|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-J-010-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 30 September – 2 October 2020 |
| **DOCUMENT** |
| **Source:** | TG-Dental Topic Driver |
| **Title:** | Att.1 – TDD update (TG-Dental) |
| **Purpose:** | Discussion |
| **Contact:** | Falk SchwendickeCharité – Universitätsmedizin BerlinGermany | Tel: +4930450662556Fax: +40304507562556Email: falk.schwendicke@charite.de |
| **Contact:** | Joachim KroisCharité – Universitätsmedizin BerlinGermany | Tel: +4930450562328Fax: +40304507562932Email: Joachim.krois@charite.de |
| **Contact:** | Tarry Singhdeepkapha.aiNetherlands | Email: tarry.singh@deepkapha.ai |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies the state of progress of the TG-Dental with respect to a standardized benchmarking framework for AI-based systems for dental diagnostics and digital dentistry. It follows the structure defined in FGAI4H-C-105 and covers all scientific, technical and administrative aspects relevant for setting up this benchmarking. The creation of this document is an ongoing process until it will be finally approved by the Focus Group. This draft will be a continuous Input- and Output-Document. |

|  |  |
| --- | --- |
| **Changelog:** | Version 3.0 (submitted as [FGAI4H-J-010-A01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-J-010-A01.docx) for e-meeting J; remote)* Updates for section 2, AI4H Topic group.
	+ Expanding list of contributors (adding information on Dr Robert Gaudin and Dr Akhilanand Chaurasia)
* Adding section 3.5.2 Tooth level metrics for segmentation tasks

Version 2.0 (submitted as FGAI4H-I-010-A01 for e-meeting I in Geneva)* Update of TDD abstract
* Formatting of the document to emphasis sections not yet populated with text
* Updates for section 2, AI4H Topic group.
	+ Expanding list of contributors (adding information on Tarry Singh and deepkapha.ai)
	+ Defining tools for communication
	+ Setting up biweekly calls
	+ Updating the current status

Version 1.0 (submitted as [FGAI4H-H-010-A01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-H-010-A01.docx) for meeting H in Brasilia)This document served as initial draft for the topic description document (TDD) of the topic group TG-Dental, which is concerned with the standardized benchmarking of AI for dental diagnostics and digital dentistry. The focus for the first draft was on the sections 1. Introduction and 3. Methods. * Introduction
	+ Relevance
	+ Impact
	+ Existing work
* Method
	+ Anatomical structures
	+ Pathologies
	+ Data sets and format
	+ AI Output Data Structure
	+ Metrics
 |
|  |  |

Table of Contents

[1 Introduction 5](#_Toc51948473)

[1.1 Document Structure 5](#_Toc51948474)

[1.2 Topic Description 5](#_Toc51948475)

[1.2.1 Topic group outline 5](#_Toc51948476)

[1.2.2 Relevance 6](#_Toc51948477)

[1.2.3 Impact 6](#_Toc51948478)

[1.3 Ethical Considerations 6](#_Toc51948479)

[1.4 Existing AI Solutions 6](#_Toc51948480)

[1.4.1 Existing work 6](#_Toc51948481)

[1.5 Existing work on benchmarking 7](#_Toc51948482)

[2 AI4H Topic group 7](#_Toc51948483)

[2.1.1 Structure 7](#_Toc51948484)

[2.1.2 Participation 7](#_Toc51948485)

[2.1.3 Tools/process of cooperation 7](#_Toc51948486)

[2.1.4 Interaction 7](#_Toc51948487)

[2.1.5 Current status 7](#_Toc51948488)

[2.1.6 Contributors 8](#_Toc51948489)

[2.1.7 Meetings 8](#_Toc51948490)

[2.1.8 Next steps to work on this document 8](#_Toc51948491)

[3 Method 9](#_Toc51948492)

[3.1.1 Anatomical structures 9](#_Toc51948493)

[3.1.2 Pathologies 9](#_Toc51948494)

[3.2 AI Input Data Structure 9](#_Toc51948495)

[3.2.1 Inputs for benchmarking 9](#_Toc51948496)

[3.2.2 Data sets and format 10](#_Toc51948497)

[3.3 AI Output Data Structure 10](#_Toc51948498)

[3.3.1 Inputs for benchmarking 10](#_Toc51948499)

[3.3.2 Data format 11](#_Toc51948500)

[3.4 Test Data Labels 11](#_Toc51948501)

[3.4.1 Label types 11](#_Toc51948502)

