream quantification tasks and confirm the feasibility of automated parasitemia estimation.

3.7 Parasite Stage Classification and Quantification

Both models assigned life-stage labels to detected parasites in addition to classifying red blood cells and leukocytes. YOLOv8 achieved higher AP on common stages such as rings and trophozoites but showed reduced performance on rarer stages like schizonts and gametocytes. Mask R-CNN's stage-wise AP values were generally lower; for example, it did not detect any trophozoites (AP = 0), although it identified some ring-stage parasites. This suggests that earlier-stage infections were more likely to be recognized, while later stages were often missed.

The dataset exhibited substantial class imbalance, with RBCs dominant (15,570 instances) and parasite stages much less frequent (e.g., 112 rings, 285 trophozoites, 38 schizonts, 31 gametocytes). This imbalance likely contributed to weaker performance on rare classes. Both models tended to predict majority classes, particularly RBCs, reflecting the data distribution.

Figures 10 to 14 provide additional evaluation perspectives.

Figure 10 shows precision–recall curves, where YOLOv8 achieved consistently higher recall and more convex curves, particularly for rings and trophozoites.

Figure 11 shows precision as a function of confidence threshold. YOLOv8 demonstrated a monotonic increase in precision, while Mask R-CNN's curves were less stable across thresholds.

F1-score curves (Figure 12) illustrate balanced performance for YOLOv8 across thresholds, compared with greater variability for Mask R-CNN.

Recall versus confidence (Figure 13) further highlights YOLOv8's ability to retain sensitivity across thresholds, while Mask R-CNN dropped sharply at low thresholds.

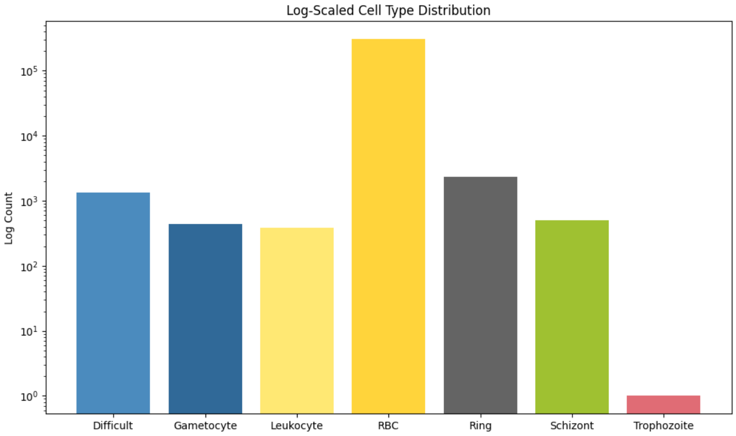


Figure 8: Class-wise detection performance of Mask R-CNN across blood cell and parasite stages.

