EFFECTIVE CREATION OF GROUND TRUTH DATASET FOR MALARIA DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence is transforming how health care is delivered across the world. This has been evident in pathology detection, surgery assistance and early detection of diseases such as breast cancer. However, these technologies often require significant amounts of data. Yet, in many developing countries, there is a shortage of available quality data for research, development and evaluation of health-based artificial intelligence tools. To address this deficiency, we have collected and annotated 10,000 images of a stained blood smear to be used to develop different artificial intelligence tools for malaria diagnosis. In the collection of this dataset, we have included demographic information of respondents such as age, gender, and location to seek any relationship with the number of parasites present in the patients. We envision the dataset to be used to improve existing algorithms in malaria diagnosis and create a new benchmark.

1 INTRODUCTION

Malaria is a very dangerous disease caused by plasmodium parasites transmitted to people through the bite of infected female Anopheles mosquitoes (WHO, 2018). While nearly half of the world's population is at risk of malaria, the Africa region is carrying a high share of malaria burden (WHO, 2018). According to the WHO (2018) malaria report, 93% of the deaths caused by malaria were reported in Africa in 2017 of which 60% of the reported cases were children under 5 years. However, prompt diagnosis of malaria and treatment reduces the chance of a mild case becoming severe and eventually fatal. (Tangpukdee et al., 2002) highlighted that having practical and effective clinical malaria diagnostic tools is an important step in eradicating malaria since it is difficult to differentiate malaria from other tropical diseases by relying only on patients' signs and symptoms.

Microscopic diagnosis is the gold standard for laboratory confirmation of malaria cases. It involves an examination of the stained thick blood smear (quantification of parasites) and thin blood smear (identification of parasites species) under the oil immersion lens of the microscoper (AR et al., 2007). The staining of the smear is done using the Giemsa solution(Warhurst DC, 1996). Although this technique provides the doctor with vital information in guiding them in making initial treatment decisions, it still requires 15 to 20 minutes for a single diagnosis and the reliability of results depends on the expertise of the lab technician. In an area with high disease burden such as Africa, with limited resources this becomes impractical. More recently, mRDTs (Malaria Rapid Diagnostic Test) have been widely adopted, and which do not require experts or laboratory equipment and which takes 10 to 15 minutes for a single diagnosis. However, mRDT is not capable of detecting the severity of malaria, and furthermore has exhibited a wide variation in its sensitivity for patients with a lower number of parasites (A, 2002).

Emerging technologies within Artificial Intelligence (AI) are being used to improve the provision of better healthcare services such as in the diagnosis of critical diseases, preliminary prediction of diseases, treatments, and surgery. Several AI techniques such as Deep Convolutional Networks have been used for the detection of malaria parasites (Sánchez-Sánchez, 2015). These techniques have great potential for improving the diagnosis and treatment of malaria, particularly in developing countries where there is a limited health experts and medical equipment.

However, one of the major drawbacks that hinder the development of AI-based applications for healthcare services in developing countries, such as Tanzania, is the lack of data for training, testing, and validation of such tools. In these countries, there is limited access to the available data from both government and non-governmental organizations. Moreover, even the little available data still lacks the necessary qualities in terms of pixels, labels that are required for the development of AI tools. In this project, we aimed to create a malaria ground-truth dataset to aspire more innovative solutions for fast, reliable and low-cost malaria diagnosis tools and other research

2 Dataset

2.1 Sample

This first phase involve capturing microscopic images of stained blood smear using a smartphone that was observed under a microscope. The images captured are from patients that had been requested, by a doctor, to receive a malaria test. In Tanzania, malaria risk is high throughout the country except for highlands such as in Arusha Regions and prominent in lowland such as Dar es Salaam. In our research, we have collected sample data from 400 patients in two regions of Tanzania namely Morogoro and Dodoma. In Morogoro region, we decided to collect positive samples (patients found with malaria parasites) because it is a malaria burden region due to suitable climate conditions for transmission, and in Dodoma region, we collected negative Samples (patients without malaria parasites) because the region has less malaria infection. We were able to collect 300 positive samples and 100 negative samples from Morogoro and Dodoma regions respectively. These patients include both female and male and their ages ranged from 2 months up to 60 years.

2.2 Reagent Preparationl

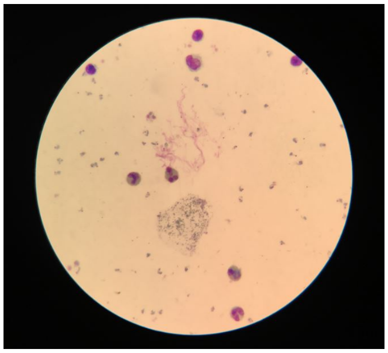
Before subjecting a blood sample to a microscope for observation and image capturing, it had to be prepared using a reagent. For that case, a buffer solution using 1 liter of distilled water and 1 buffer tablet were prepared with the aim of making a 7.2 PH solution. Also, a Giemsa working solution was prepared by taking 2.5 ml of Giemsa stain stock into 25 ml of water making a 10% concentration. The working solution was then filtered using a circle filter paper. After filtration, the working solution was ready to be used for staining the blood samples. The samples were placed horizontally and stained for 10 minutes. The stained samples were washed by using tap-water and placed vertically using a staining rack for the water to run off. At this stage, the dried stained blood samples were ready to be observed under a 100 magnification of Olympus microscope. So far, we have stained the blood samples from 50 positive patients collected from Morogoro region and from 100 negative patients collected from Dodoma.

