A COCO-Formatted Instance-Level Dataset for Plasmodium Falciparum Detection in Giemsa-Stained Blood Smears

Frauke **Wilm** [1,2], Luis Carlos Rivera Monroy [1, 2], Mathias Öttl [1, 2], Lukas Mürdter [1], Leonid Mill [1, 2], Andreas Maier [2]

1 MIRA Vision Microscopy GmbH, 73037 Göppingen, Germany
2 Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany

Abstract

Accurate detection of *Plasmodium falciparum* in Giemsa-stained blood smears is an essential component of reliable malaria diagnosis, especially in developing countries. Deep learning-based object detection methods have demonstrated strong potential for automated Malaria diagnosis, but their adoption is limited by the scarcity of datasets with detailed instance-level annotations. In this work, we present an enhanced version of the publicly available NIH malaria dataset, with detailed bounding box annotations in COCO format to support object detection training. We validated the revised annotations by training a Faster R-CNN model to detect infected and non-infected red blood cells, as well as white blood cells. Cross-validation on the original dataset yielded F1 scores of up to 0.88 for infected cell detection. These results underscore the importance of annotation volume and consistency, and demonstrate that automated annotation refinement combined with targeted manual correction can produce training data of sufficient quality for robust detection performance. The updated annotations set is publicly available via Zenodo: `https://doi.org/10.5281/zenodo.17514694`.

1. Background

Malaria is a tropical disease caused by protozoan parasites of the genus *Plasmodium*, which infect red blood cells and are primarily transmitted through the bites of female Anopheles mosquitoes. In humans, the disease is mainly associated with four species: *P. falciparum*, *P. vivax*, *P. malariae*, and *P. ovale*. In recent years, malaria has also been increasingly transmitted by a fifth species, i. e., *P. knowlesi*. Among these five species, *P. falciparum* and *P. vivax* are the most prevalent, and *P. falciparum* is responsible for the majority of malaria-related deaths (World Health Organization, 2024).

Most malaria infections are reported in tropical and subtropical regions, affecting populations in low-income countries with limited access to healthcare. Although modern treatments can effectively cure malaria, early diagnosis remains critical and delays in detection are a major contributing factor to malaria-related mortality (Sultani et al.,

2022). For parasitological diagnosis of malaria, microscopic examination of thick and thin blood smear images is routinely performed. In addition to identifying the *Plasmodium* species, light microscopy allows parasite quantification and monitoring therapy response. Therefore, it is often preferred over molecular testing (World Health Organization, 2024). Nonetheless, the parasitological assessment of blood smear images requires a high level of expertise, and trained personnel might be scarce in low-resource countries or rural areas (Poostchi et al., 2018).

Recently, machine learning-based approaches for analyzing digitized blood smear images have demonstrated promising results in parasitemia quantification (Poostchi et al., 2018). However, these methods typically rely on large, well-annotated datasets for effective training, making publicly available resources particularly valuable. Most existing work focuses on classifying individual cell patches as infected or non-infected (Kassim et al., 2020), which requires the prior extraction of single-cell crops. This step

can be challenging in densely populated blood smear images and limits the applicability of such approaches in real-world diagnostic workflows, where direct localization and accurate quantification of infected cells are essential. In contrast to patch-based classification approaches, object detection architectures require datasets with detailed instance-level annotations, typically in the form of labeled bounding boxes. However, acquiring such detailed annotations is labor-intensive and time-consuming, which limits their availability. The cro:nihNIH dataset, comprising 965 images, is one of the largest publicly available resources for *P. falciparum* detection. However, only 165 of these images include detailed polygon-based annotations, while the remaining 800 are limited to point annotations marking cell centers. This sparsity limits their suitability for training deep learning-based object detection models, which typically require bounding box annotations.

In this work, we present a revised version of the cro:nihNIH dataset with enhanced annotations. Using the Cellpose framework (Pachitariu and Stringer, 2022) and manual label correction, we converted the original point annotations into bounding box labels, which are better suited for object detection. To validate the quality of the revised dataset, we trained a Faster R-CNN (Ren et al., 2015) for parasite detection, achieving an F1 score of up to 0.88 for infected cell identification. The updated annotation set is publicly available via Zenodo: `https://doi.org/10.5281/zenodo.17514694`.

