|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-F-012 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | Zanzibar, 3-5 September 2019 |
| **DOCUMENT** |
| **Source:** | TG-Ophthalmo topic driver |
| **Title:** | TDD update: TG-Ophthalmo (Ophthalmology) |
| **Purpose:** | Discussion |
| **Contact:** | Arun ShroffXtend.aiIndia | Tel: +1 908-208-1100Fax: N/A Email: arun@xtend.ai |
| **Contact:** | Yanwu (Frank) XUArtificial Intelligence Innovation Business, BaiduChina | Tel: +86 13918541815Fax: +86 10 59922186Email: xuyanwu@baidu.com |

|  |  |
| --- | --- |
| **Abstract:** | This document is the Topic Description Document (TDD) containing the standardized benchmarking approach for the use of AI for Ophthalmology (Retinal Imaging Diagnostics). It follows the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6) relevant for setting up this benchmarking. This document is a draft and a work-in-progress until it is finally approved by the Focus Group. |

Table of Contents

[1 Introduction 3](#_Toc9613838)

[1.1 Topic Description 3](#_Toc9613839)

[1.1.1 Relevance 3](#_Toc9613840)

[1.1.2 Current approaches and gold standards for detection 4](#_Toc9613841)

[1.1.3 Impact of AI 4](#_Toc9613842)

[1.1.4 Impact of benchmarking AI Solutions 5](#_Toc9613843)

[1.2 Ethical considerations 5](#_Toc9613844)

[1.3 Existing AI solutions (includes datasets, systems and benchmarks) 5](#_Toc9613845)

[1.3.1 DR 5](#_Toc9613846)

[1.3.2 GC 5](#_Toc9613847)

[1.3.3 AMD 6](#_Toc9613848)

[2 AI4H Topic Group 6](#_Toc9613849)

[2.1 General mandate of the Topic Group 6](#_Toc9613850)

[2.2 Topic description document 7](#_Toc9613851)

[2.3 Subtopics 7](#_Toc9613852)

[2.4 Topic group participation 7](#_Toc9613853)

[2.5 Status of this Topic Group 7](#_Toc9613854)

[2.6 Next meetings 8](#_Toc9613855)

[3 Method 8](#_Toc9613856)

[3.1 Overview of the benchmarking 8](#_Toc9613857)

[3.2 AI Input Data Structure 8](#_Toc9613858)

[3.3 AI Output Data Structure 8](#_Toc9613859)

[3.4 Test Data Labels 9](#_Toc9613860)

[3.5 Scores and metrics 10](#_Toc9613861)

[3.6 Undisclosed test data set collection 10](#_Toc9613862)

[3.7 Benchmarking methodology and architecture 10](#_Toc9613863)

[4 Reporting methodology 11](#_Toc9613864)

[5 Results 11](#_Toc9613865)

[6 Discussion 11](#_Toc9613866)

[7 Declaration of conflict of interest 11](#_Toc9613867)

[Appendix A: Glossary 13](#_Toc9613868)

[References 14](#_Toc9613869)

# Introduction

As part of the work of the WHO/ITU Focus Group (FG) AI for health (AI4H), this document specifies a standardized benchmarking approach for AI-based applications for Ophthalmology (Retinal Imaging Diagnostics).

## Topic Description

This topic group is devoted to standardized benchmarking of artificial intelligence for Ophthalmology (Retinal Imaging Diagnostics). The specific conditions and diseases include Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), Glaucoma (GC) and and Pathological Myopia (PM).

Additional diseases and conditions that are relevant to this Topic Group may be added in the future.

DR is a serious eye-disease caused by diabetes that affects blood vessels in the light-sensitive tissue called the retina that lines the back of the eye. It is the most common cause of vision loss among people with diabetes and the leading cause of vision impairment and blindness among working-age adults worldwide.

AMD causes damage to the macula and is a leading cause of vision loss among people age 50 and older. The macula is a small spot near the center of the retina and the part of the eye needed for sharp, central vision, which lets us see objects that are straight ahead. While AMD by itself does not lead to complete blindness but loss of central vision in it can interfere with simple everyday activities.

[GC](https://nei.nih.gov/health/glaucoma/glaucoma_facts) is a group of diseases that damage the eye’s optic nerve—the bundle of nerve fibers that connects the eye to the brain and leads to vision loss and blindness. In adults, diabetes nearly doubles the risk of glaucoma.

PM represents a subgroup of myopia and affects up to 3% of the world population. Vision loss related to pathologic myopia is of great clinical significance as it can be progressive, irreversible and affects individuals during their most productive years. High myopia is defined as refractive error of at least -6.00D or an axial length of 26.5mm or more. Pathological or degenerative myopia is defined as “high myopia with any posterior myopia-specific pathology from axial elongation."

**(TBC: categorization of the topic** according to categorization guideline (currently [C-104](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-104.docx))

### Relevance

Diabetic Retinopathy (DR):

The WHO estimates that there are over 422 million people with diabetes worldwide.[[1]](#endnote-1) Of these 35% or over 148 million are estimated to have DR with potential for vision impairment and 11% or 48 million are estimated to have Vision Threating DR (VTDR) that can lead to blindness.[[2]](#endnote-2) Both the number of people with diabetes and those affected by DR are growing at alarming rates – and projected by 2040 to be 642 million with diabetes, 225 million with DR and 64 million with VTDR.

