|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2022-2024 | | FG-AI4H-S-014-A01 | |
| ITU-T Focus Group on AI for Health | |
| Original: English | |
| **WG(s):** | | Plenary | Geneva, 3-5 July 2023 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Malaria Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Malaria) [same as Meeting N] | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Rose Nakasi  Makerere University Uganda | | E-mail: [g.nakasirose@gmail.com](mailto:g.nakasirose@rki.de) |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for AI-based Malaria. It covers all scientific, technical, and administrative aspects relevant for setting up this benchmarking (and follows the template structure defined in document FGAI4H-J-105). The creation of this TDD is an ongoing iterative process until it is approved by the Focus Group on AI for Health (FG-AI4H) as deliverable No. 014. This draft will be a continuous input- and output document. |

Contributors

|  |  |
| --- | --- |
| Herilalaina Rokototarison Université Paris-Saclay France | Email: [heri@lri.fr](mailto:heri@lri.fr) |
| Fetulhak Abdurahman  Jimma University | Email: [afetulhak@yahoo.com](mailto:afetulhak@yahoo.com) |

**CONTENTS**

| Page |
| --- |
| [1 Introduction 4](#_Toc95406788)  [2 About the FG-AI4H topic group on TG-MALARIA 4](#_Toc95406789)  [2.1 Status of this topic group 5](#_Toc95406790)  [2.1.1 Status update for meeting [Discussions arising out of e-meetings] 5](#_Toc95406791)  [2.1.2 Status Update [Members] 5](#_Toc95406792)  [2.1.3 Status Update [Next steps] 6](#_Toc95406793)  [2.2 Topic group participation 6](#_Toc95406794)  [3 Topic description 7](#_Toc95406795)  [3.1 Subtopic [AI based detection of malaria] 8](#_Toc95406796)  [3.1.1 Definition of the AI task 8](#_Toc95406797)  [3.1.2 Current gold standard 8](#_Toc95406798)  [3.1.3 Relevance and impact of an AI solution 8](#_Toc95406799)  [3.1.4 Existing AI solutions 9](#_Toc95406800)  [3.2 Subtopic [B] 9](#_Toc95406801)  [4 Ethical considerations 9](#_Toc95406802)  [5 Existing work on benchmarking 10](#_Toc95406803)  [5.1 Subtopic [AI based detection of malaria] 10](#_Toc95406804)  [5.1.1 Publications on benchmarking systems 10](#_Toc95406805)  [5.1.2 Benchmarking by AI developers 11](#_Toc95406806)  [5.1.3 Relevant existing benchmarking frameworks 11](#_Toc95406807)  [5.2 Subtopic [B] 11](#_Toc95406808)  [6 Benchmarking by the topic group 11](#_Toc95406809)  [6.1 Subtopic [AI based detection of malaria] 12](#_Toc95406810)  [6.1.1 Benchmarking version [2] 12](#_Toc95406811)  [6.2 Updates on the benchmarking platform V4 for malaria detection 13](#_Toc95406812)  [6.3 Benchmarking V2 platform for malaria detection 14](#_Toc95406813)  [6.3.2 Some recent publications 21](#_Toc95406814)  [6.3.3 Benchmarking version [1] 24](#_Toc95406815)  [6.4 Subtopic [B] 27](#_Toc95406816)  [7 Overall discussion of the benchmarking 27](#_Toc95406817)  [8 Regulatory considerations 27](#_Toc95406818)  [8.1 Existing applicable regulatory frameworks 28](#_Toc95406819)  [8.2 Regulatory features to be reported by benchmarking participants 28](#_Toc95406820)  [8.3 Regulatory requirements for the benchmarking systems 28](#_Toc95406821)  [8.4 Regulatory approach for the topic group 29](#_Toc95406822)  [9 References 30](#_Toc95406823)  [Annex A: Glossary 31](#_Toc95406824)  [Annex B: Declaration of conflict of interests 32](#_Toc95406825) |

**List of Tables**

[Table 1: Topic group output documents 5](#_Toc95406826)

**List of Figures**

[Figure 1: Updated user Interface for the benchmarking platform for malaria detection 14](#_Toc95406827)

[Figure 2: Update for data upload 15](#_Toc95406828)

[Figure 3: Result report for the AI models submitted 23](#_Toc95406829)

[Figure 4: General Benchmarking pipeline framework for implementation of AI based health solution [8 ]. 25](#_Toc95406830)

[Figure 5: Benchmarking-Malaria platform implemented using Codalab 25](#_Toc95406831)

[Figure 6: Derived detection accuracies of different models. 26](#_Toc95406832)

FG-AI4H Topic Description Document

Topic group-Malaria

# Introduction

Malaria is one of the largest endemic diseases in 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. Recently AI and machine learning techniques have been successful in different medical image analysis tasks and have a capability to improve public health.

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

This topic description document specifies the standardized benchmarking for Topic Group-Malaria systems. It serves as deliverable No.014 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

# About the FG-AI4H topic group on TG-MALARIA

The introduction highlights the potential of a standardized benchmarking of AI systems for Topic Group Malaria to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Malaria at the meeting H in Zanzibar, 3-5 September 2019.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD.During FG-AI4H meeting H in Zanzibar, 3-5 September 2019, Rose Nakasi from Makerere University was nominated as topic driver for the TG-Malaria.