2. Methods

For our experiments, we generated new bounding box annotations for the cro:nihNIH dataset, which contains Giemsa-stained, thin blood smear images of *P. falciparum*. We conducted a technical validation of these annotations by training a deep learning-based object detector to identify three cell types: non-infected red blood cells, infected red blood cells, and white blood cells.

2.1 Data Details

The cro:nihNIH dataset (Kassim et al., 2020) is a thin-smear malaria image dataset acquired at Chittagong Medical College Hospital in Bangladesh and published by the National Library of Medicine, cro:nihNational Institutes of Health (NIH), Bethesda, MD, USA. It comprises Giemsa-stained, thin blood smear images from 193 patients (148 infected and 45 uninfected), with five images per patient. Each image was captured using a microscope-mounted smartphone camera at a resolution of $5\,312 \times 2\,988$ (width $\times$ height) pixels. Annotations cover three classes: non-infected red blood cells, infected red blood cells, and white blood cells. Of the 965 total images, 165 include detailed polygon-based

annotations, while the remaining 800 provide only point annotations marking cell centers. Table 1 summarizes these subsets, hereafter referred to as NIH_{polys} and NIH_{points}, respectively. Figures 1a and 1b show example regions of interest with contour and point annotations, corresponding to the NIH_{polys} and NIH_{points} subsets.

Table 1: Overview of the NIH dataset subsets. NIH_{polys} includes detailed polygon-based labels, whereas NIH_{points} was annotated with point markers indicating cell centers. $MIRA_{boxes}$ comprises revised labels for the NIH_{points} dataset with detailed bounding box annotations.

	NIH_{polys}	NIH_{points}	$MIRA_{boxes}$
patients	33	160	160
no. of images	165	800	800
annotations	contours	points	boxes
no. of annotations			
non-infected	33 071	155 640	155 201
infected	1 142	6 810	6 805
white blood cell	51	220	220
ambiguous	-	-	19 592

2.2 Annotation Revision

To enable the use of the NIH dataset for training object detection models, we converted the point annotations into detailed bounding-box annotations. For this, we first detected cell instances using Cellpose 2 (Pachitariu and Stringer, 2022), an open-source framework designed for robust, generalizable segmentation. Trained with a diverse dataset of more than $70\,000$ cells, Cellpose offers strong performance across a wide range of cell types and imaging modalities, making it well suited for segmenting Giemsa-stained blood smear images.

Following cell instance segmentation, we assigned labels to detected cells by overlaying the original point annotations. If a point annotation fell within a predicted bounding box, that box was assigned the corresponding cell class. However, Cellpose occasionally detected cells, which were not annotated in the original dataset. These were often partially visible cells at the edge of the field of view. In the updated annotation set, these detections were labeled as *ambiguous*. Figure 2 shows an example with ambiguous cells at the border of the field of view. Overall, the updated annotations comprise $19\,592$ ambiguous cells, which makes up around $10\,\%$ of the original NIH_{points} subset.

Due to its reliance on an average cell size, Cellpose sometimes fragmented larger cells and particularly white blood cells into multiple instances. To address this, we manually reviewed and merged these fragmented detections. Additionally, Cellpose occasionally misclassified artifacts or blood platelets as cells. These false positives were also