[3.4.2 Data format 11](#_Toc51948503)

[3.5 Scores & Metrics 11](#_Toc51948504)

[3.5.1 Metrics 11](#_Toc51948505)

[3.5.2 Tooth level metrics for segmentation tasks 12](#_Toc51948506)

[3.6 Undisclosed Test Data Set Collection 13](#_Toc51948507)

[3.7 Benchmarking Methodology and Architecture 13](#_Toc51948508)

[3.8 Reporting Methodology 13](#_Toc51948509)

[4 Results 13](#_Toc51948510)

[5 Discussion 14](#_Toc51948511)

[6 Declaration of Conflict of Interest 14](#_Toc51948512)

# Introduction

## Document Structure

* *overview of the whole document*

## Topic Description

* *description of topic*
* *categorization of the topic according to categorization guideline (currently C-0xx)*
* *relevance of the health topic*
* *gold standard of current health topic handling*
* *possible impact of AI in this topic*
* *expected impact of the benchmarking*

### Topic group outline

The topic group **Dental Diagnostics and Digital Dentistry (*TG-Dental*)** focusses on diagnostics in dentistry, dental and oral medicine, and associated disciplines (oral and maxillofacial surgery, orthodontics, dental and oral traumatology). Specifically, it deals with the following (not exhaustive) ICD-10 code headings:

* K00 Disorders of tooth development and eruption
* K01 Embedded and impacted teeth
* K02 Dental caries
* K03 Other diseases of hard tissues of teeth
* K04 Diseases of the pulp and periapical tissues
* K05 Gingivitis and periodontal diseases
* K06 Other disorders of the gingiva and edentulous alveolar ridges
* K07 Dento-facial anomalies including malocclusion
* K08 Other disorders of the teeth and supporting structures
* K09 Cysts of oral origin, not elsewhere classified
* K10 Other diseases of jaws
* K11 Diseases of the salivary glands
* K12 Stomatitis and related lesions
* K13 Other diseases of the lip and oral mucosa
* K14 Diseases of the tongue

Diagnostics include the detection, assessment and prognosis of and on these conditions, as well as diagnostics on patients’ risk to experience a condition or its progression. Diagnostics also includes the (image or non-image) assessment of anatomic structures or physiologic functions for the purpose of both detection of conditions, but also treatment planning for various therapies (e.g. implantology).

The TG falls in the following FOG AI4H classes:

* 1b.6 Dentistry
* 2.1.1 Machine learning, 2.1.2 Regression tasks, 2.1.7 Anomaly detection, 2.5.1 Visual recognition
* 3.1-3.4, with a specific focus on 3.2 Medical Images
* 4.1 PACS, 4.2 EHR, 4.3 PHR
* 5.1 Service providers

### Relevance

Dental conditions, like caries, periodontitis or tooth loss, are among the most prevalent diseases of humankind, affecting up to 98% of a population. Direct treatment costs due to dental diseases globally were estimated at US $298 billion annually, corresponding to an average of 4.6% of global health expenditure. The burden emanating from oral diseases is comparable to that from diabetes or cardiovascular diseases. The majority of the world's 1.6 million dentists are based in Europe and the Americas, such that 69% of the world's dentists serve 27% of the global population. Africa has only 1% of the global workforce. The overall workforce in dentistry exceeds 10 million worldwide. Diagnostics in dentistry largely relies on dentists diagnosing diseases via a combination of dental history taking, clinical investigation and imaging as well, if required, further physical or (bio)chemical or microbiologic assessments.

### Impact

AI will help to (1) improve the accuracy of each of these individual tasks, (2) allow the integration of different data with higher effectiveness than the individual can do this, (3) without ease also longitudinally assess these data, compare them over time, and hence allow predictions, (4) reduce the reliance of diagnosis making from the dentist, expanding the scope of dental auxiliary staff, thereby increasing the access and efficiency of dental services, and (5) enable patients and healthy individuals to better participate into their dental health experience and management. AI will pave the way to a more personalized, precise, preventive and participatory dentistry for more people worldwide. It has the potential to aid in overcoming current ineffective, expensive care models.

Expected impact of benchmarking: Benchmarking is expected to yield more robust models and algorithms, with initially lower accuracy compared with current validation strategies (largely in-sample). Benchmarking is further expected to allow transparent comparisons of different models and algorithms.

