INTERNATIONAL TELECOMMUNICATION UNION

**TELECOMMUNICATION STANDARDIZATION SECTOR**

STUDY PERIOD 2022-2024

**FGAI4H-P-013-A02**

**ITU-T Focus Group on AI for Health**

**Original: English**

| | | |
|---|---|---|
| **WG(s):** | Plen | Helsinki, 20-22 September 2022 |

**DOCUMENT**

| | |
|---|---|
| **Source:** | TG-Histo Topic Driver |
| **Title:** | Att.2 – CfTGP (TG-Histo) [same as Meeting E] |
| **Purpose:** | Engagement |

**Contact:** Frederick Klauschen Email: frederick.klauschen@charite.de
Charite Berlin, Germany

**Abstract:** Calling on members of the medical and artificial intelligence communities with a vested interest in AI for histopathology! Become engaged in the group dedicated to establishing a standardized benchmarking platform for AI for histopathology within the International Telecommunication Union (ITU)/World Health Organization (WHO) Focus Group on "Artificial Intelligence for Health" (FG-AI4H).

This document is the same as FG-AI4H-E-005-A06 (the latest version), reproduced as Meeting N document for easier reference.

## Call for Topic Group Participation: AI for Histopathology

The International Telecommunication Union (ITU)/World Health Organization (WHO) Focus Group on "Artificial Intelligence for Health" (FG-AI4H; https://www.itu.int/go/fgai4h) seeks engagement from members of the medical and artificial intelligence (AI) communities (including clinicians, technologists, entrepreneurs, potential benchmarking data providers, machine learning experts, software developers, researchers, regulators, policy-makers, companies/institutions, and field experts) with a vested interest in shaping the benchmarking process of AI for histopathology.

### 1 About FG-AI4H

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 complexity 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 topic groups. These are concerned with: AI and cardiovascular disease risk prediction, child growth monitoring, dermatology, falls among the elderly, histopathology, neuro-cognitive diseases, ophthalmology (retinal imaging diagnostics), psychiatry, radiotherapy, snakebite and snake identification, symptom checkers, tuberculosis, and volumetric chest computed tomography.

Each topic group agrees upon representative benchmarking tasks in a pragmatic, best-practice approach, which can later be scaled and expanded to similar tasks. Every benchmarking task should address a health problem of relevance (e.g. impacting a large and diverse part of the global

population or challenging to treat) and for which AI technology would provide a tangible improvement relative to the current practice (e.g. better care, results, and/or cost/time effectiveness).

For a rigorous and sound evaluation, undisclosed test data sets must be available (or have to be collected) for each task. All data must be of high quality and compliant with ethical and legal standards. In addition, the data must originate from a variety of sources so that it can be determined whether an AI algorithm can generalize across different conditions, locations, or settings (e.g. across different people, hospitals, and/or measurement devices). The format/properties of the data serving as input to the AI and of the output expected from the AI, as well as the benchmarking metrics are agreed upon and specified by the topic group.

Finally, the AI-to-be-evaluated will be benchmarked with the undisclosed test data on FG-AI4H computing infrastructure. Here, the AI will process single samples of the undisclosed test data set and predict output variables, which will be compared with the "ground truth." The results of the benchmarking will be provided to the AI developers and will appear on a (potentially anonymized) leaderboard.

## 2 Topic group: AI for Histopathology

A topic group is a community of stakeholders from the medical and AI communities with a shared interest in a topic. The objectives of the topic groups are manifold:

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

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

This topic group is dedicated to AI for histopathology. Tumor Infiltrating Lymphocytes (TILs) are emerging as a very promising biomarker in solid tumors such as breast cancer, lung cancer and melanoma. TILs have been shown to be marker of tumor immunogenicity in breast cancer. It is clear that higher levels of TILs are associated with improved prognosis in certain subtypes of breast cancer while their presence indicates a decreased survival in other subtypes. TILs also indicate a higher probability of achieving therapy response in the neoadjuvant setting. Analysis of TILs in residual disease specimens after neoadjuvant therapy has also been shown to have prognostic value. The evaluation of TILs as a biomarker in breast cancer is expected to be extended from the research domain to the clinical setting in the near future. While TILs are normally assessed by manual estimation, efforts are ongoing for the assessment of TILs by image analysis methods. These methods, and among them particularly AI-based methods, are still experimental and not sufficiently documented and standardized for introduction into clinical trial and daily practice.

We therefore propose to establish a data set for the benchmarking of machine learning based tumor cell detection and TIL quantification algorithms.