(a) Sample from NIH$_{polys}$ (b) Sample from NIH$_{points}$ (c) Bounding box annotations
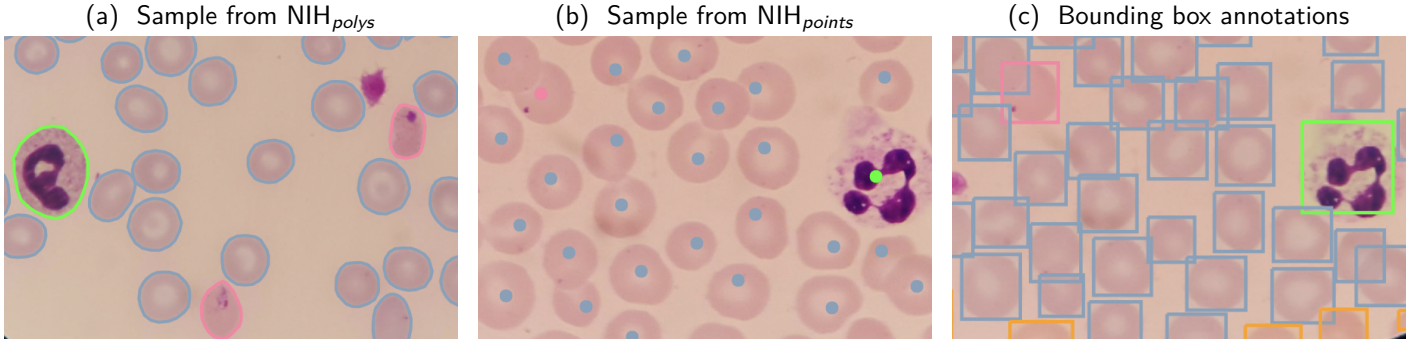

Figure 1: Different annotation types provided by the NIH dataset. (a): contour annotations, (b): point-only annotations, (c) bounding box annotations created with Cellpose (Pachitariu and Stringer, 2022). Blue: non-infected red blood cells, pink: infected cells, green: white blood cells, orange: ambiguous cells.

removed during manual post-processing. Figure 1c shows a representative region of interest after bounding box detection, with ambiguous cells highlighted in orange, and the last column of Table 1 summarizes the number of cell instances after this annotation revision.
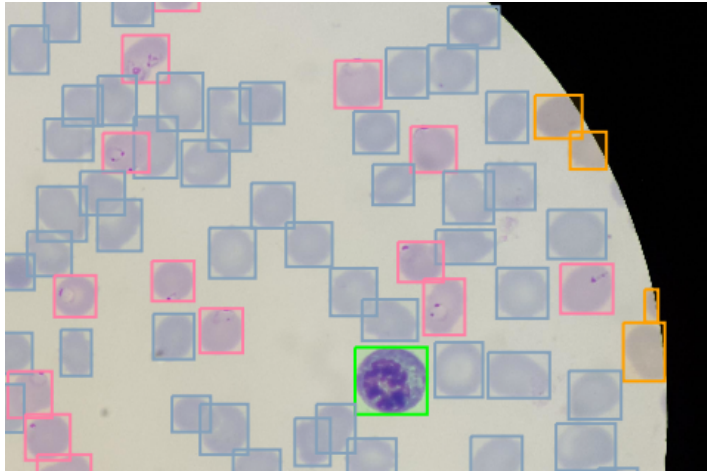

Figure 2: During label cleaning, non-annotated cells at the border of the field of view were labeled as *ambiguous* (orange). Blue: non-infected red blood cells, pink: infected cells, green: white blood cells.

3. Technical Validation

To validate the revised annotations, we trained a Faster R-CNN model (Ren et al., 2015) to detect three cell classes: non-infected red blood cells, infected red blood cells, and white blood cells. We conducted cross-validation experiments by training the model on either the NIH$_{polys}$ or the revised MIRA$_{boxes}$ subset and evaluating its detection performance on the other, respectively.