## Ethical Considerations

* *ethical considerations on usage of AI*
* *ethical consideration of and benchmarking including its data acquisition*

## Existing AI Solutions

* *current systems available with their inputs, output, focus/bias*
* *existing benchmarking including self-stated performance*

### Existing work

The project starts from scratch. However, there are currently a range of AI solutions developed, focusing on (1) automated detection and segmentation of structures and conditions on 2-D radiographs, (2) automated detection and segmentation of structures on 3-D radiographs, (3) automated detection and segmentation of structures on surface scans (STL), and (4) on predictive modelling for various dental conditions (e.g. caries, periodontitis). The former three are situated both in academia and technology startups, while the last is mainly an academic exercise at present. There is currently only one validated, certified AI solution available in dental diagnostics focusing on landmark detection on cephalometric radiographs for orthodontic purposes.

## Existing work on benchmarking

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

# AI4H Topic group

* *Topic group structure*
	+ *Subtopic 1*
	+ *Subtopic 2*
* *Topic group participation*
* *Tools/process of TG cooperation*
* *TG interaction with WG, FG*
* *Current topic group and topic status*
* *Contributors so far*
* *Next meetings*
* *Next steps for the work on this document*

### Structure

* The group will have a speaker, who is elected for 1 year by the group members. The speaker coordinates the group’s activities. A wider board may be implemented if needed.
* The group will consist of subgroups along diagnostic strategies and data, and – if needed – conditions.
* The first subgroup to be established is AI in dental imagery analysis, with a focus on dental radiographs.
* Further subgroups are subject to contributors, and will be agreed to by the group.
* The group will seek financial support from sponsors to fund its activities and support it. Funding will be made transparent and must be unrelated to the group’s activities.

### Participation

The group will be open for participation for everyone.

### Tools/process of cooperation

We use [sharepoint provided by ITU](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx) as our collaborative document repository. Further, we use the slack group (AI4H-TG-Dental) for communications.

### Interaction

* Group members will regularly interact via email or short messenger systems.
* A regular quarterly report will be provided by email by the speaker to the group’s members.
* The group will convene in irregular intervals according to availability.

### Current status

* The group applied for participation at meeting G in New Delhi.
* The group was officially presented at meeting H in Brasilia.
* The group is actively looking for contributors from academia and industry.

### Contributors

* Prof. Dr. Falk Schwendicke, Charité - Universitätsmedizin Berlin, Germany
Prof. Dr. Falk Schwendicke is Head of the Department of Oral Diagnostics, Digital Health and Health Services Research at Charité - Universitätsmedizin Berlin, Europe's largest university hospital, and founding director of the Berlin Institute for AI and Health Policy. He is a specialist in dental diagnostics, preventive and operative dentistry, has extensive experience in both practice and university dentistry and is author or editor of >300 scientific articles and books.
* Dr. Joachim Krois, Charité - Universitätsmedizin Berlin, Germany
Dr. Joachim Krois is senior data scientist at Charité – Universitätsmedizin Berlin and project lead for the project “Deep Learning for Dental Image Diagnostics”. He is an academic lecturer and cooperate trainer for coding, statistics, machine learning and scientific computing.
* Tarry Singh, deepkapha.ai, The Netherlands
Tarry Singh, the CEO of deepkapha.ai, is a visiting faculty lecturer at University of Dallas, Texas and at University of Utrecht and he is a startup mentor at Hult Prize Foundation, which is sponsored by President Bill Clinton. The company deepkapha.ai was founded with the sole goal of solving problems with Artificial Intelligence and Deep Learning. Deepkapha.ai develops niche algorithms that outperform leading state-of-the algorithms. The company focuses on solving extremely hard and complex problems within healthcare with the use of AI.
* Prof. Jae-Hong Lee, Wonkwang University College of Dentistry, Daejeon, Korea

Prof. Jae-Hong Lee is Associate Professor at the Department of Periodontology, Daejeon Dental Hospital, Wonkwang University College of Dentistry, Daejeon, Korea, Associate Director of Information Correspond Affairs at Korean Association of Periodontology and Associate Director of General Affairs at Korean Academy of Laboratory and Diagnostic Dentistry

* Dr. Akhilanand Chaurasia, King George Medical University, Lucknow, India
Dr. Akhilanand Chaurasia is a clinical consultant, academician and researcher at Department of Oral Medicine and Radiology, King George Medical University, Lucknow, India. He has remarkable credit of contribution to academics by publishing more than 75 research papers, Books and Book chapters, peer reviewer of more than 50 international journals. He is editor in chief of 2 national journals, member of editorial board of 14 international journals and associate editor of 1 international journal. For his elite contribution to academics and research, he has been awarded by many pristine fellowships and memberships like Union for International Cancer Control technical Fellowship (FUICC), MFDT, RCSEd (Edinburgh), MFDS, RCPS(Glasgow), MFDS, RCS (England) and many more accolades from world renowned Institutions.