Documentation

This document is the TDD for the TG-Malaria. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for TG-Malaria. It describes the existing approaches for assessing the quality of TG-Malaria systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group’s benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable “DEL 014. AI based detection of Malaria (TG-Malaria).” The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

The final version of this TDD will be released as deliverable “DEL 014. AI based detection of Malaria (TG-Malaria).” The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic group output documents

| Number | Title |
| --- | --- |
| FGAI4H-O-014-A01 | Latest update of the Topic Description Document of the TG-Malaria |
| FGAI4H-O-014-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| FGAI4H-O-014-A03 | The presentation summarizing the latest update of the Topic Description Document of the TG-Malaria |

The working version of this document can be found in the official topic group SharePoint directory.

Select the following link:

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-L-014-A01.docx>

## Status of this topic group

The following subsections describe the update of the collaboration within the TG-Malaria for the official focus group meetings.

### Status update for meeting [Discussions arising out of e-meetings]

* Discuss updates on benchmarking platform improvements (data, AI models, Interface)
* Discuss technical implementation details that come with improvements
* Develop simple models for testing the updated benchmarking platform
* Platform beta testing
* \*Launching the challenge

### Status Update [Members]

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 of Machine learning from the Université Paris-Saclay). Herilalaina has been pivotal in implementing the benchmarking platform.
* Fetulhak Abdurahman, is a Lecturer in Jimma University of Electrical and Computer Engineering Faculty, Ethiopia.
* William Mangion, Healthcare and Robotics Computer vision consultant, V7 Labs, London.

### Status Update [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 to seek for 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.

-The group also intends to develop robust, reliable and low-cost AI solutions which can be deployed in a real-time environment. To this end, the group aims at providing Malaria parasite detection AI algorithms that can represent the real-time clinical setting. Our platform enables users to access benchmarking solutions and can integrate into their organization.

-The group will also participate at conferences to present the progress of our achievements and publish results of challenge-based benchmarking in reputable conferences and journals.

- Launch of challenge to a wider community.

-Iterate with the Focus Group benchmarking platform.

## Topic group participation

The participation in both, the Focus Group on AI for Health and in a TG is generally open to anyone (with a free ITU account). For this TG, the corresponding ‘Call for TG participation’ (CfTGP) can be found here:

* <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Malaria.pdf>

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Malaria.aspx>

For participation in this topic group, interested parties can also join the regular online meetings. For all TGs, the link will be the standard ITU-TG ‘zoom’ link:

* <https://itu.zoom.us/my/fgai4h>

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list [fgai4h@lists.itu.int](mailto:fgai4h@lists.itu.int).

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the ‘Call for Topic Group participation’ and this link:

* <https://itu.int/go/fgai4h/join>

In addition to the general FG-AI4H mailing list, each topic group can create an *individual mailing list:*

* [fgai4htgmalaria@lists.itu.int](mailto:fgai4htgmalaria@lists.itu.int)

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

* <https://itu.int/go/fgai4h>

# Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in TG Malaria and how this can help to solve a relevant ‘real-world’ problem.

Topic groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-Malaria currently has no subtopics. Future subtopics for AI based Malaria Surveillance might be introduced.

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 the 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 with patient care and treatment.

## Subtopic [AI based detection of malaria]

### Definition of the AI task

This section provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to [DEL03](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7997F2C1-5A1D-4409-B2A0-CBC4E9CE8CDA%7D&file=DEL03.docx&action=default) *“AI requirements specifications*,” which describes the functional, behavioural, and operational aspects of an AI system.

A use case on Artificial Intelligence-based Diagnosis of Malaria is presented here. Malaria is one of the major diseases causing death in sub Saharan Africa according to WHO report. Part of the reason for endemicity is poor diagnosis at the laboratory level which may lead to misdiagnosis of the disease as well as drug resistance. The burden is further increased because of lack of enough skilled lab technologists in health facilities to diagnose the disease through the gold standard method of conventional microscopy. However conventional microscopy is subjective and results vary significantly by different Microscopists thus inaccurate and low throughput screening. Therefore, timely and accurate diagnostic interventions are necessary to reduce cases of misdiagnosis, drug residence burden.

Advances in technology help to push forward in the provision of health care facilities in the form of automated diagnosis of diseases, telemedicine, 3D-printing of medical devices, and mobile health. AI-based Detection of Malaria therefore focuses on use of artificial intelligence techniques to detect plasmodium pathogens in blood smear images in a timely and more accurate manner. Here we propose machine learning methods that deal with all aspects related to improving the conventional malaria diagnosis on blood films. Machine learning methodologies learn good representations of data directly from the pixel data thus providing a more reliable, fast and accurate diagnosis helping to provide confidence of a diagnosis to the lab technicians.

### Current gold standard

This section provides a description of the established gold standard of the addressed health topic.