3.1 Implementation Details

We employed a Faster R-CNN model (Ren et al., 2015) with a ResNet34 (He et al., 2016) backbone, pretrained on ImageNet (Russakovsky et al., 2015). The datasets were split into 70% for training and 30% for validation. On the NIH$_{polys}$ dataset, the model was trained for $1\,000$ epochs using a cosine annealing learning rate schedule with linear warm-up over the first 50 epochs and a maximum learning rate of 10^{-4}. For the MIRA$_{boxes}$ dataset, the training time was lowered to 200 epochs, to match the almost five-fold size of the data subset. For optimization, the Adam optimizer and standard Faster R-CNN loss functions were used. Training patches of $1\,280 \times 960$ pixels were sampled from the original $5\,312 \times 2\,988$ pixel images. This resolution was chosen to match the 4:3 aspect ratio typical of microscopy images, while ensuring that each patch contained a sufficient number of cells for effective training. The patches were then downscaled by a factor of 2 to a final size of 640×480 pixels, enabling a batch size of 32 without exceeding memory constraints. To address class imbalance, we applied a custom patch sampling strategy that oversampled regions containing underrepresented classes, such as white blood cells. Model performance was monitored using the cro:mapmean average precision (mAP) on the validation set, and the final model was selected based on the best validation mAP.

For inference on the full-resolution $5\,312 \times 2\,988$ pixel images, we used the SAHI framework (Akyon et al., 2021, 2022), which performs sliding-window predictions and applies cro:nmsnon-maximum suppression (NMS) to eliminate duplicate detections across overlapping patches. As a post-processing step, we removed all predicted bounding boxes with an area smaller than $2\,500$ pixels or larger than $140\,000$ pixels. These thresholds were determined based on the minimum and maximum annotation sizes observed in the original NIH dataset.

Training was performed on an NVIDIA A100 GPU. Ex-

periments were implemented using the torchvision Faster R-CNN model, with PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) for streamlined training and Hydra (Yadan, 2019) for configuration management.

3.2 Evaluation

For evaluation, we computed class-wise F1 scores from the instance-level confusion matrices. Cells that were detected by Cellpose but not labeled by human annotators (i. e., ambiguous cells) were excluded from the evaluation. Annotated cells that were not detected by the model were considered cro:fn-detfalse negatives due to detection failure (FN_{det}), while model predictions that were not annotated and not labeled as ambiguous were considered cro:fp-detfalse positives due to detection failure (FP_{det}). The class-wise F1 score for class c was computed as:

$$F1(c) = 2 \cdot \frac{\mathsf{Prec}(c) \cdot \mathsf{Rec}(c)}{\mathsf{Prec}(c) + \mathsf{Rec}(c)} \ , \text{ with} \tag{1}$$

$$\mathsf{Prec}(c) = \frac{TP(c)}{TP(c) + FP_{\mathsf{cls}}(c) + FP_{\mathsf{det}}(c)}$$
$$= \frac{M_{cc}}{\sum_{i=1}^{N+1} M_{ic}} \ , \text{ and} \tag{2}$$

$$\mathsf{Rec}(c) = \frac{TP(c)}{TP(c) + FN_{\mathsf{cls}}(c) + FN_{\mathsf{det}}(c)}$$
$$= \frac{M_{cc}}{\sum_{i=1}^{N+1} M_{ci}} \ . \tag{3}$$

Here, M_{ij} denotes the element in the i-th row and j-th column of the confusion matrix, i. e., the number of cells labeled as class i and predicted as class j. N is the number of cell classes, and the $(N+1)$-th row and column represent false positive (FP_{det}) and false negative (FN_{det}) detections, respectively.

3.3 Results

Figure 3a presents the confusion matrix of the Faster R-CNN model trained on the NIH$_{polys}$ subset and evaluated on the MIRA$_{boxes}$ subset, and vice versa. Results are displayed as row-normalized percentages along with absolute cell counts. Overall, the model performs better when trained on the MIRA$_{boxes}$ and evaluated on the NIH$_{polys}$ subset than the other way round, reflected by a lower proportion of off-diagonal entries in the confusion matrix. When training on NIH$_{polys}$ and testing on MIRA$_{boxes}$, a comparably high ratio ($> 20\,\%$) of infected cells was misclassified as non-infected, indicating reduced recall for malaria detection. Furthermore, $81.8\,\%$ of the cells annotated as *ambiguous* were