Age-related Macular Degeneration (AMD):

According to Lancet research, the number of people living with macular degeneration is expected to reach 196 million worldwide by 2020 and increase to 288 million by 2040 [[3]](#endnote-3) And AMD is a leading cause (3rd) of vision loss worldwide, by 2010, it has been responsible for approximately 5% of all blindness globally [[4]](#endnote-4). Age is a prominent risk factor for AMD. The risk of getting advanced AMD increases from 2% for those ages 50-59, to nearly 30% for those over the age of 75. Studies suggest in China the prevalence of early AMD in Chinese persons aged 50 years or older was 9.5% and that of late AMD was 1.0%[[5]](#endnote-5).

Glaucoma (GC):

There are nearly 40 million blind people in the world today, according to World Health Organization [[6]](#endnote-6). Another 285 million have visual impairment. Globally, 8% of all blindness is attributable to glaucoma, making it the leading cause of global irreversible blindness [[7]](#endnote-7). There were 60 million people with glaucoma in the world in 2010 and will be nearly 80 million by 2020. Of these 60 million, 7.4 million were bilaterally blind from glaucoma in 2010 and 11.2 million (14%) will be bilaterally blind in 2020.

In China, according to a study, it was estimated that 9.4 million (2.6%) people aged 40 years and older have glaucomatous optic neuropathy [[8]](#endnote-8). Of this number, 5.2 million (55%) are blind in at least one eye and 1.7 million (18.1%) are blind in both eyes.

Pathological myopia (PM):

PM has become a global burden of public health. Among myopic patients, about 35% have high myopia. Myopia leads to elongation of axial length, potentially causing pathological changes in retina and choroid. With an increase in myopic refraction, high myopia will develop into pathologic myopia, which is characterized by formation of pathologic changes at: (1) posterior pole, including tessellated fundus, posterior staphyloma, retino-choroidal degeneration, etc; (2) optic disc, including parapapillary atrophy, tilting, etc; (3) myopic maculopathy, including lacquer crack, Fuchs spot, CNV, etc. Pathologic myopia causes irreversible visual impairment to patients. Therefore, it’s important to have early diagnosis and regular follow-up.

The overall global prevalence is estimated to be 0.9-3.1% with regional variability. The prevalence of pathological myopia-related visual impairment has been reported as 0.1%-0.5% in European studies and 0.2% to 1.4% in Asian studies.

### Current approaches and gold standards for detection

DR detection requires capturing a photograph of the retina using specialized equipment such as a slit-lamp and fundus camera. The image is then examined by an ophthalmologist, optometrist or a trained professional to detect abnormalities such as microaneurysms, exudates, haemorrhages, macular edema, etc. to determine if DR is present and its severity and stage of progression.

In general DR can be classified as mild, moderate or vision-threatening, which includes severe non-proliferative DR (NPDR), proliferative DR (PDR) and diabetic macular edema (DME). Accurate diagnosis of DR from fundus camera images and grading its severity requires professional expertise and training.

The UK National Institute for Clinical Excellence (NICE) guideline states that a DR screening test should have sensitivity and specificity of at least 80% and 95% respectively, with a technical failure rate of less than 5%.[[9]](#endnote-9)

The gold standard photography method for the detection of DR is stereoscopic color fundus photography in 7 standard fields (30°) as defined by the Early Treatment Diabetic Retinopathy Study (ETDRS) group. [[10]](#endnote-10)

### Impact of AI

For all the diseases described above, vision loss and blindness can be delayed or prevented by early detection and treatment of the condition. This requires an examination and screening by a trained ophthalmologist or an eye care professional.

However, given the large numbers of people affected worldwide by these conditions, there are not sufficient specialists globally to screen everyone at risk. The shortfall is particularly acute in developing countries, including India, China and many countries in Asia and Africa. In addition to the dire shortage of trained professionals, many of the affected people live in remote areas with little or no access to an eye care clinic or a screening center.

In India, for example, there are over 72 million people with diabetes and an estimated 25 million have some stage of DR and about 7 million have VTDR. However, India only has 15,000 trained ophthalmologists, which in a nation with 1.3 billion people amounts to a mere 9 specialists per million. Kenya, with a population of 48 million has less than 100 ophthalmologists, and Angola, less than 20 for 29 million people.[[11]](#endnote-11)

In recent years, many AI systems using deep learning have been very successful in image recognition and classification tasks. For example, in the ImageNet challenge, requiring identification of objects in 1000 categories, the best models achieve a classification error rate of less than 5%. – exceeding the best human accuracy levels.

Many of these models, have now been adapted successfully for use in a variety of medical image diagnosis tasks such as melanoma, breast, lung cancer detection and DR, AMD, and Glaucoma.

AI and deep learning-based systems offer the following benefits:

* Bridge the shortage of healthcare professionals and provide access to screening where none exists.
* Increase overall efficiency and scalability of current screening methods.
* Provide earlier detection of many eye diseases thereby preventing vision loss for millions.
* Decrease overall health-care costs via earlier interventions when it is easier and less expensive to treat these diseases.