Conventional light microscopy remains the gold standard method of diagnosis of malaria. Microscopy is particularly well adapted to low-resource, high disease burden areas, being both simple and versatile. In contrast to alternatives such as rapid diagnostic tests, however, microscopy-based diagnosis does depend on the availability of skilled technicians, of which there is a critical shortage. As a result, diagnoses are often made on the basis of clinical signs and symptoms alone, which is error-prone and leads to higher mortality, drug resistance, and the economic burden of buying unnecessary drugs. There is therefore a need for alternatives which help to provide access to fast and quality diagnosis.

### Relevance and impact of an AI solution

This section addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.

AI based diagnosis of Malaria aims to reduce the need for many human Microscopists by providing a consistent and accurate diagnosis with minimum human intervention. This is because AI algorithms can accurately learn a good representation of data directly from the annotated datasets. Automation presents a significant advantage over a human Microscopists by potentially increasing the speed and accuracy for blood film analysis, reducing the turnaround time, and significantly improving diagnostic performance.

Currently malaria diagnosis is by use of symptoms and signs, Rapid Diagnostic Tests (RDTs) and conventional microscopy which methods are prone to human error, slow and lack specificity details and thus accuracy is based on human judgment which is usually biased. Therefore, the need to benchmark AI algorithms for malaria diagnosis. Automated AI based microscopy for malaria diagnosis maintains the benefits of manual microscopy (gold standard) by incorporating them in a machine vision platform which helps to provide the access to fast and quality diagnosis that is currently routinely unavailable.

In principle, new benchmarking AI Methods should focus on providing a robust, fast, low cost and more accurate malaria diagnosis that reduces biases through capturing all the necessary implementation parameters that provide a good representation of a dataset, and implementation platform.

The added advantage is that the solution can be applied to any microscopical assessment and in different implementation environments.

### Existing AI solutions

This section provides an overview of existing AI solutions for the same health topic that are already in operation. It should contain details of the operations, limitations, robustness, and the scope of the available AI solutions. The details on performance and existing benchmarking procedures will be covered in chapter 6.

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 thus 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.

## Subtopic [B]

**TBC**

# Ethical considerations

**TBC**

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable [DEL01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default) “*AI4H ethics considerations,”* which was developed by the working group on “Ethical considerations on AI4H” (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-Malaria.

* What are the ethical implications of applying the AI model in real-world scenarios?
* What are the ethical implications of introducing benchmarking (having the benchmarking in place itself has some ethical risks; e.g., if the test data are not representative for a use case, the data might create the illusion of safety and put people at risk)?
* What are the ethical implications of collecting the data for benchmarking (e.g., how is misuse of data addressed, is there the need for an ethics board approval for clinical data, is there the need for consent management for sharing patient data, and what are the considerations about data ownership/data custodianship)?
* What risks face individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground?
* How is the privacy of personal health information protected (e.g., in light of longer data retention for documentation, data deletion requests from users, and the need for an informed consent of the patients to use data)?
* How is ensured that benchmarking data are representative and that an AI offers the same performance and fairness (e.g., can the same performance in high, low-, and middle-income countries be guaranteed; are differences in race, sex, and minority ethnic populations captured; are considerations about biases, when implementing the same AI application in a different context included; is there a review and clearance of ‘inclusion and exclusion criteria’ for test data)?
* What are your experiences and learnings from addressing ethics in your TG?

# Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and TG-Malaria for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

## Subtopic [AI based detection of malaria]

### Publications on benchmarking systems

While a representative comparable benchmarking for TG-Malaria does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL07 *“AI for health evaluation considerations,”*[DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“AI4H evaluation process description,”* [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification*,*”* [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM),”* and [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default) *“Clinical Evaluation of AI for health”*.

* What is the most relevant peer-reviewed scientific publications on benchmarking or objectively measuring the performance of systems in your topic?
* State what are the most relevant approaches used in literature?
* Which scores and metrics have been used?
* How were test data collected?
* How did the AI system perform and how did it compare the current gold standard? Is the performance of the AI system equal across less represented groups? Can it be compared to other systems with a similar benchmarking performance and the same clinically meaningful endpoint (addressing comparative efficacy)?
* How can the utility of the AI system be evaluated in a real-life clinical environment (also considering specific requirements, e.g., in a low- and middle-income country setting)?
* Have there been clinical evaluation attempts (e.g., internal and external validation processes) and considerations about the use in trial settings?
* What are the most relevant gaps in the literature (what is missing concerning AI benchmarking)?

### Benchmarking by AI developers

TBC

All developers of AI solutions for TG-Malaria implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

* What are the most relevant learnings from the benchmarking by AI developers in this field (e.g., ask the members of your topic group what they want to share on their benchmarking experiences)?
* Which scores and metrics have been used?
* How did they approach the acquisition of test data?

### Relevant existing benchmarking frameworks

TBC

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“FG-AI4H assessment platform”* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

* Which benchmarking platforms could be used for this topic group (e.g., EvalAI, AIcrowd, Kaggle, and CodaLab)?
* Are the benchmarking assessment platforms discussed, used, or endorsed by FG-AI4H an option?
* Are there important features in this topic group that require special attention?
* Is the reporting flexible enough to answer the questions stakeholders want to get answered by the benchmarking?
* What are the relative advantages and disadvantages of these diverse solutions?