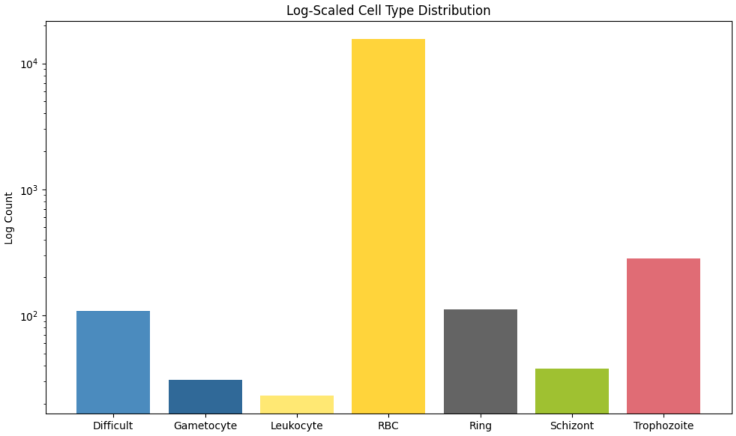
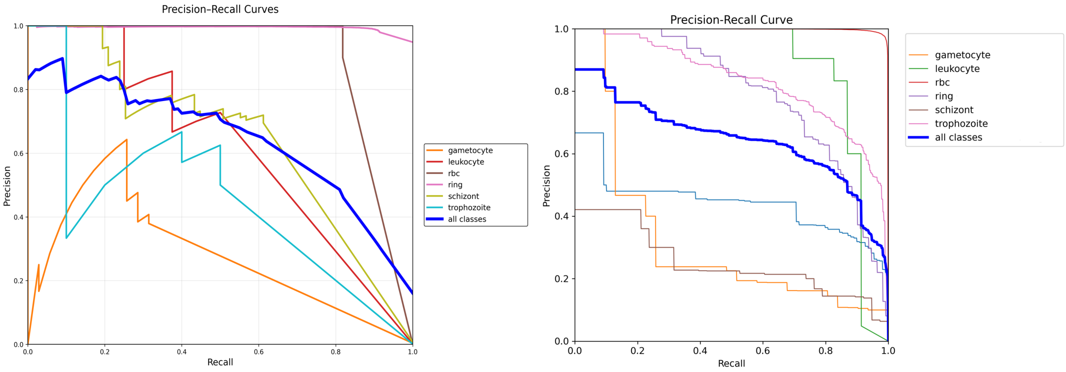
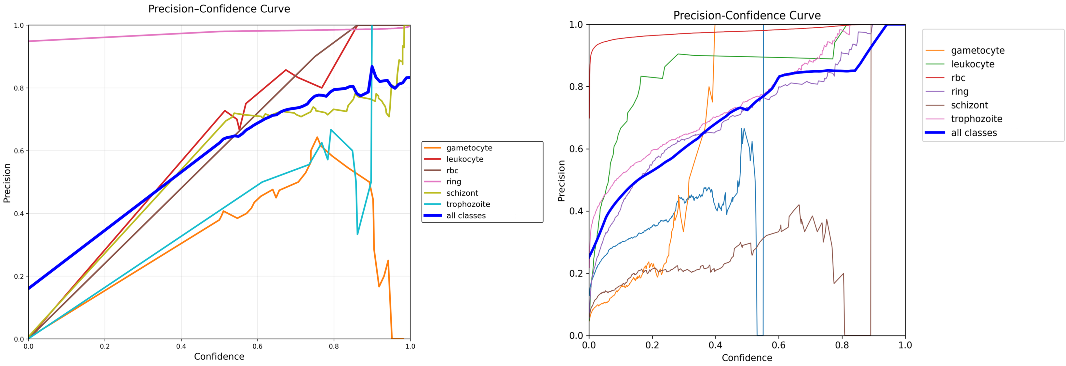


Figure 9: Class-wise detection performance of YOLOv8 across blood cell and parasite stages.



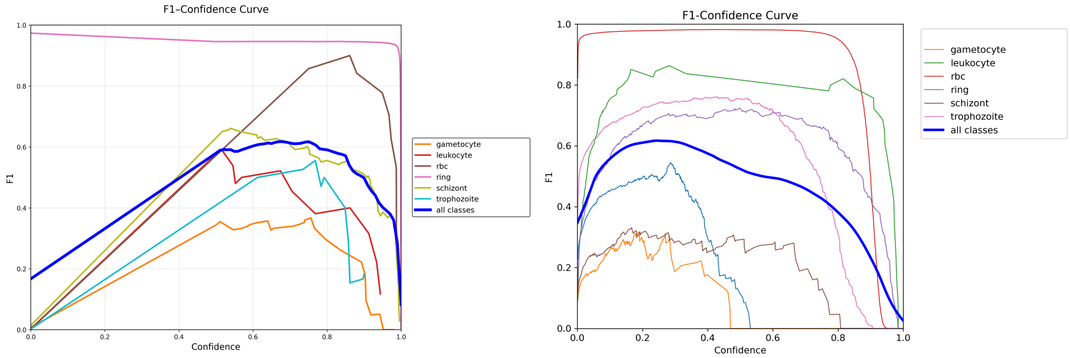
(a) (b)

Figure 10: Precision–Recall curves by class, with overall mean in bold. YOLOv8 shows higher recall and more stable curves, especially for ring and trophozoite classes.