### Impact of benchmarking AI Solutions

An accurate way of benchmarking the performance of AI solutions to detect and diagnose DR, AMD and GC can have a major impact on selecting and implementing the best solution to address the global healthcare challenge posed by these diseases specially in the LMICs. This can in turn improve the lives of millions at risk for vison impairment and vision loss globally because they do not have access to human experts and infrastructure to get screened. This also fulfils the important objective of achieving the UN’s SDGs in health.

## Ethical considerations

( to be completed )

## Existing AI solutions (includes datasets, systems and benchmarks)

(The following section is sourced from the FG-AI4H-C-022 document from Baidu, available at the FG4AIH Site)

### DR

Publicly available datasets include the EyePACS dataset (around 90,000 fundus images, 5 levels of severity), [[12]](#endnote-12)

MESSIDOR dataset (1,200 images, 4 levels of severity), [[13]](#endnote-13)

the DIARETDB dataset (around 200 images marked with lesions), etc.[[14]](#endnote-14)

A team at Google published results in 2016 of a study for detecting DR working with doctors in India and the US. The results show that their AI model’s performance for DR detection and grading its severity was on-par with that of ophthalmologists. Their model had a combined accuracy score of 0.95, which was slightly better than the median of the 8 ophthalmologists consulted (measured at 0.91). [[15]](#endnote-15)

Currently, IDx-DR is the first FDA approved device for AI DR screening. Based on a customized CNN architecture and lesion characteristics, this device can achieve a sensitivity of 96.8% and a specificity of 87%.[[16]](#endnote-16)

The best reported performance on binary classification of no DR/non-referable DR vs. referable DR is a sensitivity of 94% and specificity of 98% [[17]](#endnote-17)

This work combined features both from deep ResNet and from meta-data, and classified the features with a gradient boosting decision tree.

For five level classification of no DR, mild, moderate, severe non-proliferative DR, and proliferative DR [[18]](#endnote-18) [[19]](#endnote-19) [[20]](#endnote-20), the best accuracy reported is 96% by a combination of GoogleNet and ResNet model.

### GC

Existing datasets include Online retinal fundus image dataset for glaucoma Analysis (ORIGA, 650 fundus images), Retinal fundus images for glaucoma analysis (RIGA, 760 images), ACHIKO-K (258 images), DRISHTI-GS (100 images mainly for optic disk and cup segmentation), etc.

AI practice on suspected glaucoma classification generally follow two approaches, i.e. an end-to-end whole image classification [[21]](#endnote-21) [[22]](#endnote-22), or a classification based on optic disk and cup information .[[23]](#endnote-23) For the end-to-end approach, a resulting AUC of 0.986 by training an inception-v3 network on their private dataset of 48000+ images was reported. [[24]](#endnote-24) A multitask deep CNN model based on a U-net sharing features for the glaucoma classification task was set up and the disc and cup segmentation task, achieving an AUC of 0.95 while providing some medical interpretability.

### AMD

Currently, most existing work of detecting AMD in fundus images address the problem as a binary classification between no/early stage AMD and intermediate/advanced stage AMD. The two commonly used datasets are the Age-Related Eye Disease Study (AREDS) dataset, which consists of fundus images from around 4,700 participants, and the Cooperative Health Research in the Region of Augsburg (KORA) dataset, which consists of fundus images from 2,840 patients. Most state-of-the-art methods for AMD binary classification are in one of the three following categories:

1. Using CNNs of existing architectures such as GoogleNet, VGG, etc. [[25]](#endnote-25) [[26]](#endnote-26). The best reported performance of this type of method is 94.3% accuracy, using an ensemble of several CNNs.
2. Using customized deep CNN models [[27]](#endnote-27) [[28]](#endnote-28) [[29]](#endnote-29). The best reported result is an AUC of 0.96 and an accuracy of 91.6% on AREDS dataset.
3. Using deep image features from pretrained CNN model and then classify with SVM or Random Forest model [[30]](#endnote-30) [[31]](#endnote-31). The best reported accuracy is 93.4%.

### PM:

Now there is only the PALM challenge which focuses on the investigation and development of algorithms associated with the diagnosis of Pathological Myopia (PM) and segmentation of lesions in fundus photos from PM patients. The goal of the challenge is to evaluate and compare automated algorithms for the detection of pathological myopia on a common dataset of retinal fundus images. The medical image analysis community were invited to participate for developing and testing existing and novel automated fundus classification and segmentation methods

# AI4H Topic Group

Over the past decade, considerable resources have been allocated to exploring the use of AI for health, which has revealed an immense potential. Yet, due to the c<https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx>omplexity of AI models, it is difficult to understand their strengths, weaknesses, and limitations. If the technology is poorly designed or the underlying training data are biased or incomplete, errors or problematic results can occur. AI technology can only be used with complete confidence if it has been quality controlled through a rigorous evaluation in a standardized way. Towards developing this standard assessment framework of AI for health, the ITU has established FG-AI4H in partnership with the WHO.