### Meetings

A biweekly video call schedule is currently established.

### Next steps to work on this document

The document is first to be populated section by section. Then, specific subsections are to be expanded on, depending on the specific activities and directions this group takes.

# Method

* *Overview of the benchmarking*

### Anatomical structures

In dentistry the assessment of anatomical structures is related to the segmentation of bones (e.g. jaws), soft tissue (e.g. gingiva) and teeth (or a at more detailed level enamel, dentin and pulp). In particular the detection and classification of teeth is mandatory as most clinical reporting schemas are based on tooth level. Hence, object detection and instance segmentation models deem most appropriate. For instance, our study group achieved good results using U-Net, Mask R-CNN and RetinaNet. With respect to structures we recommend the metrics Intersection over Union (IoU), sometimes referred to as Jaccard Index, or the Dice Coefficient (DC). For particular tasks such as the determination of the gingiva’s edge mean squared error (MSE) computed for each pixel column is a reasonable choice. Further, for tooth detection/segmentation, a multi-class classification task, mean average precision (MAP) or a binarization e.g. based on IoU>0.5 followed by the classification accuracy deem appropriate metrics too.

### Pathologies

Similar to anatomical structures some dental pathologies (e.g. caries) can be seen as polygonal structures or pixel blobs, but with less distinct boarders compared to anatomical structures. In principle it would be possible to formulate dental pathology diagnostics as a binary classification task, we however advocate to frame it as an object detection and/or instance segmentation task. First, most dental imagery covers a larger part of the patient’s dentition, hence more than one part/tooth covered by the image may be affected by the disease. Not accounting for the location would diminish the value of an AI-based diagnosis tool. Second, though it would be possibly to crop an image to tooth level and to make predictions on an image segment, we found that the models we applied perform better if provided a larger view (this may be related to the intra-patient correlational structure of particular pathologies or their spatial distribution). Consequently, similar as outlined above, metrics such as IoU or DC deem most appropriate for dental pathologies. However, we want to stress out that according to the practiced tooth level-based reporting schemas and the familiarity of the dental researcher with metrics such as the classification accuracy, AUC, sensitivity (aka recall), specificity, PPV (aka precession) and NPV, among others, we advocate a tooth based or finding based evaluation. This however increases the complexity of the model evaluation process as it may become necessary to split apart the polygonal structures or pixel blobs predicted by the AI model and in case assign these to a particular tooth or if not feasible compute those metrics on an instances level. Finally, as the prevalence of pathologies on tooth level is low we advocate to leverage metrics such as F1-score and the precision-recall-curve (PRC), which are more robust with respect imbalanced classes.

## AI Input Data Structure

* *possible inputs for benchmarking*
* *ontologies, terminologies*
* *data format*

### Inputs for benchmarking

* Imagery data
* Patient records and clinical data
* *A priori* knowledge. Implementing a prior knowledge will increase the predictive power of the algorithms.
	+ As many dental diseases are observed and reported tooth wise or even side-wise, many of those diseases show considerable intra-patient correlational structures and clustering effects.
	+ Many dental diseases are corelated with patient level variables such as age or smoking status.

### Data sets and format

The datasets available consist:

* Imagery data
* Patient records and clinical data
* Claims data

The data formats are as follows:

* Image data
	+ In dentistry DICOM is a standard for imaging information and related data. That data format allows to store the raw image data together with meta information such as date of survey, device information and pixel size, among others. It should be noted that in most cases sensitive patient information is included by default. Hence a rigor standardization should be applied to convey as much as information as possible e.g. device details or (anonymized) patient data such as age, at the same time sensible patient information should be excluded.
	+ Standard image files such as jpg, tif or png. In principle standard image files are sufficient to train prediction models, however without additional population data and information about the image generation process it becomes more difficult to account for potential bias.
* Patient records, clinical data and unstructured text data
	+ Patient records and other text data sources such as treatment plans or notes taken during the examination are in principle suited to train models on. However, there is little to no standardization how to document and store relevant clinical and therapeutic information. Hence, the considerable efforts need to be taken for data retrieval and preprocessing.
* Claims data
	+ Claims data are available on clinic/practice level but as well on insurance level. Including those longitudinal and patient centered data would allow to leverage epidemiological patterns.

