|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-I-014-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 7-8 May 2020 |
| **DOCUMENT** |
| **Source:** | TG-Malaria Topic Driver |
| **Title:** | Att.1 – TDD update (TG-Malaria) |
| **Purpose:** | Discussion |
| **Contact:** | Rose Nakasi Makerere UniversityUganda | E-mail: g.nakasirose@gmail.com  |

|  |  |
| --- | --- |
| **Abstract:** | This document contains an update of the previous topic description document (TDD) (FGAI4H-H-014-A01) that was presented at meeting H in Brasilia. It covers all scientific, technical and administrative aspects relevant for setting up TG Malaria benchmarking system. |

**Table of Contents**

1 Introduction 3

1.1 Document structure 3

1.2 Topic description 3

1.2.1 Impact of the benchmarking AI Solution 4

1.3 Ethical considerations 4

1.4 Existing AI solutions 4

1.5 Existing work on benchmarking 4

1.6 AI4H Topic Group: Current topic group and its mandate 5

1.6.1 Topic group structure 5

1.6.2 Topic group participation 5

1.6.3 Tools/process of TG cooperation~~.~~ 6

1.6.4 TG interaction with WG and FG 6

1.6.5 Current Status 6

1.6.6 Next meetings 6

1.6.7 Next steps 7

Procedure. 7

1.7 Business case 8

2 Method 8

2.1 AI input data structure 8

2.2 AI output data structure 9

2.3 Test data labels 9

2.4 Scores & metrics 9

2.5 Undisclosed test data set collection 9

2.6 Benchmarking methodology and architecture 10

2.7 Reporting methodology 12

3 Results 13

4 Discussion 14

5 Declaration of Conflict of Interest 14

References 15

# Introduction

Malaria is one of the largest endemic diseases in the Sub Saharan Africa [5]. In Low developed countries (LDCs), the scourge is further buttressed by the lack of enough skilled lab technologists in health centers to detect the disease using the widely accepted gold standard Microscopy method. Thus, the need for reliable detection interventions. This explains the birth of Automated malaria detection using Artificial Intelligence (AI). The aim is to harness AI to automate the detection of malaria in a more fast, accurate and cost-effective manner. Of recent AI and machine learning techniques have been successful in different medical image analysis tasks and have a capability to improve public health.

The is document therefore aims at developing a standardised benchmarking approach for AI based detection of Malaria.

## Document structure

This TDD is dedicated to support standardised benchmarking of AI-Based detection of Malaria. The document will cover the core aspects relevant to topic including data collection and annotation, selection of AI model for Malaria detection, Algorithm performance and evaluation, collection of test datasets and benchmarking methodologies.

## Topic description

According to the World Health Organization report of 2016, nearly half of the world population is at risk of malaria [5]. Records from the WHO report of 2015 indicate that in 2015, 212 million cases reported, Malaria accounted for over 480,000 deaths, 90% of which were from Africa, 7% from S.E Asia and 2% from Eastern Mediterranean region [6]. Although there were fewer Malaria cases in 2017 than in 2010 according to WHO report of 2017, data for the period 2015-2017 highlighted that no significant progress in reducing global Malaria cases was made in this timeframe [7]. Malaria is thus of major concern to public health and therefore the need for early, fast and accurate diagnosis.

The gold standard method for detection of Malaria is microscopy of blood smear slides. Unlike Rapid Diagnostic Tests (RDTs), microscopy supports direct parasite detection and identification and provides monitoring of systemic inflammation and its response to therapy [9]. Detection of malaria requires examination of thin and thick blood smear images through conventional light microscopy. In general, Malaria parasite detection, species identification, and parasitemia determination requires expertise from trained Microscopists (lab technicians).

Effective Malaria control can be achieved by a fast, consistent and accurate diagnosis. However, this requires the expertise of Microscopists to operate the gold standard method of microscopy screening of Malaria. Unfortunately, highly Malaria endemic Countries have very few expert Microscopists to diagnose and interpret the results of the huge numbers of malaria patients.