## Subtopic [B]

TBC

# Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the TG-Malaria AI task including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default) *“Data specification”* (introduction to deliverables 5.1-5.6), [DEL05\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B19830259-F63B-42D4-A408-48C854D6C124%7D&file=DEL05_1.docx&action=default)*“Data requirements”* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), [DEL05\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B25141F77-E59A-45F1-B081-185C2194FE67%7D&file=DEL05_2.docx&action=default) *“Data acquisition”*, [DEL05\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B05D8938E-BC2A-4A62-BCB0-1FD46AA72235%7D&file=DEL05_3.docx&action=default) *“Data annotation specification”*, [DEL05\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF267A95C-4C5B-4D63-A135-58AF487C3AD3%7D&file=DEL05_4.docx&action=default) *“Training and test data specification”* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) *“Data handling”* (which outlines how data will be handled once they are accepted), [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) *“Data sharing practices”* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), [DEL06](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF5967277-90C8-4252-A0B9-43A5692F35E2%7D&file=DEL06.docx&action=default) *“AI training best practices specification”* (which reviews best practices for proper AI model training and guidelines for model reporting), [DEL07](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B47E77197-F87B-49F4-80B3-2DD949A5F185%7D&file=DEL07.docx&action=default)*“AI for health evaluation considerations”* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), [DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“AI4H evaluation process description”* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification”* (which specifies how an AI can and should be tested *in silico*), [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM)”* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default)*“Clinical Evaluation of AI for health”* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“FG-AI4H assessment platform”* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), [DEL09](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3E940987-8D75-44B8-85E4-F0E475964F15%7D&file=DEL09.docx&action=default) *“AI for health applications and platforms”* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), [DEL09\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1A2EC8D5-53CA-4C8C-9B09-B61CA6F428C5%7D&file=DEL09_1.docx&action=default) *“Mobile based AI applications,”* and [DEL09\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3B5A31DE-D3B1-4EC1-A261-2C2E19F73810%7D&file=DEL09_2.docx&action=default) *“Cloud-based AI applications”* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

## Subtopic [AI based detection of malaria]

The benchmarking of TG-Malaria is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent sections, where the actual benchmarking methodology for each version will be described.

* Which benchmarking iterations have been implemented thus far?
* What important new features are introduced with each iteration?
* What are the next planned iterations and which features are they going to add?

### Benchmarking version [2]

This section includes all technological and operational details of the benchmarking process for the latest benchmarking version [2].

#### Overview

This section provides an overview of the key aspects of this benchmarking iteration, version [2].

* What is the overall scope of this benchmarking iteration (e.g., performing a first benchmarking, adding benchmarking for multi-morbidity, or introducing synthetic-data-based robustness scoring)?
* What features have been added to the benchmarking in this iteration?

#### Benchmarking methods

This section provides details about the methods of the benchmarking version 2. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

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.

##### Benchmarking system architecture

This section covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required.

## Updates on the benchmarking platform V4 for malaria detection

The initial version of the competition has been migrated to Codabench (https://www.codabench.org) - a successor version of Codalab focusing on benchmarking instead of competition. This choice is motivated for several reasons. Firstly, the improvement of Codabench over Codalab eases benchmark management tasks such as incorporating new metrics in the leaderboard or proposing new problems. Secondly, the functionality enabling participants to submit a new dataset is straightforward to implement in Codabench. And lastly, participants are given the flexibility to choose the set of datasets on which their submission will be benchmarked on. The figure below depicts the submission page on Codabench.

Note that our benchmarking is not published yet as it is still under development; however, it can be accessible through the link: <https://www.codabench.org/competitions/389/?secret_key=2f16646d-f5db-40c1-ae57-6a89c6802551>

Graphical user interface, text, application, email

Description automatically generated

The next milestone is to incorporate object detection tasks to extend the available classification tasks. Since the processing pipeline for object detection is crucial, we plan to start with the VIPriors challenge template, which provides an end-to-end training of the model. The latter template thus enables participants to focus their effort on the model design part. However, some difficulties arise, especially in adapting the Microscopy dataset format into the VIPriors template and integrating the overall pipeline into Codabench. Therefore, future works will focus on addressing these issues.

## Benchmarking V2 platform for malaria detection

Building from the first version 1 of the benchmarking platform, the new updated version 2 of the platform for benchmarking malaria detection (available at:<https://codalab.lri.fr/competitions/775>) with improvements is ready for an alpha test phase. To this end, a call for participation was drafted in our previous CfTGP document we are sought for active participation from people with background not only in computer vision, machine learning and artificial intelligence, but also data submission from microscopists to take part in our malaria detection challenge.

Some of the changes and updates are highlighted in the benchmarking Interface as shown in the Figure below.

Graphical user interface, text, application

Description automatically generated

*Figure 1: Updated user Interface for the benchmarking platform for malaria detection*

The updates to the platform that have been affected are summarised below;

1. Adding support for uploading datasets;

Unlike the first version of the platform which never had provision for data submission, Participants are now able to submit their own datasets through the “Upload dataset” to enrich current benchmark datasets (see figure 2). Submission contains several files: dataset information (name of features, target variable, copyright), dataset file (in Codalab format) and train-test split indices. To ensure data quality, the platform verifies uploaded dataset and rejects incorrect submission. The latter validation script is available for download by participants (reference is given in the platform). Note that at the moment, the uploading module only works for classification tasks.