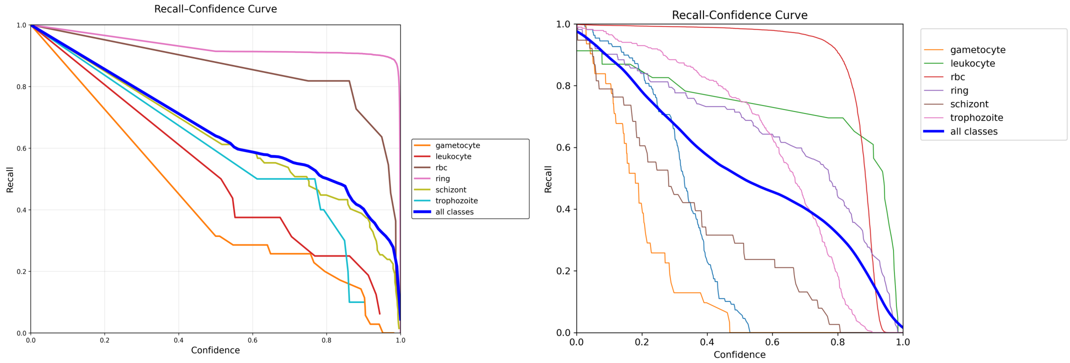
(a) (b)

Figure 11: Precision as a function of confidence threshold. YOLOv8 shows consistent calibration, while Mask R-CNN varies across thresholds.



(a) (b)

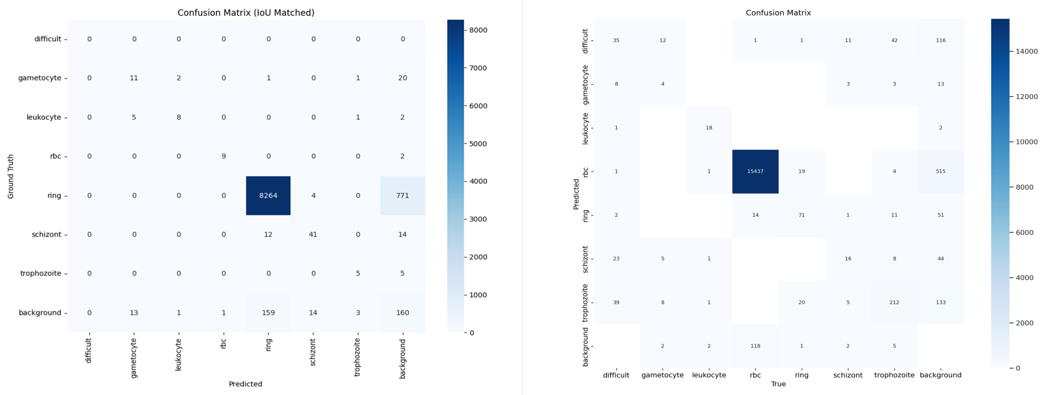
Figure 12: F1-score across confidence thresholds. YOLOv8 maintains balanced performance; Mask R-CNN shows greater variability.



(a) (b)

Figure 13: Recall as a function of confidence threshold. YOLOv8 sustains higher recall across thresholds, which is advantageous in screening contexts.

Finally, confusion matrices (Figure 14) show YOLOv8 detecting a broader range of parasite stages with fewer misclassifications. Mask R-CNN performed reliably on RBCs and rings but struggled on less frequent classes such as gametocytes.



Ground Truth \ Predicted	difficult	gametocyte	leukocyte	rbc	ring	schizont	trophozoite	background
difficult	0	0	0	0	0	0	0	0
gametocyte	0	11	2	0	1	0	1	20
leukocyte	0	5	8	0	0	0	1	2
rbc	0	0	0	9	0	0	0	2
ring	0	0	0	0	8264	4	0	771
schizont	0	0	0	0	12	41	0	14
trophozoite	0	0	0	0	0	0	5	5
background	0	13	1	1	159	14	3	160

(a)

Predicted \ True	difficult	gametocyte	leukocyte	rbc	ring	schizont	trophozoite	background
difficult	35	12		1	1	11	42	116
gametocyte	8	4				3	3	13
leukocyte	1		18					2
rbc	1		1	15437	19		4	515
ring	2			14	71	1	11	51
schizont	23	5	1			16	8	44
trophozoite	39	8	1		20	5	212	133
background		2	2	118	1	2	5	

(b)

Figure 14: Confusion matrices showing true vs. predicted class distributions. YOLOv8 detects a broader range of parasite stages with fewer misclassifications.