A nationwide study in Ghana, for example, found 1.72 microscopes per 100,000 population, but only 0.85 trained laboratory technicians per 100,000 population [1] which is grossly inadequate. As a result, diagnoses are often made on the basis of clinical signs and symptoms alone, which are error-prone and leads to higher mortality, drug resistance, and the economic burden of buying unnecessary drugs [2].

Computational Microscopy using Artificial Intelligence technologies which is the backbone of this TDD aims to reduce the need for many human Microscopists by providing a fast, consistent and accurate diagnosis with minimum human intervention. AI models have the capability to learn good representations of image data with reduced turnaround time bridging the gap for lack of enough skilled Microscopists and significantly improving diagnostic performance and reducing health costs associated to patient care and treatment.

### Impact of the benchmarking AI Solution

The benchmark solution for AI based detection of malaria should focus on developing AI tools that can adequately detect malaria in a fast, accurate, cost effective and reliable manner. The benchmark will be impactful if the selected solutions are robust enough and relevant especially in developing but highly Malaria Endemic Countries. This is envisioned to improve public health and to synergise with goal 3 Target 3.3 of the Sustainable Development Goal (SDG) which aims to have Malaria endemic end by 2030.

## Ethical considerations

TBC

**Ethical consideration of benchmarking including its data acquisition**:

* Ethical consideration of data collection must follow Ethical practises like acquisition of ethical approvals from Country medical Authorities through working IRBs.
* Data capture and annotation must be done by qualified medical experts.
* Data anonymity must be considered by removing all references to personal identifiers.
* Integration of other datasets must also prevent personal identifiers.

**Ethical consideration on usage of AI**:

* For any testing task, patients must be informed about the use of the AI tool in detection of malaria and consent from them must be sought.
* Best scientific practices and accuracies of the model must be assured.

## Existing AI solutions

At the AI and Data Science lab of Makerere University, we have deployed both traditional machine learning and deep learning algorithms for pathogen detection in thick blood smear samples and improvements in detection accuracies have been registered. We have also extended this to other related microscopy diagnosis challenges for example in the detection of tuberculosis and intestinal parasites [4].

An extensive study by Rosado et al [3] reviewed the various image processing and analysis approaches for the automated detection of Malaria with the conclusion that improvements in accuracy are still needed. Some AI tools tend to fail on the undisclosed data sets due to false alarms. There is currently no certified AI based solution for Malaria diagnosis. A major factor contributing to this is the lack of availability of a bigger and diverse standardised dataset from which to infer and draw comparison from the different AI Solutions. Existing AI solutions have also focussed on single detection goals rather than learning complex relationships between different datasets that could provide a more representative diagnosis approach for better realistic results.

 There is need to collect a large sample dataset that captures different settings of the image to create a wide array of data complexities that depict real life implementations. Datasets such as demographics, environment and any other contributory factor to Malaria prevalence could be captured to assure a more dependable analysis

## Existing work on benchmarking

TBC

There has been discussion on having a need for a benchmarking in area of AI for health but for the Malaria detection, there is no benchmarking done till now. The tools available have been trained on the data set that was made available to them mostly hospital based.

Items to be covered:

* papers on existing attempts to benchmark solutions on the topic
* clinical evaluation attempts, RCT, etc.

including existing numbers

Items to be covered:

* papers on existing attempts to benchmark solutions on the topic
* clinical evaluation attempts, RCT, etc.
* including existing numbers

## AI4H Topic Group: Current topic group and its mandate

The topic group (TG-Malaria) is specific and relevant to AI4H. The objectives are:

1. to provide a forum for open communication among various stakeholders,
2. to agree upon the benchmarking tasks of this topic and scoring metrics,
3. to facilitate the collection of high-quality labelled test data from different sources,
4. to clarify the input and output format of the test data,
5. to define and set-up the technical benchmarking infrastructure, and
6. to coordinate the benchmarking process in collaboration with the Focus Group management and working groups.