2.3 Image collection

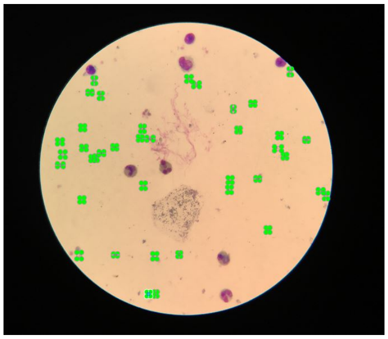
A small portion of immersion oil was applied to the stained thick blood smear to enhance visibility and then the slide containing blood smear was placed under an Olympus CX 21 microscope for observation. The lens used had 100x magnification as recommended by the WHO (Payne, 1988). The microscope was continually adjusted by a lab technician to ensure proper focus. Then, the iPhone 6s+ mobile phone was mounted to the microscope using the Labcam Microscope Adapter. Microscopic images were then transferred directly to the mobile phone for capturing and storage. For malaria diagnosis, a lab technician is required to go around not less than 100 fields for a single slide under observation (Payne, 1988)). Therefore, approximately 100 images were captured for every slide under observation. For the 50 positive patients that we have started with, each patient contributed a single slide of blood smear and a single slide of blood smear produced 100 images making a total of 5000 images from affected patients. Likewise, 100 negatives (un-infected) patients produced a total of 10000 images. All 5000 images from 50 positive(infected) patients required annotation. On the other hand, the 10000 images from uninfected patients did not require annotation. The images captured are in JPG format, had a resolution of 4302 X 3204 pixels and a size of approximately 1 MB. The images were stored in a folder labeled with a date the slide was taken followed by a sample number for identification of the image.

2.4 Image annotation

A team of experts from the College of Health Science of the University of Dodoma and Benjamin Mkapa Hospital performed the annotation of the 5000 images. The annotated images were then



(a) Captured image by a smart phone camera installed to a microscope lens



(b) Annotated Image

verified by two lab technicians from Benjamin Mkapa Hospital. The images were annotated using the LabelIMG annotation tool. The images are annotated by bounding boxes for the plasmodium object class. Annotators were instructed to label a target class by drawing the smallest possible box that contains all the visible parts of the plasmodium. The output of the annotation is an XML file with specific details on the position of the bounding box. The time taken to annotate a single image with a fewer number of parasites, less than 20 parasites, was less than 2 minutes while for a case with a higher number of parasites, more than 100 parasites, was around 15 minutes. In general for one patient, it took about 8 hours to annotate the blood sample.

2.5 Verification description

Annotated images were verified by the lab technician to ensure the accuracy of the labels. Verification was done by three experts and each verified file was used to calculate the deviation from the original annotation as shown on table 1. The deviation was calculated by counting the number of added/deleted bounding box over the initial number of the bounding box in the original annotation. The final output of the verified dataset was a bounding box that occurred more than twice in the four sets (1 original annotation, and 3 expert verification set).

Table 1: Summary of deviation percentage during verification process.

Verifiers	**Avg. deviation (%)**
Verifier 1	3
Verifier 2	4
Verifier 3	2

3 Conclusion

The research study aimed at creating an image data set for malaria diagnosis by creating a bounding box around a parasite. Three experts verified the dataset and average was calculated to have an optimal annotation on 5000 images. In order to resolve the annotation variation of different experts, we computed an average of the annotation from three lab technicians who were doing the final verification of the same images of slides. The future work is to include labeling of red blood cells so as to facilitate automatic quantification and calculation of parasitemia. This dataset will be used to develop more artificial intelligence based malaria diagnostic tools. It will also be sent to the WHO/ITU Malaria AI Focus Group responsible for regulating standards of malaria tools worldwide. In addition, we will develop a semi-automatic tool that will enhance the data creation process as has been outlined in this paper.

REFERENCES

Moody A. Rapid diagnostic tests for malaria parasites. *Clin Microbiol Rev*, 15:66–78, 2002.

Bharti AR, Patra KP, Chuquiyauri R, Kosek M, Gilman RH, Llanos-Cuentas A, and Vinetz JM. Polymerase chain reaction detection of plasmodium vivax and plasmodium falciparum dna from stored serum samples: implications for retrospective diagnosis of malaria. *Am J Trop Med Hyg*, 77:444–446, 2007.

Payne. Use and limitations of light microscopy for diagnosing malaria at the primary healthcare level, bulletin of the world health organization. *n/a*, 66:621–626, 1988. URL https://apps.who.int/iris/bitstream/handle/10665/264613/PMC2491188.pdf?sequence=1&isAllowed=y.

C Sánchez-Sánchez. Deep learning for identifying malaria parasites in images. *n/a*, 2015. URL GoogleScholar:https://scholar.google.com/citations?user=eDdZ3hcAAAAJ&hl=en#d=gs_md_cita-.

N. Tangpukdee, Duangdeem C., Wilairatana P., and Krudsood S. Malaria diagnosis: a brief review. *The Korean Journal of parasitology*, 47:93–102, 2002.

Williams JE Warhurst DC. Laboratory diagnosis of malaria. j clin pathol. *n/a*, 49:533–538, 1996.

WHO. Who malaria 2018 report. *n/a*, 2018. URL https://apps.who.int/iris/bitstream/handle/10665/275867/9789241565653-eng.pdf?ua=1.