(a) train: NIH$_{polys}$, test: MIRA$_{boxes}$

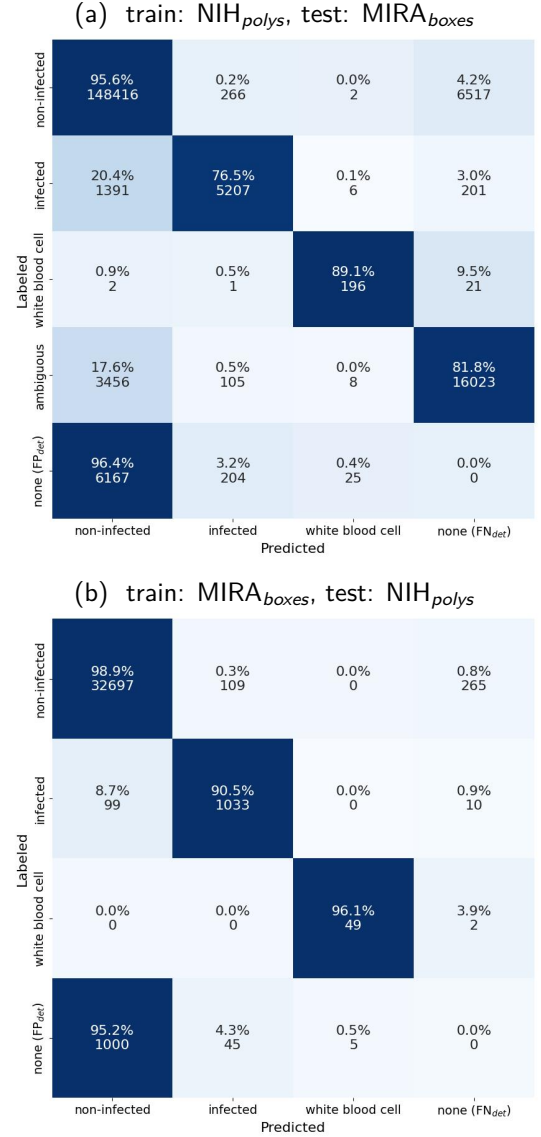(b) train: MIRA$_{boxes}$, test: NIH$_{polys}$

Figure 3: Confusion matrices for Faster R-CNN predictions on the NIH subsets. Each matrix shows row-normalized percentages along with absolute cell counts. The last row indicates cro:fpfalse positives (FPs), i. e., cell instances detected by the model but not annotated in the dataset. The last column indicates cro:fnfalse negatives (FNs), i. e., annotated cell instances that were not detected by the model.

not detected by the model. Closer inspection of these cells revealed that ambiguous cells were often located near the field-of-view borders, where annotations were inconsistently applied. Specifically, these border cells were frequently unannotated in both, the NIH$_{polys}$ and NIH$_{points}$ dataset. This suggests a possible labeling bias, which is further illustrated in Fig. 4, where white arrows indicate unlabeled yet clearly visible cells.

Table 2 summarizes the detection performance, reported as precision, recall, and F1 scores computed from the confusion matrices according to Eqs. (1) to (3). For each dataset,