3.8 Summary of Statistical and Visual Evidence

Together, these results show that YOLOv8 achieved higher sensitivity, more consistent class-wise performance, and greater stability across thresholds compared to Mask R-CNN. Bootstrap analysis confirmed that these differences were statistically supported and reflected distinct detection strategies: Mask R-CNN's conservative, high-precision behavior versus YOLOv8's broader, recall-oriented approach. Visual inspection of segmentation overlays corroborated these quantitative trends, demonstrating that both models provide complementary strengths, YOLOv8 for initial screening and Mask R-CNN for confirmatory diagnosis.

4 Conclusion

This study compared two deep learning instance segmentation models, Mask R-CNN and YOLOv8, for automated detection, segmentation, and life-stage classification of *Plasmodium* parasites in Giemsa-stained blood smear images. Using a dataset of 1,328 annotated thin-smear images, both models were evaluated using standard COCO-style instance segmentation metrics. YOLOv8 achieved higher mean average precision and substantially greater recall than Mask R-CNN, confirming its stronger sensitivity and suitability for broad screening applications. Mask R-CNN produced spatially precise masks on detected cells, supporting accurate morphological delineation but with limited recall, particularly for rare parasite stages.

Beyond recall, YOLOv8 exhibited smoother precision–recall characteristics and more stable confidence calibration across thresholds, indicating robust performance under variable decision criteria. Its lightweight architecture enabled near real-time inference (approximately 9 ms per image), making it particularly suitable for deployment in point-of-care or high-throughput diagnostic settings. Mask R-CNN, although slower, generated high-fidelity segmentation masks that may be advantageous for downstream tasks such as parasite morphology analysis, stage quantification, and infection-density estimation.

The results demonstrate that instance segmentation offers tangible clinical value beyond object detection by enabling pixel-level quantification of parasitemia and stage-specific morphology, both important for disease monitoring and treatment response assessment. The complementary behaviors of YOLOv8 and Mask R-CNN high recall versus fine-grained precision, suggest their combined use could provide a balanced diagnostic workflow: YOLOv8 for initial screening and Mask R-CNN for confirmatory analysis.

This study was limited by dataset imbalance, particularly underrepresentation of rare parasite stages, and by reliance on a single public dataset, which may constrain generalization across staining conditions and imaging hardware. Future work should include domain adaptation, larger multi-source datasets, and evaluation of hybrid architectures integrating both detection speed and mask precision.

The findings confirm that deep learning instance segmentation models can enhance malaria diagnostics by uniting detection accuracy, morphological segmentation, and computational efficiency. Their integration into automated microscopy workflows has strong potential to support scalable, cost-effective malaria screening and improve diagnostic capacity in resource-limited regions.

5 Declarations

5.1 Conflict of Interest

The authors declare that they have no commercial or financial relationships that could be construed as potential conflicts of interest.

5.5 Ethics Approval and Consent to Participate

This study did not involve new data collection from human participants. It used pre-existing, anonymized Giemsa-stained blood smear images made publicly available by the Roboflow repository (Penelitian, 2024). In accordance with institutional and national research guidelines, studies using publicly available and fully anonymized secondary datasets do not require formal ethics approval. Because no identifiable information was present, informed consent to participate was not required. The study was conducted in accordance with the principles of the Declaration of Helsinki.

5.6 Consent for Publication

This work does not contain any individual person's data in any form (including images, videos, or personal details); therefore, consent for publication was not applicable.

5.7 Availability of Data and Materials

The dataset analyzed in this study is publicly available through the Roboflow repository: `https://universe.roboflow.com/penelitian-hnhst/malaria-egy2k`. The code used for model training and evaluation is available from the corresponding author upon reasonable request.

References

Britton, S., Cheng, Q., and McCarthy, J. S. (2016). Novel molecular diagnostic tools for malaria elimination: a review of options from the point of view of high-throughput and applicability in resource limited settings. *Malaria journal*, 15(1):88.

Bronzan, R. N., McMorrow, M. L., and Patrick Kachur, S. (2008). Diagnosis of malaria: challenges for clinicians in endemic and non-endemic regions. *Molecular diagnosis & therapy*, 12(5):299–306.