Graphical user interface, text, application, email

Description automatically generated

*Figure 2: Update for data upload*

2. Add New public dataset of thin blood smear dataset to the benchmark

A second aspect that was updated on the benchmarking platform is the provision for a new dataset. factors. In the initial version of the platform, only thick blood smear dataset was used. For test purposes, another dataset that comprises both infected and uninfected cell images obtained from [13] is added to have multiple tasks on the platform. The latter dataset is a classification task.

3. Adding support for a deep learning library (pytorch and tensorflow) and setting up time budget for 1 hour/submission.

One of the limitations for the implementation of the initial benchmarking platform was on the implementation environment and time which allowed submission for only traditional machine learning models. With the update, comes an improvement with support for deep learning libraries and a submission time budget of up to 1 hour to enable bulky models.

##### Benchmarking system dataflow

TBC

This section describes the dataflow throughout the benchmarking architecture. The benchmarking platform has three components: data storage server, compute worker and frontend web application. Participants / organizers interact via the web application. It includes uploading data, configuring competition, manager list of participants, submitting AI models. On Codalab, data formats are flexible and only depend on the ingestion program. We may recall that the overall processing step is executed by the ingestion program, termed IP. In our competition, we stored data in libsvm format since it is the default format of Codalab datasets. Versioning of datasets are also handled by Codalab. Submissions (AI models) are processed as follows: first, they are handled by the ingestion program for an initial verification of its format and create the AI model (Python object). Then, the ingestion program reads data from storage (according to the path specified in the configuration files). The latter extracted dataset is fed to the AI model for training. The ingestion program monitors memory consumption and time budget during the training. If successful, IP fetches predictions and calls the script to compute the metrics. In our challenge, we computed predictive accuracy, ROC-AUC and Recall. Finally, IP pushed the scores on the leaderboard.

##### Safe and secure system operation and hosting

Dataset can be stored on Codalab or within another cloud server. Since the initial version of our challenge is hosted at the official platform Codalab server (codalab.org), we also opt to store our data on the same server. Codalab.org also hosts many AI challenges with well-known machine learning conferences suggesting that the platform has some level of security. All the security aspects of data and source code are managed by the platform and the administrators of Codalab.org.

##### Benchmarking process

TBC

This section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

* How are new benchmarking iterations scheduled (e.g., on demand or quarterly)?
* How do possible participants learn about an upcoming benchmarking?
* How can one apply for participation?
* What information and metadata do participants have to provide (e.g., AI autonomy level assignment (IMDRF), certifications, AI/machine learning technology used, company size, company location)?
* Are there any contracts or legal documents to be signed?
* Are there inclusion or exclusion criteria to be considered?
* How do participants learn about the interface they will implement for the benchmarking (e.g., input and output format specification and application program interface endpoint specification)?
* How can participants test their interface (e.g., is there a test dataset in case of file-based offline benchmarking or are there tools for dry runs with synthetic data cloud-hosted application program interface endpoints)?
* Who is going to execute the benchmarking and how is it ensured that there are no conflicts of interest?
* If there are problems with an AI, how are problems resolved (e.g., are participants informed offline that their AI fails to allow them to update their AI until it works? Or, for online benchmarking, is the benchmarking paused? Are there timeouts?)?
* How and when will the results be published (e.g., always or anonymized unless there is consent)? With or without seeing the results first? Is there an interactive drill-down tool or a static leader board? Is there a mechanism to only share the results with stakeholders approved by the AI provider as in a credit check scenario?
* In case of online benchmarking, are the benchmarking data published after the benchmarking? Is there a mechanism for collecting feedback or complaints about the data? Is there a mechanism of how the results are updated if an error was found in the benchmarking data?

#### AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of TG-Malaria. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking.

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 the 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

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. It covers the details of the data format, coding, and error handling at the level of detail needed for an AI to participate in the benchmarking.

* What are the general data output types returned by the AI and what is the nature of the output (e.g., classification, detection, segmentation, or prediction)?
  + How exactly are they encoded? Discuss points like:
    - The exact data format with all fields and metadata (including examples or links to examples)
    - Ontologies and terminologies
* What types of errors should the AI generate if something is defective?

#### Test data label/annotation structure

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called ‘labels’) for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following section.

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 and metrics

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics used to measure the performance, robustness, and general characteristics of the submitted AI systems.

To evaluate AI tool’s performance, labelled Dataset of blood smear images would be taken and tested against the 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 be based on the algorithm used and purpose of the task.

#### Test dataset acquisition

TBC

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure including quality control of the dataset, control mechanisms, data sources, and storage.

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.