The primary output of a topic group is one document that describes all aspects of how to perform the benchmarking for this topic. (The document will be developed in a cooperative way by suggesting changes as input documents for the next FG-AI4H meeting that will then be discussed and integrated into an official output document of this meeting. The process will continue over several meetings until the topic description document is ready for performing the first benchmarking.)

### Topic group structure

TBC

The group has a Topic Group Driver who will moderate the activities of the Topic Group. Topic Groups summarize use cases of a certain health topic or problem and similar AI benchmarking requirements. However, inside a Topic Group different Sub-topic Groups can be established to pursue different topic-specific specializations. TG-Malaria will start without separate subtopic Groups. However, it is possible that during the process subtopics will be introduced as contributors will deem fit.

### Topic group participation

The topic group on “AI based detection of Malaria” currently involves an interdisciplinary network between the health team at Mulago Referral Hospital with expertise in Microscopy and the technical AI and 3D printing experts from the AI and Data Science research group at Makerere University.

The participation in both the focus and Topic Group is generally open and free of charge. To participate, one can follow the “call for participation” document outlining the process for joining the Focus Group and the Topic Group. For this topic, the corresponding call can be accessed through the official website (<https://itu.int/go/fgai4h>).

### Tools/process of TG cooperation~~.~~

* Group members will regularly interact via email, skype, LinkedIn or other short messenger systems. The topic group now has a group mail that is up and running. This was set up by the FG upon request by the Topic group ( fgai4htgmalaria@lists.itu.int).
* A monthly report will be provided by email by the Topic Group Driver to the group’s members.
* Regular group meetings will be held according to availability.

### TG interaction with WG and FG

As will be regulated by the FG.

### Current Status

Response to call for contribution to the TG-Malaria;

* Laura Moro, PhD. Researcher, science & medical writer. Co-founder of AI Scope. AI Scope a non-profit organization working in AI for improved diagnosis (mostly malaria for now) in low-resource settings.
* Dr. Helmi Zakariah. AIME inc.

Is a cofounder of AIME company and they work primarily in Forecasting Vector-Borne Disease Outbreak by using AI & ML. While their focus is in Dengue and West Nile Virus, they have begun work in Malaria through collaboration with APMEN members in Malaysia.

* Martha Shaka, a researcher at University of Dodoma and leads a team focusing on the automation of malaria diagnosis using deep learning. They are made up of 2 organization with medical and computer science experts. The team has collaboration with local researchers in the field of malaria diagnosis and their next step is on creating ground truth data sets in Tanzania.
* Phil Verstraete. Co-Managing Director, Milan & Associates
* Ana Rivière Cinnamond, Advisor and Pubic Health Expert in disease surveillance and prevention, DMAP under Health Emergency Information & Risk Assessment Department with [PAHO/WHO.](http://www.paho.org/hq/index.php?lang=en)
* Herilalaina RAKOTOARISON, PhD student in Machine learning from the Université Paris-Saclay).

**Progress in terms of collaboration for development of AI tool**

* FIND **(Foundation for Innovative New Diagnostics), these are a global non-profit organization driving innovation in the development and delivery of diagnostics to combat major diseases affecting the world’s poorest populations. Discussion**s on how best to collaborate are under way. Our key contacts at FIND are Rigveda Kadam, Senior Access Officer and Seda Yerlikaya, Scientific Officer.

### Next meetings

The Focus Groups meets about every two months at changing locations. The upcoming meetings are:

* I: May 2020

An up to date list can be found at the official [ITU FG AI4H website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).

### Next steps

We aim to extend the topic of Malaria detection to all Malaria endemic Countries, while bringing together AI solutions and data from different countries. Next steps for the group can be of different forms:

The group equally intends to undertake supervision of retraining and retooling of microscopists in endemic countries on AI based detection of Malaria.

The group further intends the creation of data centers for annotated data of thick blood smears from the different medical centers in endemic countries. This will help in creation of data repository center hence access to big data for further research.