None of these data are yet available for public use, but can be made available by our group under data protection restraints. It is not clear how far these data can be made available to AI developers due to data protection and privacy laws.

## AI Output Data Structure

* *outputs to benchmark*
* *ontologies, terminologies*
* *data format*

### Inputs for benchmarking

* Binary outcome labels (e.g. 0/1) on patient and/or on tooth level
* Logits for each class on patient and/or on tooth level
* Pixel-wise classification on patient/image level, e.g. allows to apply IoU (intersection over union)
* Pixel-wise classification on tooth level, may need as a prerequisite a heuristic to assign each pixel blob to a specific tooth

### Data format

* Text based, e.g. csv table
* Base64 encoded image, may be stored in a JSON type data structure

## Test Data Labels

* *label types*
* *ontologies, terminologies*
* *data format*

### Label types

* Binary patient or tooth level labels such as positive/negative, affected/not affected or prevalent/not prevalent
* Bounding boxes, capturing the structure/pathology of interest.
* Pixel-wise segmentation of the structure/pathology of interest.

### Data format

* Binary labels (e.g. 0/1 stored in a text file)
* Bounding box coordinates (xy-pixel coordinate of upper left corner, width, height, angle in rad)
* Segmentation masks. Pixel-blobs masking the structure/pathology of interest. A base64 encoded image (string).

## Scores & Metrics

* *which metrics & scores to use for benchmarking*
* *considering relation to parameters stakeholders need for decision making*
* *considering scores that providers use*
* *considering the scope providers designed their solutions for*
* *considering the state of the art in RCT, statistics, AI benchmarking etc.*
* *considering bias transparency*

### Metrics

* Accuracy
* Sensitivity (Recall)
* Specificity
* F1-score
* AUC
	+ ROC (Receiver operating characteristic curve)
	+ PRC (Precision Recall curve)
* PPV (Positive predictive value aka Precision)
* NPV (Negative predictive value)

### Tooth level metrics for segmentation tasks

The presence of a caries lesion on a tooth is determined by the intersection of the corresponding pixel segmentation and a pixel mask of the corresponding tooth (the masks being generated by human annotation and by the output of a tooth segmentation model).

Quantification of teeth level metrics of caries segmentation was achieved by computing, for every tooth:

* $N\left(Ref∩Pred\right)$: the total number of intersections between pixel blobs in the reference test and those corresponding to model predictions (or dentists’ annotations).
* $N\left(Ref\right)$the number of overlapping free pixel blobs from the reference set.
* $N\left(Pred\right)$the number of overlapping free predicted pixel blobs.

Based on these quantities, the elements of the confusion matrix are defined as follows:

* True Positive:

$$TP=\frac{N\left(Ref∩Pred\right)}{N\left(Ref\right)+N\left(Pred\right)+N\left(Ref∩Pred\right)}$$

* False Positive:

$$FP=\frac{N\left(Pred\right)}{N\left(Ref\right)+N\left(Pred\right)+N\left(Ref∩Pred\right)}$$

* False Negative:

$$FN=\frac{N\left(Ref\right)}{N\left(Ref\right)+N\left(Pred\right)+N\left(Ref∩Pred\right)}$$

* True Negative:

$$TN=\left\{\begin{array}{c}1 if N\left(Ref\right)=N\left(Pred\right)=0\\0 otherwise\end{array}\right\}$$

According to these definitions, a single tooth can exhibit positive values of TP, FN and FP, thereby capturing the different types of model error at the pixel blob level. The different teeth level scores are finally computed by taking the sum of the above defined quantities over all the teeth and images (see illustration in Fig. XX).



***Figure XX.*** Schematic illustration of the computation of teeth-level confusion matrix elements on simple representative cases of prediction and annotation blob occurrences.

## Undisclosed Test Data Set Collection

* *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

* *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

* *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

* *insert here the reports of the different benchmarking runs*

# Discussion

* *Discussion of the insights from executing the benchmarking on*
	+ *external feedback on the whole topic and its benchmarking*
	+ *technical architecture*
	+ *data acquisition*
	+ *benchmarking process*
	+ *benchmarking results*

*○ field implementation success stories*

# Declaration of Conflict of Interest

* *by each contributor to this document*

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