* How does the overall dataset acquisition and annotation process look?
* How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
* Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
* How was the dataset documented and which metadata were collected?
  + Where were the data acquired?
  + Were they collected in an ethical-conform way?
  + Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
  + Do the data contain ‘sensitive information’ (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
  + What kind of data anonymization or deidentification has been applied?
  + Are the data self-contained (i.e., independent from externally linked datasets)?
  + How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?
  + What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?
* Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
* Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
* How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
* Specific data governance derived by the general data governance document (currently [F-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-F-103-DataPolicy.pdf) and the deliverables beginning with [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default))
* How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
* Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
* How was the annotation organized?
  + How many annotators/peer reviewers were engaged?
  + Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
  + How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
  + If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
  + How was the annotation of each case documented?
  + Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
* Were data/label update/amendment policies and/or criteria in place?
* How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
* How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?
* Is there assurance that the test dataset is undisclosed and was never previously used for training or testing of any AI model?
* What mechanisms are in place to ensure that test datasets are used only once for benchmarking? (Each benchmarking session will need to run with a new and previously undisclosed test dataset to ensure fairness and no data leakage to subsequent sessions)

#### Data sharing policies

TBC

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) on *data handling* and [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) on *data sharing practices*).

* Which legal framework was used for data sharing?
* Was a data sharing contract signed and what was the content? Did it contain:
  + Purpose and intended use of data
  + Period of agreement
  + Description of data
  + Metadata registry
  + Data harmonization
  + Data update procedure
  + Data sharing scenarios
    - Data can be shared in public repositories
    - Data are stored in local private databases (e.g., hospitals)
  + Rules and regulation for patients’ consent
  + Data anonymization and de-identification procedure
  + Roles and responsibilities
    - Data provider
    - Data protection officer
    - Data controllers
    - Data processors
    - Data receivers
* Which legal framework was used for sharing the AI?
* Was a contract signed and what was the content?

#### Baseline acquisition

TBC

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

* Does this topic require comparison of the AI model with a baseline (gold standard) so that stakeholders can make decisions?
* Is the baseline known for all relevant application contexts (e.g., region, subtask, sex, age group, and ethnicity)?
* Was a baseline assessed as part of the benchmarking?
* How was the process of collecting the baseline organized? If the data acquisition process was also used to assess the baseline, please describe additions made to the process described in the previous section.
* What are the actual numbers (e.g., for the performance of the different types of health workers doing the task)?

#### Reporting methodology

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

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.

### Some recent publications

With the growing interest in research around automated detection of malaria, some research has been conducted around improvement of malaria detection using AI with respect to assessing of data quality and use of new models and platforms for detection of malaria. Some publications are discussed below;

1. An approach for assessing quality of labelled Data for a machine learning task in Malaria detection [12].

While microscopy diagnosis through supervised learning for image analysis notably contributes to malaria detection, it has limitations. Among its principal challenges is the manual and tiresome process of data annotation for the classification task. The manual annotation of data is prone to inaccuracy defects due to bias, subjectivity and unclear images resulting in many false positives. This is normally due to personal independent judgements that vary from individual microscopists hence summatively affecting the accuracy of the model. In this study, we sought to investigate the possibility of classifying the negative far examples and the positive near examples from the positives in thick blood smear images for malaria detection. Assessing the classification performance could potentially inform us of the quality of training dataset and guide on selecting the best training dataset for a malaria parasite detection task. We employed the Mean Squared Error (MSE) to distinguish between positive and negative images. We later investigate the performance of the VGG-16 classification model based on how close or far negative examples are from positives. Experimental results showed that negative examples far from the positives produce better results than those near and that the proposed method could potentially be used to reduce false positives and bias in the training data.

2. A new approach for microscopic diagnosis of malaria parasites in thick blood smears using pre-trained deep learning models [11].

This research was motivated by the emerging technologies of machine learning that can learn complex image patterns and have accelerated research in medical image analysis. In this study, on a dataset of thick blood smear images, we evaluate and compare performance of three pre-trained deep learning architectures namely; faster regional convolutional neural network (faster R-CNN), Single-Shot multibox Detector (SSD) and RetinaNet through a Tensorflow object detection API. Data augmentation method was applied to optimise performance of the meta architectures. The possibility for mobile phone detector deployment was also investigated. The results revealed that faster R-CNN was the best trained model with a mean average precision of over 0.94 and SSD, was the best model for mobile deployment. We therefore deduce that faster R-CNN is best suited for obtaining high rates of accuracy in malaria detection while SDD is best suited for mobile deployment.

3. A web-based intelligence platform for diagnosis of malaria in thick blood smear images: A case for a developing Country [10].

The study was motivated by the need for development of remote systems that can provide fast, accurate and timely diagnosis of Malaria. With availability of internet, mobile phones and computers, rapid dissemination and timely reporting of medical image analytics is possible. This research aimed at developing and implementing an automated web-based Malaria diagnostic system for thick blood smear images under light microscopy to identify parasites. We implemented an image processing algorithm based on a pre-trained model of Faster Convolutional Neural Network (Faster R-CNN) and then integrated it with web-based technology to allow easy and convenient online identification of parasites by medical practitioners. The developed system holds the potential to improve the efficiency and accuracy in malaria diagnosis, especially in remote areas of developing countries that lack adequate skilled labour.

4. Mobile-Aware Deep Learning Algorithms for Malaria Parasites and White Blood Cells Localization in

Thick Blood Smears [14].