 **Creation of public dataset repository data bank**.

To this end we propose acquisition of the first public dataset repository for multiple datasets such as image (both thick and thin blood smear images and other datasets such as Electronic Health Records, demographics, environmental). From this different AI based analyses for Malaria detection can be drawn from a large pool of data.

## Procedure.

We intend to seek expertise from laboratory technicians at different health centers to collect and annotate both thick and thin blood smear images and other data sets such as electronic health records, environment data. This data collected and annotated from the different health centers following the ethical procedures outlined above will then be posted to a central public repository data center.

This will then be replicated in other countries and creation of public repository data centers that will be managed by appointed data center managers.

Research Assistants, Data entrants, Data Center managers will be required in consultation with ITU/WHO management team responsible for this particular program.

**Sources of potential support and collaboration**.

The topic group will work hand in hand with ITU/WHO in lobbying support since the data center will be a public good. The group will seek participation, collaboration and financial support from potential sponsors to run its activities (for the start, an independent highly quality dataset is needed for the benchmarking process).

**Activities of the Topic group**

1. Collection of labelled data from different sources. Microscopic Image datasets (both from thin blood smear image and thick blood smear images) and any other data directly linked to malaria endemicity from sources like (environment, clinical records) that could improve the detection confidence is of high value. The ultimate goal is to make a strong malaria detection model given data from different systems and countries.
2. Selection and implementation of AI models and approaches related to malaria detection.
3. Suggestions on scoring metrics.
4. Contributing to the development of a viable and accepted benchmarking framework.
5. Support to the group on different aspects (data, methods, benchmarking, etc.) of this topic
6. Extension of the solution to improve disease surveillance and prediction.

**Next steps on the benchmarking platform for detection of malaria**

We have achieved a first attempt on the benchmarking platform for Malaria and promising results have been noted. As this gives us a positive headway, our next iterations will focus on improvement of the platform and as such is work in progress This will be in terms of:

* Finding a way to allow participants contribute data through the system.
* Allowing for multiple algorithms to be used. Apparently shallow learning models are expected. Deep learning that is implemented on modules like Tensorflow are not permissible.
* Adding more evaluation metrics that are acceptable for different AI models used.
* Once more and different datasets come in, the need to standardise it to uniform format in terms of image specification arises.
* A need to acquire secret test data
* Finding a flexible way that enables participants to choose their own development environment for their developing and executing their AI.

## Business case

TBC

Describe possible / relevant business case(s) applicable to this topic group.

# Method

The benchmarking method will consider all aspects of Input data requirements, how data will be annotated and annotation formats, AI analysis engine requirements, output and test data formats and scoring metric requirements.

Blood smear Images of both thick and thin blood smear slides that have been annotated by laboratory experts from different Health facilities in different Malaria endemic countries would be required and an undisclosed test data for evaluation of the tool.

The labels will depend on the specific attribute to be investigated.

All data will be subject to permissions from the different country authorities.

## AI input data structure

For our first attempt on prototyping a benchmarking platform, the TG has leveraged on the existing dataset available (1182 images of thick blood smear slides that have been annotated by laboratory experts from Mulago referral hospital.

To this end, only image data of thick blood smears of image format .jpg is sufficient to build malaria detection models. This is currently because we do not have any other dataset at hand. We believe that future iterations will allow multiple datasets (thin blood smear images, demographic data, environment data) to allow derivation of more accurate predictive models.

Our first benchmarking task is built in form of a codalab competition challenge in which we provide input dataset (thick blood smear images) to participants.

Image data together with corresponding labels (specifying presence or absence of malaria parasites) is provided. The TG envisions to attract machine learning experts who are particularly interested in automated malaria diagnosis to tune their models on the prototype dataset provided.

However, for a feasible and reliable solution, large amounts of data of both thick blood smear and thin blood smear images from different Health facilities in different malaria endemic countries would be required for machine learning models and an undisclosed test data for evaluation of the tool.

