Data Sheet for Malaria Image Data Collected in Uganda Under the Lacuna Project

We present the Lacuna Malaria data sheet created by Makerere Artificial Intelligence Lab created by a group of researchers from the Makerere Artificial Intelligence Lab in Makerere University in Uganda. We follow the datasheet for dataset framework created by (Gebru et al. 2021).

Motivation	
For what purpose was the data set created?	
The dataset presents a set of blood slide images captured from a microscope using a smartphone camera. The images are annotated to show different objects like malaria parasites and white blood cells and this dataset can be used to develop automated solutions for mobile microscopy. Was there a specific task in mind?	The dataset was primarily created to be used to develop machine learning models for the detection of malaria parasites and other properties from images of thick and thin blood slides.
Who created the dataset?	The dataset was created by the Makerere Artificial Intelligence Lab at Makerere University. The images were captured and annotated at Kiruddu National Referral Hospital by microscopists at the hospital.
Composition	
What do the instances that comprise the dataset represent?	The dataset contains images(JPEG) and annotations in YOLO format (Redmond et al. 2016).
How many instances are there in total (of each type, if appropriate)?	4000 images and 4000 annotations.
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset contains all instances.

What data does each instance consist of? "Raw" data or features?	The image is JPEG file (Christopoulos et al. 2000) showing the blood slide as captured by the smart phone camera on the microscope.
Is there a label or target associated with each instance? If so, please provide a description.	The annotation is a text file showing the normalized bounding box coordinates and the ID of the object type for each bounding box. There is also a excel file showing metadata for some of the images. This includes a slide ID representing the physical slide from which the image was captured, the x and y stage micrometer readings and the type of phone from which the image was captured.
Is any information missing from individual instances?	No
Are relationships between individual instances made explicit?	Yes
Are there recommended data splits (for example, training, devel- opment/validation, testing)?	No
Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.	No
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	The dataset is self-contained
Does the dataset contain data that might be considered confidential?	No
Does the dataset contain data that, if viewed directly, might be of- fensive, insulting, threatening, or might otherwise cause anxiety?	No
Collection Process	
How was the data associated with each instance acquired?	The images in the dataset were captured by placing a smartphone over a microscope to show the view of the blood slide through the eyepiece of the microscope. The phone was placed in a custom D printed adapter to support and simplify the process of the data capture. Along with the image, the slide from which the image was captured, the stage micrometer readings of the microscope and the objective lens settings were recorded.

What mechanisms or procedures were used to collect the data?	The data collection process was done following a well defined protocol developed in conjunction with the expert microscopists. The protocol defined the criteria for selection of candidate slides from which to capture the images, the preparation of the slides for the image capture, the number of images to be capture per slide and the phone to be used. There was also a need to design a 3D printable smartphone adapter to be used for the setup.
If the dataset is a sample from a larger set, what was the sampling strategy?	The dataset is not from a larger sample but the slides from which the images were captured were randomly sampled and evaluated for presence of malaria parasites through the microscopes before the final selection. This was done to ensure that the images captured had the objects of interest.
Who was involved in the data collection process?	Researchers from the Makerere Artificial Intelligence Lab at Makerere University and microscopists from Kiruddu National Referral Hospital in Uganda.
Over what timeframe was the data collected?	The data was collected over 6 months from August 2022 to February 2023.
Were any ethical review processes conducted (for example, by an institutional review board)?	Yes. The data collection was approved and guided by an established Institutional Review Board (IRB).
Preprocessing, cleaning, and labelling	
Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?	The data was visually inspected to remove images that were considered to be of poor quality.
Was the "raw" data saved in addition to the preprocessed/cleaned/ labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.	No
Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.	The VGG Image Annotator[1] (Dutta et al. 2019).

[1] https://www.robots.ox.ac.uk/ vgg/software/via/

Uses	
Has the dataset been used for any tasks already? If so, please provide a description.	No
Is there a repository that links to any or all papers or systems that use the dataset?	No
What (other) tasks could the dataset be used for?	The dataset can be used to study other microscopy tasks such as determination of parasitemia.
Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses?	No
Distribution	
Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.	No
How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?	https://doi.org/10.7910/DVN/VEADSE
When will the dataset be distributed?	The first version of the dataset was published in June 2023 and consequent revisions will show up as version updates.
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?	No
Have any third parties imposed IP-based or other restrictions on the data associated with the instances?	No
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	No
Maintenance	
Who will be supporting/hosting/maintaining the dataset?	The main version of the dataset is hosted on the Harvard Dataverse and will be maintained by the Makerere Artificial Intelligence Lab.

How can the owner/curator/ manager of the dataset be contacted (for example, email address)?	g.nakasi.rose@gmail.com
Is there an erratum?	No.
Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)?	Yes the dataset will be updated.
Will older versions of the data- set continue to be supported/hosted/ maintained? If so, please describe how.	Older versions of the dataset will remain available for purposes of tracking changes but support will mainly be offered only on the newest version of the dataset. If any one wishes to get support on the older versions of the dataset, they can inform the point of contact in writing.
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?	The dataset has been published under the CC BY 4.0 creative commons license which allows for remixing and adaptation for any purpose.

References

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

R. Joseph, D. Santosh, G. Ross, and F. Ali, “You Only Look Once: Unified, Real-Time Object Detection,” arXiv (Cornell University), Jun. 2015, doi: 10.48550/arxiv.1506.02640.

C. Christopoulos, T. Ebrahimi, and A. N. Skodras, JPEG2000. 2000. doi: 10.1145/357744.357757.

A. Dutta and A. Zisserman, The VIA Annotation Software for Images, Audio and Video. 2019. doi: 10.1145/3343031.3350535.