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

We present the Lacuna Malaria data sheet created by MinoHealth AI Labs. We follow the datasheet for the dataset framework created by (Gebru et al. 2021).

Motivation

For what purpose was the data set created? Was there a specific task in mind?

The dataset was created with the specific purpose of advancing research and development in the field of malaria diagnosis and detection. By collecting the dataset from Ghana, a diverse and well-annotated collection of malaria thin and thick blood smear images will be made available to support the development of accurate, efficient, and scalable methods for malaria diagnosis.

Who created this data set (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

The dataset was created by a team of data scientists from the minoHealth AI Labs, with support from medical officers, medical laboratory scientists and officers.

Composition

What do the instances that comprise the data set represent (e.g., documents, photos, people, countries)?

Each instance in the dataset includes thick and thin blood smear images captured through the lens of a microscope (JPEG), image status for thick smear (Parasite, White Blood Cells), image status for thin smear (Gametocytes, Trophozoites, White Blood Cells, Artifacts, Ring) file type (images and bounding box annotations) and location (this variable though is without values).

How many instances are there in total (of each type, if appropriate)?

There are 3,000 instances of thick smear images and “1,000” instances of thin smear images.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of the cases from a larger set?

The dataset contains various instances that were captured in two hospitals in the Greater Accra region, Princess Marie Louise Hospital.

What data does each instance consist of? “Raw” data or features?

Each instance includes: the blood smear image (thick, thin)

Is there a label or target associated with each instance? If so, please provide a description.

Based on their presence in a blood smear image, each instance is associated with a class label of parasites and white blood cells (WBC) for thick smear images and growth stages (gametocytes, trophozoites, Ring stage), WBCs and artifacts.

Is any information missing from individual instances?

Are relationships between individual instances made explicit?

Yes, there are two sets of data, the thick blood smear set and the thin blood smear set.

Are there recommended data splits (for example, training, development/validation, testing)?

No, the data exists as two sets, a thick blood smear set and a thin blood smear set

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Due to the quality of the mobile devices used to collect the data (capture the images), some instances have low quality making it difficult to observe classes in the data.

Is the dataset self-contained, or does it link to or otherwise rely on external resources?

No, the dataset is self-contained, it does not rely on any other external sources

Does the dataset contain data that might be considered confidential?

No, the dataset does not contain data that might be considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No, the dataset does not contain data that might be offensive, insulting, threatening or data that may cause anxiety.

Collection Process

How was the data associated with each instance acquired?

The data associated with each instance was acquired from the Princess Marie-Louis Children's Hospital in Accra, Ghana.

What mechanisms or procedures were used to collect the data?

The images were collected manually using a mobile phone camera. The images were taken at different angles to provide a good balance and introduce variation in the dataset.

If the dataset is a sample from a larger set, what was the sampling strategy?

The final dataset is the complete dataset and not a sample of any other dataset

Who was involved in the data collection process?

The minoHealth AI team with supervision from lab technicians and scientists participated in the collection of this data.

Over what timeframe was the data collected?

The data was collected over a total period of 8 months.

Preprocessing, cleaning, and labeling

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 annotated with bounding boxes using annotation tools (makesense.ai). This was the only preprocessing step done on the dataset.

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.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

Yes, the annotation tool *makesense.ai* can be accessed *here*

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

During the annotation, the dataset was used to develop models to train object-detection models to speed up annotation in a semi-supervised approach.

Is there a repository that links to any or all papers or systems that use the dataset?

None at the moment

What (other) tasks could the dataset be used for?

1. Cell counting
2. Anaemia Detection
3. Blood smear quality assessment
4. Diagnosis of hematological malignancies (e.g. leukemia)

Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses?

Nothing about the composition of the dataset would affect future use for the use case/task the dataset was curated for.

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. How will the dataset be distributed (for example, tarball on website, API, GitHub)?

Does the dataset have a digital object identifier (DOI)?

When will the dataset be distributed?

This will be determined by the Makerere team in Uganda.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

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 MinoHealth AI team as well as the Makereke team wil be responsible for maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (for example, email address)?

Darlington Akogo can be be contacted on his email address - darlington@gudra-studio.com

Is there an erratum?

No

Will the dataset be updated (for example, to correct labelling errors, add new instances, or delete instances)?

Updates to the dataset will be communicated to the public through the datasheet or data cards on data hosting websites.

Will older versions of the data- set continue to be supported/hosted/ maintained? If so, please describe how.

The data which is publicly available will be maintained by MinoHealth AI Labs and Makerere University. Information regarding dataset version will be communicated through datasheets and data cards on online hosting platforms

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

The dataset and the datasheet will be made publicly available. Any contribution can be directed to the authors, MinoHealth AI Labs and Makerere University.

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.