Thus far, FG-AI4H has established thirteen (13) topic groups. These are concerned with: AI and cardiovascular disease risk prediction, dermatology, falls among the elderly, histopathology, neuro-cognitive disorders, outbreak detection, ophthalmology (retinal imaging diagnostics), psychiatry, radiotherapy, snakebite and snake identification, symptom assessment, tuberculosis and Volumetric chest computed tompgraphy. A current list can be found at the FGAI4H website at:

<https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx>

As the work by the Focus Group continues, new Topic Groups will be created. To organize the Topic Groups, for each topic the Focus Group chose a topic driver. The exact responsibilities of the topic driver are still to be defined and are likely to change over time. The preliminary and yet-to-confirm list of the responsibilities includes:

* Creating the initial draft version(s) of the topic description document.
* Reviewing the input documents for the topic and moderating the integration in a dedicated session at each Focus Group meeting.
* Organizing regular phone calls to coordinate work on the topic description document between meetings.

During meeting C in Lausanne, Arun Shroff from Xtend.AI was selected as topic driver for the “Topic Group - AI for Ophthalmology (retinal imaging diagnostics).

## General mandate of the Topic Group

The Topic Group is a concept specific to the AI4H-FG. The preliminary responsibilities of the Topic Groups are:

1. Provide a forum for open communication among various stakeholders
2. Agree upon the benchmarking tasks of this topic and scoring metrics
3. Facilitate the collection of high quality labeled test data from different sources
4. Clarify the input and output format of the test data
5. Define and set-up the technical benchmarking infrastructure
6. Coordinate the benchmarking process in collaboration with the Focus Group management and working groups

## Topic description document

The primary output of each Topic Group is the topic description document (TDD) specifying all relevant aspects of the benchmarking for the individual topics. **This document is the TDD for the Topic Group on “AI for Ophthalmology (retinal imaging diagnostics)” (TG-Ophthalmo)** The document has been developed cooperatively over several FG-AI4H meetings starting from meeting D in Shanghai. Suggested changes to the document will be submitted as input documents for each meeting. The relevant changes will then be discussed and integrated into an official output document until the TDD ready for the first official benchmarking.

## Subtopics

Topic groups summarize similar AI benchmarking use cases to limit the number of use case specific meetings at the Focus Group meetings and to share similar parts of the benchmarking. However, in some cases, it is expected that inside a Topic Group different subtopic Groups can be established to pursue different topic-specific specializations. TG-Ophthalmo will start without separate subtopic Groups. They will introduce once participants with different benchmarking requirements join the Topic Group. It is expected to introduce subtopic Groups for “DR” and “AMD” and “Glaucoma” maybe added as soon as partners that are interested in benchmarking systems for these conditions.

## Topic group participation

The participation in both the focus and Topic Group is generally open and free of charge. Anyone who is from a member country of the ITU may participate. On the 14. of March 2019 the ITU published an official “call for participation” document outlining the process for joining the Focus Group and the Topic Group. For this topic, the corresponding call can be found [here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/ophthalmo.aspx).

## Status of this Topic Group

During meeting D it was discussed that the TDD should contain an explicit section describing the progress since the last meeting for the upcoming meeting. The following subsections serve this purpose:

### Status Update for Meeting D (Shanghai):

1. The first draft of the Topic Group Document was created and shared with members of the topic group.
2. We promoted the topic group via email and our networks to invite anyone interested to join and contribute.
3. Topic Group Member, Baidu.com provided some suggestions on including AMD, Glaucoma and other imaging methods that were incorporated into the TDD
4. The first version of the TDD was published (as [FGAI4H-D-038](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-038.docx) ) .

### Status Update for Meeting E (Geneva)

1. The updated Call for Topic Group participation for TG-Ophthalmo was published on the ITU website and can be [downloaded here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Ophthalmo.pdf).
2. We had several email exchanges with the topic group members to request inputs and updates to the TDD. Yanwu XU indicated that Dr Xingxing Cao from Baidu would provide topic group updates on their behalf.
3. Dr. Xingxing Cao, Baidu added the condition Pathological Myopia (PM) to the TDD.
4. An updated and revised TDD was published (as [FGAI4H-E-014-R01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-E-014-R01.docx) )
5. We reached out to our networks via email and social media (LinkedIn, Twitter), sharing the call for topic group participation and to spread the word.
6. We had email exchanges, calls and discussions with several groups and individuals interested in contributing to the topic group including the following:
	* Pearse A Keane, MD MSc FRC Ophth MRCSI, Consultant Ophthalmologist, Moorfields Eye Hospital, U.K.
	* Prof Leo Celi M.D. M.S. M.P.H., Clinical Research Director, Laboratory for Computational Physiology, Harvard-MIT Division of Health Science and Technology, Open Access (MIT) Project
	* Ashley Kras, M.D. M. S., Ophthalmologist & Bioinformatician (Harvard Medical School)

### Status Update for Meeting F (Zanzibar):

1. We continued to promote the topic group via email to our networks and via social media.
2. As a result, we received several inbound emails and interest in joining/contributing to the group. We had two new Topic group members who joined the group:
	* Dr Covadonga Bascaran, PHEC MSc Programme Director, International Centre for Eye Health (ICEH), London School of Hygiene & Tropical Medicine
	* Inês Sousa, Head of Intelligent Systems, Fraunhofer.
3. We had several online Meetings / calls and discussion with:
	* Prof Leo Celi, Clinical Research Director, Harvard MIT Division of Health Science and Technology and Ash Krasley: (online meeting on June 22, 2019)
		1. Details about MIT Open Access Project
		2. Potential collaboration with FGAI4H / Contribution of Data
	* Dr. Jorge Cuadros, O.D., PhD. (CEO), Founder & CEO of EyePACS LLC, USA

(online meeting on July 22, 2019).