The research was motivated by the need for effective determination of malaria parasitemia is paramount in aiding clinicians to accurately estimate the severity of malaria and guide the response for quality treatment. This study presents an end-to-end deep learning approach to automate the localization and count of P.falciparum parasites and White Blood Cells (WBCs) for effective parasitemia determination. The method involved building computer vision models on a dataset of annotated thick blood smear images. These computer vision models were built based on pre-trained deep learning models including Faster Regional Convolutional Neural Network (Faster R-CNN) and Single Shot multibox Detector (SSD) models that help process the obtained digital images. A mobile smartphone-based inference app to detect malaria parasites and WBCs in situ was developed. The proposed method can be applied to support malaria diagnostics in settings with few trained Microscopy Experts yet constrained with large volumes of patients to diagnose.

#### Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

Results are based on agreed upon evaluation metrics for an AI tool’s ability to detect the presence of malaria parasites. We have so far developed our first and second version of the TG benchmarking system prototypes, TG-Malaria hinged on the following evaluation metrics;

ROC AUC, precision, recall, Average precision.

**Evaluation report for each AI solution submitted for in our trial versions.**

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;

Graphical user interface, application

Description automatically generated

*Figure 3: Result report for the AI models submitted*

#### Discussion of the benchmarking

TBC

This section discusses insights of this benchmarking iterations and provides details about the ‘outcome’ of the benchmarking process (e.g., giving an overview of the benchmark results and process).

* What was the general outcome of this benchmarking iteration?
* How does this compare to the goals for this benchmarking iteration (e.g., was there a focus on a new aspect to benchmark)?
* Are there real benchmarking results and interesting insights from this data?
  + How was the performance of the AI system compared to the baseline?
  + How was the performance of the AI system compared to other benchmarking initiatives (e.g., are the numbers plausible and consistent with clinical experience)?
  + How did the results change in comparison to the last benchmarking iteration?
* Are there any technical lessons?
  + Did the architecture, implementation, configuration, and hosting of the benchmarking system fulfil its objectives?
  + How was the performance and operational efficiency of the benchmarking itself (e.g., how long did it take to run the benchmarking for all AI models vs. one AI model; was the hardware sufficient)?
* Are there any lessons concerning data acquisition?
  + Was it possible to collect enough data?
  + Were the data as representative as needed and expected?
  + How good was the quality of the benchmarking data (e.g., how much work went into conflict resolution)?
  + Was it possible to find annotators?
  + Was there any relevant feedback from the annotators?
  + How long did it take to create the dataset?
* Is there any feedback from stakeholders about how the benchmarking helped them with decision-making?
  + Are metrics missing?
  + Do the stakeholders need different reports or additional metadata (e.g., do they need the “offline capability” included in the AI metadata so that they can have a report on the best offline system for a certain task)?
* Are there insights on the benchmarking process?
  + How was the interest in participation?
  + Are there reasons that someone could not join the benchmarking?
  + What was the feedback of participants on the benchmarking processes?
  + How did the participants learn about the benchmarking?

#### Retirement

TBC

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) “*AI software lifecycle specification”* (identification of standards and best practices that are relevant for the AI for health software life cycle).

* What happens with the data after the benchmarking (e.g., will they be deleted, stored for transparency, or published)?
* What happens to the submitted AI models after the benchmarking?
* Could the results be reproduced?
* Are there legal or compliance requirements to respond to data deletion requests?

### Benchmarking version [1]

This section includes all technological and operational details of the benchmarking process for the benchmarking version **[1]**

**Technical architecture**

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

A picture containing logo

Description automatically generated

*Figure 4: 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>

Graphical user interface, text, application

Description automatically generated

*Figure 5: 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 6).

Table, website

Description automatically generated

*Figure 6: 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, a minimal benchmarking system is created by the organizers for prototyping. The latter system will process submitted files in a Python 3 environment with a 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 allows participants to decide whether they want to make submissions publicly available or not.

## 

## Subtopic [B]

TBC

# Overall discussion of the benchmarking

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in 4.2.11).

* What is the overall outcome of the benchmarking thus far?
* Have there been important lessons?
* Are there any field implementation success stories?
* Are there any insights showing how the benchmarking results correspond to, for instance, clinical evaluation?
* Are there any insights showing the impact (e.g., health economic effects) of using AI systems that were selected based on the benchmarking?
* Was there any feedback from users of the AI system that provides insights on the effectiveness of benchmarking?
  + Did the AI system perform as predicted relative to the baselines?
  + Did other important factors prevent the use of the AI system despite a good benchmarking performance (e.g., usability, access, explainability, trust, and quality of service)?
* Were there instances of the benchmarking not meeting the expectations (or helping) the stakeholders? What was learned (and changed) as a result?
* What was learned from executing the benchmarking process and methodology (e.g., technical architecture, data acquisition, benchmarking process, benchmarking results, and legal/contractual framing)?