(a) Sample from NIH$_{polys}$
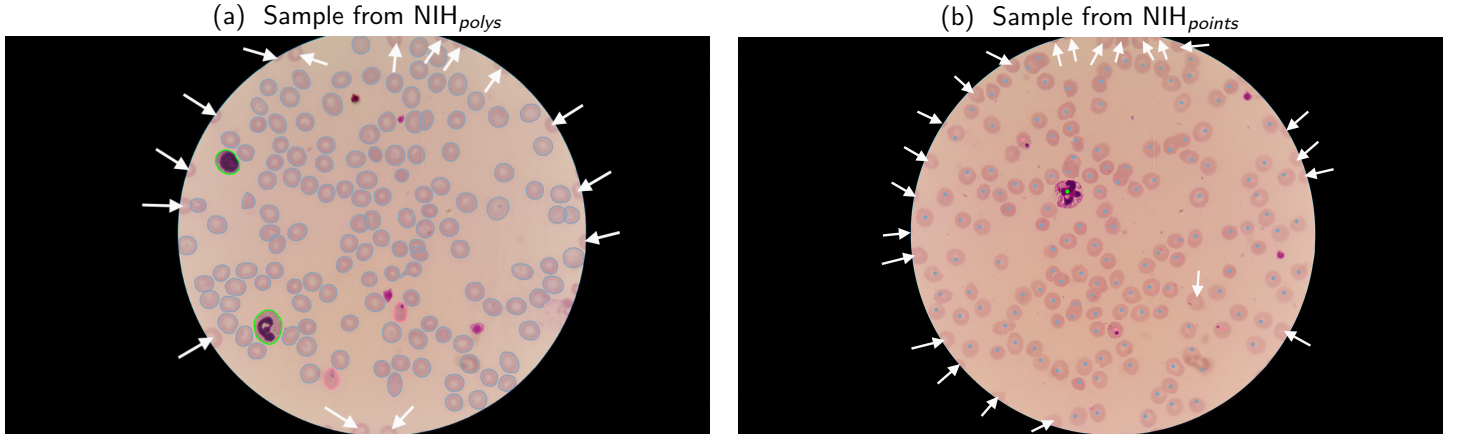
(b) Sample from NIH$_{points}$

Figure 4: Representative samples from NIH subsets with white arrows indicating non-annotated cells at the border of field of view: (a) sample from the polygon subset with detailed contour annotations, (b) sample from the point subset with spot annotations in the cell center.

training was repeated with three different random seeds, and we report the average performance as mean $\pm$ standard deviation ($\mu \pm \sigma$).

Table 2: Class-wise F1 score ($\mu \pm \sigma$) of detection model trained on NIH$_{polys}$ subset and evaluated on the MIRA$_{boxes}$ subset and vice versa.

	NIH$_{polys} \rightarrow$ MIRA$_{boxes}$	MIRA$_{boxes} \rightarrow$ NIH$_{polys}$
Precision		
non-infected cells	0.96 ± 0.01	0.97 ± 0.00
infected cells	0.91 ± 0.01	0.86 ± 0.01
white blood cells	0.90 ± 0.04	0.88 ± 0.03
Recall		
non-infected cells	0.97 ± 0.01	0.99 ± 0.00
infected cells	0.77 ± 0.01	0.91 ± 0.01
white blood cells	0.92 ± 0.03	0.96 ± 0.00
F1 score		
non-infected cells	0.96 ± 0.01	0.98 ± 0.00
infected cells	0.84 ± 0.01	0.88 ± 0.00
white blood cells	0.91 ± 0.04	0.92 ± 0.02

The results demonstrate high performance for the detection of non-infected red blood cells and white blood cells, with F1 scores above $90\,\%$. Infected cells were detected with an average F1 score of 0.84, when training on NIH$_{polys}$ and 0.88 when training on MIRA$_{boxes}$, indicating good but comparatively lower performance. The repeated training runs demonstrate low variability, indicated by a low standard deviation of performance results.

The performance metrics again demonstrate a superior performance of the model trained on MIRA$_{boxes}$. This especially holds for the recall of infected cells, with average values of 0.77 for training on NIH$_{polys}$ and 0.91 for training on MIRA$_{boxes}$. This observation could be attributed to discrepancies in labeling consistency, but also to the higher volume of annotated instances ($6\,810$ vs. $1\,142$), which provides more diverse training examples to the model.

4. Discussion and Summary

This study presents a revised version of the NIH malaria dataset with instance-level annotations in COCO format, facilitating the development of deep learning-based object detection models for the automatic detection of infected cells. We validated these annotations by training a Faster R-CNN to detect infected and non-infected red blood cells, as well as white blood cells, achieving an F1 score of up to 0.88 for the detection of infected cells. For trained microscopists, the cro:whoWorld Health Organization (WHO) guidelines recommend a minimum recall of infected malaria samples of 0.90 (World Health Organization, 2009), which our system achieves on a cellular level. Therefore, the system meets the minimum competency level required in a diagnostic setting. Nevertheless, our analysis of ambiguous cells revealed inconsistencies in the original annotations, where especially at the image borders cells were not labeled by the pathologists. However, it is difficult to tell whether these cells were simply overlooked or not labeled on purpose as a reliable malaria diagnosis might not be possible on partially visible cells. This raises broader concerns about ground truth quality in biomedical datasets likely caused by a trade-off of labeling precision and time investment. To the best of our knowledge, the original dataset was annotated by a single expert, which can introduce a considerable labeling bias. Future work could address this with additional manual annotation rounds with consensus labeling by multiple experts or introducing a separate class for partially visible cells. For evaluating the performance of machine learning models, we recommend excluding these cells from evaluation.