EyePACS LLC is a US company with a telemedicine platform for ophthalmology deployed in more than 600 community health and primary care centers across the US and internationally.

They have a database of over 5 million diagnostic retinal images. We discussed possibility of collaborating with the topic group and contributing data. They expressed interest and discussions are ongoing.

1. Topic Group Meeting (Jul 31, 2019)
	* + Participants:
			- Dr. Covadonga Bascaran, PHEC MSc Programme Director, International Centre for Eye Health (ICEH), London School of Hygiene & Tropical Medicine
			- Inês Sousa , Head of Intelligent Systems, Fraunhofer Portugal
		+ Discussion about DR-Net:
		+ DR-NET is run and coordinated by the VISION 2020 LINKS team at the London School of Hygiene & Tropical Medicine in the UK. It brings together a LMIC eye department with a partner eye department in the UK with the aim of improving the quality and quantity of eye care services in the poorest countries in the world.
		+ The DR-NET data may be available for sharing and benchmarking on an anonymized basis subject to permissions which vary for each country. None of the data has been disclosed so far. However, the images are not labeled/ annotated and would need to be labeled in order to be used in benchmarking, once appropriate permissons are obtained,
2. Current members of the topic group include:
* Arun Shroff, CEO, Xtend.AI (USA)& CTO, Medindia.net (India) Topic Driver,
* Yanwu XU, Artificial Intelligence Innovation Business, Chief Scientist, Baidu, China
* Jingyu WANG, Artificial Intelligence Group, Baidu, China
* Shan Xu, CAICT, China
* Ashley Kras, M.D. M. S., Ophthalmologist & Bioinformatician (Harvard Medical School)
* Dr Covadonga Bascaran, PHEC MSc Programme Director, International Centre for Eye Health (ICEH), London School of Hygiene & Tropical Medicine
* Inês Sousa, Head of Intelligent Systems, Fraunhofer.

## Next meetings

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

* G: New Delhi, India; November 2019
* H: Brasilia, Brazil; January 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).

Tools/process of TG cooperation - to be filled out according to FG regulations

TG interaction with WG, FG- to be filled out according to FG regulations

# Method

## Overview of the benchmarking

The benchmarking of the algorithms for detecting DR, GC, AMD will be done on a sufficiently large and previously undisclosed test data set. All data will be provided as per the data acceptance guidelines published by the focus group. All data will be labelled by licensed ophthalmologists or eye-care professionals.

## AI Input Data Structure

The following input data structure is being proposed for all eye conditions - DR, AMD, GC.

Images of each retina captured with fundus cameras should be submitted as separate files in the following format:

* Image File Format: JPG or PNG format
* Image File Names: Images names will be anonymised to exclude any patient identifying information.
* Image Resolution: the images will be supplied in their original resolution as captured from the fundus cameras.

(TBD: do we need to include if image is of Right or Left Eye? If so it could be the first letter of the file name: example R23456.JPG )

## AI Output Data Structure

The output of the algorithm should be a CSV file in text format with the following columns:

* Name of the image file processed (for example: I705656.JPG OR L566768.PNG)
* The diagnosis of the retinal image as per the algorithm. The labels will depend upon the specific condition and the type of classification that is being benchmarked:
1. **DR:**
2. Classification: All DR severity levels:
* 0 (Non-gradable Image) (TBD: should this be one of the classifications.)
* 1(No DR)
* 2 (Mild)
* 3 (Moderate NPDR)
* 4 (Severe NPDR)
* 5 (PDR)
1. Classification: Referable or Non-referable DR:
* 0 (Ungradable Image. TBD)
* 1 (Non-referable Retinopathy – No DR or Mild DR)
* 2 (Referable Retinopathy - Moderate, Severe, PDR)
1. **AMD:**
* 0 (Image Ungradable. TBD)
* 1 (No/early stage AMD
* 2 (Intermediate/advanced stage AMD)
1. **GC:**
* 0 (Image Ungradable. TBD)
* 1 (No GC)
* 2 (GC)
1. **PM:**
* 0 (Image Ungradable. TBD)
* 1 (No PM/HM)
* 2 (HM: high myopia)
* 3 (PM)

## Test Data Labels

A separate CSV file in text format will be provided containing the following columns:

* Name of the Image File (example: R705656.JPG OR L566768.PNG)
* Label or Annotation of the Image that contains the diagnosis of the retinal image. The labels will depend upon the specific condition that is being benchmarked and also the type of classification. Currently, the following are being proposed:
1. **DR:**
2. Classification: All DR severity levels:
* 0 (Non-gradable Image) (TBD: should this be one of the classifications.)
* 1(No DR)
* 2 (Mild)
* 3 (Moderate NPDR)
* 4 (Severe NPDR)
* 5 (PDR)
1. Classification: Referable or Non-referable DR:
* 0 (Ungradable Image. TBD)
* 1 (Non-referable Retinopathy – No DR or Mild DR)
* 2 (Referable Retinopathy - Moderate, Severe, PDR)
1. **AMD:**
* 0 (Image Ungradable. TBD)
* 1 (No/early stage AMD
* 2 (Intermediate/advanced stage AMD)
1. **GC:**
* 0 (Image Ungradable. TBD)
* 1 (No GC)
* 2 (GC)
1. **d) PM:**
* 0 (Image Ungradable. TBD)
* 1 (No PM/HM)
* 2 (HM: high myopia)
* 3 (PM)

## Scores and metrics

All metrics will be computed based on the performance of the algorithm on the undisclosed test data-set. For the ophthalmological conditions in this topic group, the task may be either classification or segmentation. For example, in the case of DR an image may be classified as having No DR, mild, moderate, severe or PDR. In the case of segmentation tasks, each pixel of an image is classified into various classes. For example, in the case of GC, usually segmentation is performed on the image by identifying pixels comprising the optic disc and optic cup areas of the retina and calculating the ratio of the areas to detect GC.

We start with a few definitions:

* An instance is either a single image (for classification tasks), or a patch or a pixel of an image (for segmentation tasks).
* True Positive (TP) is the number of positive (disease) instances which are correctlyclassified.
* True Negative (TN) is the number of negative (normal) instances which are correctly classified.
* False Positive (FP) is the number of positive (disease) instances which are incorrectly classified.
* False Negative (FN) is the number of negative (normal) instances which are incorrectly classified.

Based on the above definitions, the following are the most common metrics used to evaluate performance of algorithms:

* Sensitivity(Recall) or True Positive Rate is the proportion of correctly classified positive (disease) instances. It is calculated as TP / (TP + FN)
* Specificity or True negative rate is the proportion of correctly classified negative (normal) instances. It is calculated as TN / (TN + FP)
* Precision or Positive Predictive Value is the fraction of positive (disease) instances that are correctly classified. It is calculated as TP / (TP + FP).
* F1-Score combines Precision and Recall into a single metric. It is calculated as the harmonic mean of Precision and Recall. It is calculated as 2 x (Precision x Recall) / (Precision + Recall)
* Accuracy is the proportion of instances that are correctly classified. It is calculated as (TP + TN) / (TP + FP + TN + FN)
* AUC (Area Under Receiver Operating Curve or ROC): The ROC is a plot of True Positive Rate (Sensitivity) vs. False Positive Rate (1- Specificity)) at different predictive thresholds of the classifier. The AUC has a value between 0 and 1. The closer it is to 1 the better the performance.
* Intersection over Union (IOU): This metric is used only for segmentation tasks. IOU is defined as follows: IOU = Area (A ∩ G) / Area ( A ∪ G)

where A indicates the segmentation from the algorithm and G indicates the manual ground truth segmentation of an image.

In all the above cases – higher values are better and algorithms would be ranked in descending order of these metrics.

## Undisclosed test data set collection

(TBC)

* 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 C-004)

## Benchmarking methodology and architecture

(TBC)

* technical architecture
* hosting (IIC, etc.)
* possibility of an online benchmarking on a public test dataset
* protocol for performing the benchmarking (who does what when etc.)
* AI submission procedure including contracts, rights, IP etc. considerations

# Reporting methodology

(TBC)

* 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

# Results

Chapter 6 will outline the results from performing the benchmarking based on the methodology specified in this document. Since the benchmarking is still in its specification phase, there are no results available yet. Depending on the progress made on this document, first preliminary test benchmarking results on small public data sets are expected by the end of 2019. The first official results form an MVB are expected in early 2020.

# Discussion

(TBC)

# Declaration of conflict of interest

In accordance with the ITU rules in this section working on this document should define his conflicts of interest that could potentially bias his point of view and the work on this document.

Xtend.AI

Xtenda.ai is a start-up focused on using AI for solving global challenges in health and other domains. It is backed by Medindia.net – a leading online publisher of health information, and a developer of health applications and services for consumers, doctors, healthcare professionals globally. Medindia’s website and applications are visited by over 4 million visitors each month from over 230 countries. Medindia offers almost 1 million pages of trusted health and wellness information including news, special reports, articles, animations, slides, infographics, videos, health directories, drug information, calculators, personalized health record, mobile apps, interactive tools, applications and much more. All of Medindia’s content is edited and authenticated by doctors and healthcare professionals. It is certified to comply with the HONCode standard for trustworthy health information. Medindia.net is headquartered in India and owned operated by Medindia4u.com Pvt. Ltd. – a private limited company based in Chennai, India. It has a marketing and support offices in USA.

Involved people: Arun Shroff, CEO of Xtend.AI and CTO of Medindia.net. Topic Driver for this topic group.

