# NLM Malaria Screener - Sudan Dataset

This dataset contains thick and thin blood smear microscopy images from 190 patients, including 100 positive patients (*P. falciparum* and *P. vivax*) and 90 negative patients. It was acquired at two primary hospitals in Sudan, one in the Alsororab (SOR) area and another in the Gezira Slanj (GS) area, where Giemsa-stained blood smears were photographed using a smartphone camera. Each site contributed images from 95 patients, one slide per patient, with both a thick and a thin smear. The blood smear images were manually annotated by WHO Level 1 expert microscopists, de-identified, and archived in PNG formats. The expert microscopists provided image-level counts, including parasite counts and white blood cell (WBC) counts for thick smears, and red blood cell (RBC) counts and infected RBC counts for thin smears.

A total of 2944 images were gathered from thick blood smears (15.5 images/patient) and 875 images from thin blood smears (4.6 images/patient). For thick smear images, 34256 parasites and 38216 WBCs were counted, while 395 infected RBCs and 220686 RBCs were counted for thin smear images.

## Folder Structure
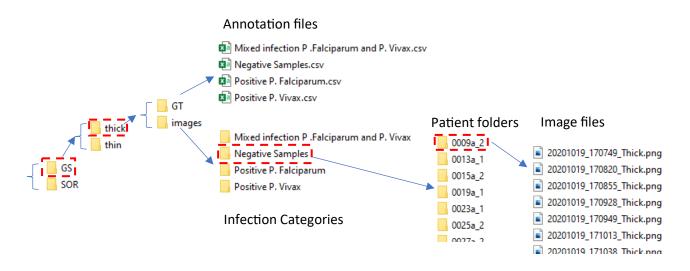
Data folders are organized as shown in the image below:



*Figure 1. Folder structure*

# Data statistics

*Table 1. Image counts*

| | | Number of patients/slides | Thick smear images | Thin smear images |
|---|---|---|---|---|
| **Site GS** | P. Falciparum | 21 | 395 | 90 |
| | P. Vivax | 28 | 409 | 125 |
| | Negative | 45 | 705 | 197 |
| | Mixed infection of P. F. & P. V. | 1 | 28 | 4 |
| | **Sum** | **95** | **1537** | **416** |
| | | | | |
| **Site SOR** | P. Falciparum | 40 | 609 | 197 |
| | P. Vivax | 10 | 124 | 53 |
| | Negative | 45 | 674 | 209 |
| | **Sum** | **95** | **1407** | **459** |

*Table 2. Manual annotation counts*

| | | Thick | | Thin | |
|---|---|---|---|---|---|
| | | Parasite counts | WBC counts | Infected RBC counts | RBC counts |
| **Site GS** | P. Falciparum | 10412 | 4003 | 75 | 24758 |
| | P. Vivax | 4126 | 5610 | 44 | 30978 |
| | Negative | 0 | 8758 | 0 | 51567 |
| | Mixed infection of P. F & P. V | 61 | 170 | 0 | 1184 |
| | **Sum** | **14599** | **18541** | **119** | **108487** |
| | | | | | |
| **Site SOR** | P. Falciparum | 18622 | 8314 | 261 | 45095 |
| | P. Vivax | 1035 | 2013 | 15 | 11432 |
| | Negative | 0 | 9348 | 0 | 55672 |
| | **Sum** | **19657** | **19675** | **276** | **112199** |

## Image files

All images are color images with RGB channels in PNG format. Because images have been captured through the eyepiece of a microscope, the visible region is circular.

Image dimensions:

GS site, thick smear: 4216 x 3184

       thin smear: 4128 x 3096

SOR site, thick smear: 3184 x 4216

       thin smear: 3096 x 4128

Thick smear images have slightly larger dimensions because they include a 44-pixel wide zero-padding at their borders.

Image orientation:

GS site images are saved in landscape mode.

SOR site images are saved in portrait mode.

## Annotation files

| Patient ID | Slide ID | Image ID | Manual parasite counts | Manual WBC counts |
|---|---|---|---|---|
| Patient_id_thick | slide_id_thick | image_id_thick | parasite_count_gt | WBC_count_gt |
| 0001a | | 1 20201015_212811_Thick | 0 | 48 |
| 0001a | | 1 20201015_212848_Thick | 0 | 42 |
| 0001a | | 1 20201015_212940_Thick | 0 | 43 |
| 0001a | | 1 20201015_213010_Thick | 0 | 51 |

| Patient ID | Slide ID | Image ID | Manual total RBC counts | Manual infected RBC counts |
|---|---|---|---|---|
| Patient_id | slide_id | image_id | cell_count_gt | infected_count_gt |
| 0011b | | 1 20201019_173100_Thin | 240 | 3 |
| 0011b | | 1 20201019_173115_Thin | 265 | 3 |
| 0011b | | 1 20201019_173137_Thin | 274 | 5 |
| 0011b | | 1 20201019_173203_Thin | 278 | 5 |

*Thin smear images of positive patients can have a zero count for infected RBCs. This is because the expert microscopists used thick smears to determine whether a patient was positive or negative.

Mapping between annotation and image files:

Each line in an annotation file points to one image. The entries in the Patient ID and Slide ID columns can be concatenated to generate the patient folder name; then, the Image ID can be used to identify the image file within this folder.



# Acknowledgement

# Reference

We request that publications resulting from the use of this data attribute the source (National Library of Medicine, National Institutes of Health, Bethesda, MD, USA) and cite the following publication:

*Yu H, Mohammed FO, Abdel Hamid MM, Yang F, Kassim YM, Mohamed AO, Maude RJ, Ding XC, Owusu EDA, Yerlikaya S, Dittrich S, Jaeger S. Patient-level performance evaluation of a smartphone-based malaria diagnostic application. Malaria Journal. 2023.*

# Contact

**Hang Yu** [email: hang.yu@nih.gov, phone: (301) 827-5527] or
**Dr. Stefan Jaeger** [email: stefan.jaeger@nih.gov, phone: (301) 435-3198]

Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bethesda, MD 20894, USA