On the side of participants therefore, the input to our first benchmarking platform is a model to train on the available protype dataset available.

## AI output data structure

TBC

The AI output should include AI tools detecting malaria parasites, species, Test data labels

Confidentiality of gold standard testing data results would be maintained.

Items to be covered:

* outputs to benchmark
* ontologies, terminologies
* data format

## Test data labels

TBC

A label/ annotation will be given of the blood smear Image that contains the malaria parasites. The labels will depend upon the specific condition that is being benchmarked and also the type of AI task.

For our first iteration, a binary task is considered with positive (parasite) and negative (no parasite) patches from an image are used.

## Scores & metrics

TBC

To evaluate AI tool’s performance, labelled Dataset of blood smear images would be taken and tested against performance of AI. The algorithm evaluation mechanism should include metrics like ROC accuracy, precision, recall, specificity F1 scores, specificity, sensitivity, mean Average Precision (mAP), average precision and the choice will base on the algorithm used and purpose of the task.

## Undisclosed test data set collection

TBC

1. In order to assess algorithm robustness, sufficient undisclosed image data would be collected. This is envisioned to come from different health facilities both public and private.
2. There is need for examination of the quality of undisclosed dataset by a panel of experienced and skilled lab technicians. Bias in data will be considered.
3. An agreeable number of test data for a benchmarking task will be specified.
4. Annotation process of data with expert labels should also be agreed upon.

TBC

Items to be covered:

* raw data acquisition / acceptance
* test data source(s): availability, reliability,
* labelling process / acceptance
* bias documentation process
* quality control mechanisms
* discussion of the necessary size of the test data set for relevant benchmarking results
* specific data governance derived by general data governance document (currently F-103)

## Benchmarking methodology and architecture

Technical architecture

The general benchmarking architecture (see figure 1) advanced by FGAI4H [8] will provide guidance to TG Malaria in the development of the benchmarking platform.

 

Figure 1 –General Benchmarking pipeline framework for implementation of AI based health solution [8 ].

To implement our first benchmarking task on detection of malaria in thick blood smear images, the benchmarking platform used is Codalab. It is an open source framework designed for enhancing reproducibility of machine learning algorithms. We adopt this for benchmarking malaria detection modeling (see figure 2). Also available at;

<https://codalab.lri.fr/competitions/748#learn_the_details>



 Figure 2 –Benchmarking-Malaria platform implemented using Codalab

The overall process of benchmarking is handled on the server side. Codalab allows organizers to define;

1. **How submission files will be handled, processed and scored;**

The benchmarking system in its current state has a prototype dataset stored at the site of the benchmarking system. Participants are required to use the available dataset send in their AI model by fine tuning a variant sample code on the leaderboard. A submission fails once it doesn’t meet the submission criteria defined. At the organizers’ site(s), derived detection accuracies of different models are shown (see figure 3).



Figure 3 – Derived detection accuracies of different models.

**2) In which environment (programming language, time constraint, memory constraint) are submission files run?**

For our first prototype, participants will need to set up their local environment by following the prerequisites below;

Install Anaconda Python 3.6.6, opencv-python (4.0.1), scikit-image (0.15.0). Download the starting kit. Usage: - modify sample\_code\_submission/model.py to provide a better model - zip the contents of sample\_code\_submission (without the directory, but with metadata)

The utility of Codalab is then to;

1) get submitted algorithm

2) score algorithm with predefined metric and environment constraint

3) update leaderboard.

* hosting (IIC, etc.)

Since Codalab is an open source framework, it can be deployed to any server. In the early stage of this project, we will use codalab.lri.fr (server maintained by Paris-Sud University) for testing and prototyping.

* possibility of an online benchmarking on a public test dataset

At the moment, the platform does not allow public users to submit their own dataset to the benchmark. Otherwise, they can contact platform maintainers (TG-Malaria) to do so. In its current state, the platform has a sample public dataset for participants to prototype their ML solutions.