Baidu.com

(Following sourced from Baidu’s document: (C-022) – please edit as required)

Baidu is an international company with leading AI technology and platforms. Baidu’s retinal algorithms focus not only on inputting an image and outputting several eye-disease risks, but also building a powerful AI retinal system that integrates all related AI capacity to provide better service and enhance the end-user experience. The AI retinal system aims to build a personal eye-health management and analysis platform for each user. Baidu’s mission is to defend people’s eyes and global health with AI.

Since 2016, Baidu has positioned AI as a strategic driver for the development of its business. Under the strategy of “strengthening the mobile foundation and leading in AI”, Baidu has steadily improved its AI ecosystem, with productization and commercialization continuing to accelerate.

As integral components to its overall AI ecosystem, Baidu has developed two open ecosystems - the Apollo open autonomous driving platform and DuerOS, the company’s conversational AI system, which operates in two important scenarios – intelligent driving and smart homes. So far, with its latest iteration – “Apollo 3.0”, Baidu’s autonomous driving platform has brought together over 130 partners and has been granted the first batches of licenses for autonomous driving public road tests from Beijing, Chongqing and Fujian. In the smart living field, Baidu has co-launched over 160 DuerOS-powered hardware products, covering smart speakers, children’s wearables, televisions, automobiles, hotels and other vertical businesses. In September 2018, the install base of DuerOS reached 141 million devices with over 800 million voice queries. After years of commercial exploration, Baidu has formed a comprehensive AI ecosystem and is now at the forefront of the AI industry in terms of fundamental technological capability, speed of productization and commercialization, and “open” strategy. In the future, Baidu will continue to enhance user experience and accelerate the development of AI applications through the strategy of “strengthening the mobile foundation and leading in AI”.

Involved people:

* Yanwu XU, Artificial Intelligence Innovation Business, Chief Scientist, Baidu China
* Xingxing Cao, Artificial Intelligence Innovation Business, Baidu, China
* Jingyu WANG Artificial Intelligence Group, Baidu, China
* Shan Xu, CAICT, China

# Appendix A:Glossary

This section lists all the relevant abbreviations and acronyms used in the document. If there is an external source

* **AI** - [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) – an umbrella term that refers to one or more of the various fields of computer science including machine learning, neural networks and deep learning.
* **AI4H** - AI for health - An [ITU-T SG16 Focus Group](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx) founded in cooperation with the WHO in July 2018.
* **API** - [Application Programming Interface](https://en.wikipedia.org/wiki/Application_programming_interface) - the software interface systems communicate through.
* **FG** - [Focus Group](https://www.itu.int/en/ITU-T/focusgroups/Pages/default.aspx) - An instrument created by ITU-T providing an alternative working environment for the quick development of specifications in their chosen areas.
* **IIC** - International Computing Centre - The United Nations data center that will host the benchmarking infrastructure.
* **ITU** - [International Telecommunication Union](https://www.itu.int) - The United Nations specialized agency for information and communication technologies – ICTs.
* **LMIC** - Low and Middle Income Countries
* **MVB** - minimal viable benchmarking
* **NGO** - [Non Governmental Organization](https://en.wikipedia.org/wiki/Non-governmental_organization) - NGOs are usually non-profit and sometimes international organizations independent of governments and international governmental organizations that are active in humanitarian, educational, health care, public policy, social, human rights, environmental, and other areas to affect changes according to their objectives. (from Wikipedia.en)
* **SDG** - [Sustainable Development Goals](https://www.un.org/sustainabledevelopment/) - The United Nations Sustainable Development Goals are the blueprint to achieve a better and more sustainable future for all. Currently there are 17 goals defined. SDG 3 is to “Ensure healthy lives and promote well-being for all at all ages” and is therefore the goal that will benefit from the AI4H Focus Groups work the most.
* **TBC** - A topic group item to be completed.
* **TBD** - A topic group item to be discussed / determined
* **TDD** - Topic Description Document - Document specifying the standardized benchmarking for a topic FG AI4H Topic Group works on. This document is the TDD for the Topic Group “AI for Ophthalmology (retinal imaging diagnostics)”.
* **TG** - Topic Group - Structures inside AI4H FG summarizing similar use cases and working on a TDD specifying the setup of a standardized benchmarking for the corresponding topic. The Topic Groups have been first introduced by the FG at the Meeting C, January 2019 in Lausanne. See protocol FG-AI4H-C-10x for details.
* **WHO** - [World Health Organization](https://www.who.int) - The United Nations specialized agency for international public health.