# Regulatory considerations

TBC

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *“*[*Regulatory considerations on AI for health”* *(WG-RC)*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), [DEL02\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6AF7C004-8BCE-4151-9F44-45F041A1EB1D%7D&file=DEL02_1.docx&action=default) *“Mapping of IMDRF essential principles to AI for health software”,* and[DEL02\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1ED0D4D1-876C-4A0F-AEF7-06D3F445F5E6%7D&file=DEL02_2.docx&action=default) *“Guidelines for AI based medical device (AI-MD): Regulatory requirements”* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) identifies standards and best practices that are relevant for the “*AI software lifecycle specification*.*”* The following sections discuss how the different regulatory aspects relate to the TG-Malaria.

## Existing applicable regulatory frameworks

TBC

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for TG-Malaria.

* What existing regulatory frameworks cover the type of AI in this TDD (e.g., MDR, FDA, GDPR, and ISO; maybe the systems in this topic group always require at least “MDR class 2b” or maybe they are not considered a medical device)?
* Are there any aspects to this AI system that require additional specific regulatory considerations?

## Regulatory features to be reported by benchmarking participants

TBC

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

* Which certifications and regulatory framework components of the previous section should be part of the metadata (e.g., as a table with structured selection of the points described in the previous section)?

## Regulatory requirements for the benchmarking systems

TBC

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

* Which regulatory frameworks apply to the benchmarking system itself?
* Are viable solutions with the necessary certifications already available?
* Could the TG implement such a solution?

## Regulatory approach for the topic group

TBC

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations.”*

* Documentation & Transparency
  + How will the development process of the benchmarking be documented in an effective, transparent, and traceable way?
* Risk management & Lifecycle approach
  + How will the risk management be implemented?
  + How is a life cycle approach throughout development and deployment of the benchmarking system structured?
* Data quality
  + How is the test data quality ensured (e.g., the process of harmonizing data of different sources, standards, and formats into a single dataset may cause bias, missing values, outliers, and errors)?
  + How are the corresponding processes document?
* Intended Use & Analytical and Clinical Validation
  + How are technical and clinical validation steps (as part of the lifecycle) ensured (e.g., as proposed in the IMDRF clinical evaluation framework)?
* Data Protection & Information Privacy
  + How is data privacy in the context of data protection regulations ensured, considering regional differences (e.g., securing large data sets against unauthorized access, collection, storage, management, transport, analysis, and destruction)? This is especially relevant if real patient data is used for the benchmarking.
* Engagement & Collaboration
  + How is stakeholder (regulators, developers, healthcare policymakers) feedback on the benchmarking collected, documented, and implemented?

# 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.

[10] R. Nakasi, J.F. Tusubira, A. Zawedde, A. Mansourian, E. Mwebaze. A web-based intelligence platform for diagnosis of malaria in thick blood smear images: A case for a developing Country. Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 984-985, 2020.

[11] R. Nakasi, E. Mwebaze, A. Zawedde, J.F. Tusubira, B. Akera, G. Maiga. A new approach for microscopic diagnosis of malaria parasites in thick blood smears using pre-trained deep learning models. Springer SN Applied Sciences 2, 1255, 2020.

[12] R. Nakasi, E. Mwebaze, A. Zawedde, J.F. Tusubira, G. Maiga. An approach for Assessing quality of labeled Data for a machine learning task in Malaria detection. COMPOSS’ 20: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies: PP. 301- 304, 2020.

[13] Arunava. Malaria cell images dataset. Available at :<https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria>

[14] Nakasi, R.; Mwebaze, E.; Zawedde, A. Mobile-Aware Deep Learning Algorithms for Malaria Parasites and White Blood Cells Localization in Thick Blood Smears. *Algorithms* **2021**, *14*, 17. https://doi.org/10.3390/a14010017

Annex A:  
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

|  |  |  |
| --- | --- | --- |
| Acronym/Term | Expansion | Comment |
| TDD | Topic Description Document | Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group-Malaria |
| TG | Topic Group |  |
| WG | Working Group |  |
| FGAI4H | Focus Group on AI for Health |  |
| AI | Artificial intelligence |  |
| ITU | International Telecommunication Union |  |
| WHO | World Health Organization |  |
| DEL | Deliverable |  |
| CfTGP | Call for topic group participation |  |
| AI4H | Artificial intelligence for health |  |
| IMDRF | International Medical Device Regulators Forum |  |
| MDR | Medical Device Regulation |  |
| ISO | International Standardization Organization |  |
| GDPR | General Data Protection Regulation |  |
| FDA | Food and Drug administration |  |
| SaMD | Software as a medical device |  |
| AI-MD | AI based medical device |  |
| LMIC | Low-and middle-income countries |  |
| GDP | Gross domestic product |  |
| API | Application programming interface |  |
| IP | Intellectual property |  |
| PII | Personal identifiable information |  |
| TBC | To Be Communicated |  |

Annex B:  
Declaration of conflict of interests

In accordance with the ITU transparency rules, this section lists the conflict-of-interest declarations for everyone who contributed to this document. Please see the guidelines in FGAI4H-F-105 “ToRs for the WG-Experts and call for experts” and the respective forms (Application form & Conflict of interest form).

Company/Institution/Individual XYZ

A short explanation of the company’s area of activity and how the work on this document might benefit the company and/or harm competitors. A list of all people who contributed to this document on behalf of this company and any personal interest in this company (e.g., shares).

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