* protocol for performing the benchmarking (who does what when etc.)

At the moment, minimal benchmarking system is created by the organizers for prototyping. The latter system will process submitted files in Python 3 environment with time budget of 10min. TG-Malaria benchmarking Organizers made available a starting kit (sample code submission) to ease the task of participants.

Participants on their side need to adapt their algorithm to fit the structure of the starting kit.

The system allows a participant to submit up to 100 times but only 5 times in a day. This will enable each participant to fine tune their detection models.

* AI submission procedure including contracts, rights, IP etc. considerations

Copyright of submitted source code will remain to the participants. Codalab allow participants to decide whether they want to make submissions publicly available or not.

## Reporting methodology

TBC

Reporting would be based on the accuracy of an AI tool’s ability to detect the presence of malaria parasites,

* Public benchmarking leaderboard developed
* Making publication of the deliverables of the TG.

Items to be covered:

* Report publication in papers or as part of ITU documents
* Online reporting
* public leaderboards vs. private leaderboards
* Credit-Check like on approved sharing with selected stakeholders
* Report structure including an example
* Frequency of benchmarking
* Subtopic 1
	+ Among the aforementioned aspects of Reporting Methodology, indicate the elements that are applicable to this subtopic.
* Subtopic 2

# Results

Results are based on agreed upon evaluation accuracy metrics for an AI tool’s ability to detect the presence of malaria parasites.

For the first version of the minimal benchmarking system, TG-Malaria hinged on the following evaluation metrics; ROC AUC, precision, recall and Average precision.

 **Evaluation report for each AI solution submitted.**

The benchmarking platform computes the evaluation metrics and scores based on the public available dataset and the AI models used. Results of the different AI models in terms of evaluation metrics are finally shown on Codalab leaderboard. The system is time stamped and keeps track each time a participant submits a new entry.

Preliminary results of our prototype benchmarking project are as shown in the Figure below;



Figure 4 – Result report for the AI models submitted

# Discussion

TBC

Discussion of the insights from executing the benchmarking on

* external feedback on the whole topic and its benchmarking
* technical architecture
* data acquisition
* benchmarking process
* benchmarking results
* field implementation success stories

# Declaration of Conflict of Interest

TBC

References

[1] I. Bates, V. Bekoe, and A. Asamoa-Adu. Improving the accuracy of malaria-related laboratory tests in Ghana. Malar. J. Vol.3. No. 38, 2004.

[2] C.A. Petti, C.R. Polage, T.C. Quinn, A.R. Ronald, and M.A. Sande. Laboratory medicine in Africa: A barrier to effective health care. Clinical Infectious Diseases, Vol. 42, N0. 3, PP. 377-382, 2006.

[3] L. Rosado, J.M. Correia da Costa, D. Elias, and J.S. Cardoso. A review of automatic malaria parasites detection and segmentation in microscopic images. Anti-Infective Agents, Vol 14, No. 1, pp 11-22, 2016.

[4] J. A. Quinn, R. Nakasi, P. K. B. Mugagga, P. Byanyima, W. Lubega, and A. Andama. Deep convolutional neural networks for microscopy-based point of care diagnosis. In proceedings of International Conference on Machine Learning for Health Care, Volume 50., 2016.

[5] WHO. World malaria report. Geneva, Switzerland, World Health Organization., 2016.

[6] WHO. World malaria report. Geneva, Switzerland, World Health Organization., 2015.

[7] WHO. World malaria report. Geneva, Switzerland, World Health Organization., 2017.

[8] M. Salathé, T. Wiegand, M. Wenz. Focus Group on Artificial Intelligence for Health. Available at; <https://arxiv.org/ftp/arxiv/papers/1809/1809.04797.pdf>

[9] D. Bhattacharya. Relevance of economic field microscope in remote rural regions for concurrent observation of malaria and inflammation. Advances In Infectious Diseases (AID), Vol.2. No.1, PP. 13-18., 2012.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_