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

# References

1. <https://www.who.int/en/news-room/fact-sheets/detail/diabetes> [↑](#endnote-ref-1)
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4657234/> [↑](#endnote-ref-2)
3. 1. [https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(13)70145-1/fulltext](https://www.thelancet.com/journals/langlo/article/PIIS2214-109X%2813%2970145-1/fulltext) (global) [↑](#endnote-ref-3)
4. 1. <https://www.brightfocus.org/macular/article/age-related-macular-facts-figures> (global) [↑](#endnote-ref-4)
5. 1. <https://www.ncbi.nlm.nih.gov/pubmed/29302323> (China) [↑](#endnote-ref-5)
6. 1. [https://www.aaojournal.org/article/S0161-6420(14)00433-3/abstract](https://www.aaojournal.org/article/S0161-6420%2814%2900433-3/abstract) (global) [↑](#endnote-ref-6)
7. 1. <https://www.glaucoma.org/glaucoma/glaucoma-facts-and-stats.php> (global) [↑](#endnote-ref-7)
8. <https://bjo.bmj.com/content/85/11/1277> (China) [↑](#endnote-ref-8)
9. National Institute for Clinical Excellence. Management of Type 2 Diabetes: Retinopathy—Screening and Early Management. London, UK: Inherited Clinical Guideline; 2002 [↑](#endnote-ref-9)
10. <https://www.ncbi.nlm.nih.gov/pubmed/2062513/> [↑](#endnote-ref-10)
11. International Council of Ophthalmology. Number of Ophthalmologists in Practice and Training Worldwide. <http://www.icoph.org/ophthalmologists-worldwide.html> [↑](#endnote-ref-11)
12. <http://www.eyepacs.com/data-analysis> [↑](#endnote-ref-12)
13. <http://www.adcis.net/en/third-party/messidor/> [↑](#endnote-ref-13)
14. <http://www.it.lut.fi/project/imageret/diaretdb1/> [↑](#endnote-ref-14)
15. Varun Gulshan, PhD1; Lily Peng, MD, PhD1; Marc Coram, PhD; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. <https://jamanetwork.com/journals/jama/fullarticle/2588763> [↑](#endnote-ref-15)
16. M. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. Folk, M. Niemeijer, Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning, Investigative ophthalmology & visual science 57 (13) (2016) 5200–5206. [↑](#endnote-ref-16)
17. 1. R. Gargeya, T. Leng, Automated identification of diabetic retinopathy using deep learning, Ophthalmology 124 (7) (2017) 962–969. [↑](#endnote-ref-17)
18. 1. H. Takahashi, H. Tampo, Y. Arai, Y. Inoue, H. Kawashima, Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy, PloS one 12 (6) (2017) e0179790. [↑](#endnote-ref-18)
19. 1. G. García, J. Gallardo, A. Mauricio, J. López, C. Del Carpio, Detection of diabetic retinopathy based on a convolutional neural network using retinal fundus images, in: International Conference on Artificial Neural Networks, Springer, 2017, pp. 635–642. [↑](#endnote-ref-19)
20. V. Gulshan, L. Peng, M. Coram, M. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, Jama 316 (22) (2016) 2402–2410. [↑](#endnote-ref-20)
21. L. Zhixi, Y. He, S. Keel, W. Meng, R. Chang, M. He, Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs, Ophthalmology 125 (8) (2018) 1199–1206. doi:10.1016/j.ophtha.2018.01.023. [↑](#endnote-ref-21)
22. C. Xiangyu, X. Yanwu, W. Damon, W. Tien, L. Jiang, Glaucoma detection based on deep convolutional neural network. (Aug 2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26736362/ [↑](#endnote-ref-22)
23. 1. H. Fu, J. Cheng, Y. Xu, C. Zhang, D. Wong, J. Liu, X. Cao, Disc-aware ensemble network for glaucoma screening from fundus image, IEEE Transactions on Medical Imaging. [↑](#endnote-ref-23)
24. A. Chakravarty, J. Sivswamy, A deep learning based joint segmentation and classification framework for glaucoma assesment in retinal color fundus images, arXiv preprint arXiv:1808.01355 [↑](#endnote-ref-24)
25. F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. Zimmermann, B. Linkohr, A. Peters, I. Heid, C. Palm, B. Weber, A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography, Ophthalmology [↑](#endnote-ref-25)
26. A. Govindaiah, M. Hussain, R. Smith, A. Bhuiyan, Deep convolutional neural network based screening and assessment of age-related macular degeneration from fundus images, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE, 2018, pp. 1525– 1528. [↑](#endnote-ref-26)
27. 1. P. Burlina, N. Joshi, M. Pekala, K. Pacheco, D. Freund, N. Bressler, Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks, JAMA ophthalmology 135 (11) (2017) 1170–1176. [↑](#endnote-ref-27)
28. J. Tan, S. Bhandary, S. Sivaprasad, Y. Hagiwara, A. Bagchi, U. Raghavendra, A. Rao, B. Raju, N. Shetty, A. Gertych, et al., Age- related macular degeneration detection using deep convolutional neural network, Future Generation Computer Systems 87 (2018) 127–135 [↑](#endnote-ref-28)
29. S. Matsuba, H. Tabuchi, H. Ohsugi, H. Enno, N. Ishitobi, H. Masumoto, Y. Kiuchi, Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration, International ophthalmology (2018) 1–7. [↑](#endnote-ref-29)
30. 1. A. Horta, N. Joshi, M. Pekala, K. Pacheco, J. Kong, N. Bressler, D. Fre- und, P. Burlina, A hybrid approach for incorporating deep visual fea- tures and side channel information with applications to amd detection, in: Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on, IEEE, 2017, pp. 716–720. [↑](#endnote-ref-30)
31. 1. P. Burlina, K. Pacheco, N. Joshi, D. Freund, N. Bressler, Comparing humans and deep learning performance for grading amd: A study in using universal deep features and transfer learning for automated amd analysis, Computers in biology and medicine 82 (2017) 80–86.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ [↑](#endnote-ref-31)