Despite the challenges associated with partially labeled data, our results demonstrate that annotation conversion via existing tools such as Cellpose, followed by targeted manual curation, can yield training data of sufficient qual-

ity to support robust model performance. This finding is particularly relevant for resource-constrained settings where detailed annotations are expensive or infeasible.

In addition to annotation consistency, we also observed differences in model performance between the two subsets of the NIH dataset, likely driven by the varying number of annotated instances available for training. This highlights the importance of dataset size and diversity for learning subtle morphological features, such as the presence of ring-stage parasites. Furthermore, our initial dataset assessment demonstrated a high class imbalance of healthy and infected cells. We compensated for this to some extent by employing a customized patch sampling strategy, but in future work dedicated augmentation strategies or class-balanced loss functions could be integrated.

Overall, our work contributes an enhanced dataset and a robust pipeline for parasite detection in microscopy, supporting further research into automated malaria diagnosis.

Conflicts of Interest

The authors do not have any conflicts of interest to declare.

Acknowledgments

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, in accordance with all applicable laws and regulations regarding the treatment of human subjects. The dataset used in this study was publicly released by Kassim et al. (2020), who obtained the necessary ethical approvals as documented in the original publication. No additional ethical approval was required.

Data availability

The dataset (Kassim et al., 2020) used in this study was developed and funded by the U.S. cro:nlmNational Library of Medicine (NML), part of the National Institutes of Health (NIH), and is publicly available for commercial and non-commercial use. Use of this dataset is governed by a license that requires proper attribution. We acknowledge the source of the data as follows: "Courtesy of the U.S. National Library of Medicine." The dataset and associated information are available at: `https://lhncbc.nlm.nih.gov/publication/pub9932`. Please cite the dataset as described by Kassim et al. (2020). The updated annotation set with the modifications described in Section 2.2 is publicly available via Zenodo: `https://doi.org/10.5281/zenodo.17514694`.

References

Fatih Cagatay Akyon, Cemil Cengiz, Sinan Onur Altinuc, Devrim Cavusoglu, Kadir Sahin, and Ogulcan Eryuksel. SAHI: A lightweight vision library for performing large scale object detection and instance segmentation. *Zenodo*, November 2021.

Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 966–970, 2022.

William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. URL `https://github.com/Lightning-AI/lightning`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Yasmin M Kassim, Kannappan Palaniappan, Feng Yang, Mahdieh Poostchi, Nila Palaniappan, Richard J Maude, Sameer Antani, and Stefan Jaeger. Clustering-based dual deep learning architecture for detecting red blood cells in malaria diagnostic smears. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1735–1746, 2020.

Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12):1634–1641, 2022.

Mahdieh Poostchi, Kamolrat Silamut, Richard J. Maude, Stefan Jaeger, and George Thoma. Image analysis and machine learning for detecting malaria. *Translational Research*, 194:36–55, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Waqas Sultani, Wajahat Nawaz, Syed Javed, Muhammad Sohail Danish, Asma Saadia, and Mohsen Ali. Towards low-cost and efficient malaria detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20655–20664. IEEE, 2022.

World Health Organization. *Malaria microscopy quality assurance manual: v.1*. WHO Regional Office for the Western Pacific, 2009.

World Health Organization. *WHO guidelines for malaria*. WHO, 2024.

Omry Yadan. Hydra - a framework for elegantly configuring complex applications, 2019. URL `https://github.com/facebookresearch/hydra`.