Brostow, G. J., Fauqueur, J., and Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern recognition letters*, 30(2):88–97.

Chibuta, S. and Acar, A. C. (2020). Real-time malaria parasite screening in thick blood smears for low-resource setting. *Journal of digital imaging*, 33(3):763–775.

Davidson, M. S., Andradi-Brown, C., Yahiya, S., Chmielewski, J., O'Donnell, A. J., Gurung, P., Jeninga, M. D., Prommana, P., Andrew, D. W., Petter, M., et al. (2021). Automated detection and staging of malaria parasites from cytological smears using convolutional neural networks. *Biological imaging*, 1:e2.

Delgado-Ortet, M., Molina, A., Alférez, S., Rodellar, J., and Merino, A. (2020). A deep learning approach for segmentation of red blood cell images and malaria detection. *Entropy*, 22(6):657.

Fuhad, K. F., Tuba, J. F., Sarker, M. R. A., Momen, S., Mohammed, N., and Rahman, T. (2020). Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. *Diagnostics*, 10(5):329.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Koita, O. A., Doumbo, O. K., Ouattara, A., Tall, L. K., Konaré, A., Diakité, M., Diallo, M., Sagara, I., Masinde, G. L., Doumbo, S. N., et al. (2012). False-negative rapid diagnostic tests for malaria and deletion of the histidine-rich repeat region of the hrp2 gene. *The American journal of tropical medicine and hygiene*, 86(2):194.

Kumar, Y., Garg, P., Moudgil, M. R., Singh, R., Woźniak, M., Shafi, J., and Ijaz, M. F. (2024). Enhancing parasitic organism detection in microscopy images through deep learning and fine-tuned optimizer. *Scientific Reports*, 14(1):5753.

Maqsood, A., Farid, M. S., Khan, M. H., and Grzegorzek, M. (2021). Deep malaria parasite detection in thin blood smear microscopic images. *Applied Sciences*, 11(5):2284.

Molina, A., Rodellar, J., Boldú, L., Acevedo, A., Alférez, S., and Merino, A. (2021). Automatic identification of malaria and other red blood cell inclusions using convolutional neural networks. *Computers in biology and medicine*, 136:104680.

Moody, A. (2002). Rapid diagnostic tests for malaria parasites. *Clinical microbiology reviews*, 15(1):66–78.

Mujahid, M., Rustam, F., Shafique, R., Montero, E. C., Alvarado, E. S., de la Torre Diez, I., and Ashraf, I. (2024). Efficient deep learning-based approach for malaria detection using red blood cell smears. *Scientific Reports*, 14(1):13249.

Narayanan, B. N., Ali, R., and Hardie, R. C. (2019). Performance analysis of machine learning and deep learning architectures for malaria detection on cell images. In *Applications of Machine Learning*, volume 11139, pages 240–247. SPIE.

Penelitian (2024). Malaria dataset. `https://universe.roboflow.com/penelitian-hnhst/malaria-egy2k`. visited on 2025-08-12.

Quinn, J. A., Andama, A., Munabi, I., and Kiwanuka, F. N. (2018). 6 automated blood smear analysis for mobile. *Mobile point-of-care monitors and diagnostic device design*, page 115.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568.

Razin, W. R. W. M., Gunawan, T. S., Kartiwi, M., and Yusoff, N. M. (2022). Malaria parasite detection and classification using cnn and yolov5 architectures. In *2022 IEEE 8th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, pages 277–281. IEEE.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Siłka, W., Wieczorek, M., Siłka, J., and Woźniak, M. (2023). Malaria detection using advanced deep learning architecture. *Sensors*, 23(3):1501.

Var, E. and Tek, F. B. (2018). Malaria parasite detection with deep transfer learning. In *2018 3rd International conference on computer science and engineering (UBMK)*, pages 298–302. IEEE.

World Health Organization (2024). *World Malaria Report 2024*. World Health Organization. Accessed: 2025-04-30.

Yang, F., Quizon, N., Yu, H., Silamut, K., Maude, R. J., Jaeger, S., and Antani, S. (2020). Cascading yolo: automated malaria parasite detection for plasmodium vivax in thin blood smears. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 404–410. SPIE.