The assessment of TILs by digital image analysis will be useful for accurate and reproducible diagnostics in the future, because this approach can be used to determine the number of TILs per stromal tissue area as an exact measurement contrary to the approximate semi-quantitative

evaluation suggested at this moment. In the first International Guidelines on TIL-assessment in breast cancer (Salgado et al., Annals of Oncology 2014), an inter-laboratory quality comparison study was proposed to assess the reproducibility and clinical validity of TIL evaluation. Because conventional image analysis approaches, although capable of identifying lymphocytes relatively easily (Wienert et al., 2012, Scientific Reports), have difficulties in robustly detecting tumor cells due to their broad morphological variability, machine learning approaches have been and are currently being developed that allow for a combined detection of both lymphocytes and cancer cells required for accurate TIL scoring (reviewed in Klauschen et al., 2018, Seminars in Cancer Biology).

Currently, no high-quality annotated data sets on TILs in breast cancer are publicly available. We intend to provide a comprehensive histological image data set that allows for the evaluation of image analysis methods for tumor cell and lymphocyte detection and quantitative scoring in breast cancer (Fig. 1A,B). These Hematoxylin&Eosin (H&E) image data will be provided in an *undisclosed fashion* within a compute infrastructure that will be used for the actual benchmarking process.

We will provide a second (disjoint and smaller) data set for public download for participants to assess general features of the data used for benchmarking such as quality, staining and morphological spectrum and to compare these features to local data sets used for training their algorithms. It is important for clinical-grade validation that the data we provide for public download are not sufficient to fully train the developed algorithm de-novo, but that the classifier is benchmarked with a data set independent of that used for training.

In the benchmarking process, the participants are expected to submit AI-based solutions that will analyze the histopathological images and

- automatically detect tumor cells and lymphocytes, and/or
- quantify the lymphocyte and tumor cell density (number of cells per square millimeter in the tumor area or in the border area of the tumor), and/or
- predict the semi-quantitative score as diagnosed by pathologists after visual inspection and comparison with reference images (Salgado et al., Annals of Oncology 2014).

The submissions should be evaluated by comparing the AI-based predictions with the cell-wise manual annotations and scores given by pathologists. Different benchmarking metrics are conceivable including statistical measures such as the detection performance (accuracy, F1 score, area under the curve of the receiver operating characteristic etc.) and the quantification error (e.g., the root mean square error). Explanations in visual form that allow humans to interpret why the AI-algorithm eventually came to a conclusion or made a prediction are additional measures to be considered in the benchmarking procedure (see Fig. 1 C for an example).

More details about the activities of the topic group can be found in the documents B-014-R1 and C-018, which can be accessed with a free ITU account (cf. “Get involved”).

The topic group on AI for histopathology is lead by Frederick Klauschen (Institute of Pathology, Charité Universitätsmedizin Berlin, Germany). The Institute of Pathology runs one of the largest university diagnostics services in histopathology in Germany providing diagnostics for over 80,000 patients and more than 250,000 histological samples per year. With its participation in computational pathology projects at the publicly funded Berlin Big Data Center and the Digital Medicine platform at the Berlin Institute of Health as well as close collaboration with the Technical University Berlin and Heinrich Hertz Institute, the Institute of Pathology integrates the application side of machine learning based image analysis with the scientific and technical developments in this field. Because of increasing requirements to provide more timely, quantitative and standardized diagnostics, it is our core interest to promote and support standardization and benchmarking efforts, which are pivotal to the implementation of machine learning in diagnostics.

We encourage interested parties to join as topic group members. The topic group would benefit from further expertise of the medical and AI communities and from additional data.

## 3 Get involved

To join this topic group, please send an e-mail to the focus group secretariat (tsbfgai4h@itu.int) and the topic driver (frederick.klauschen@charite.de). Please use a descriptive e-mail subject (e.g. "Participation topic group AI for histopathology"), briefly introduce yourself and your organization, concisely describe your relevant experience and expertise, and explain your interest in the topic group.

Participation in FG-AI4H is free of charge and open to all. To attend the workshops and meetings, please visit the Focus Group website (https://www.itu.int/go/fgai4h), where you can also find the whitepaper, get access to the documentation, and sign up to the mailing list.

## References

Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., ... & Perez, E. A. (2014). The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Annals of oncology*, *26*(2), 259-271. https://doi.org/10.1093/annonc/mdu450

Wienert, S., Heim, D., Saeger, K., Stenzinger, A., Beil, M., Hufnagl, P., ... & Klauschen, F. (2012). Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. *Scientific reports*, *2*, 503. https://doi.org/10.1038/srep00503

Klauschen, F., Müller, K. R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., ... & Michiels, S. (2018, July). Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. In Seminars in cancer biology. Academic Press. https://doi.org/10.1016/j.semcancer.2018.07.001

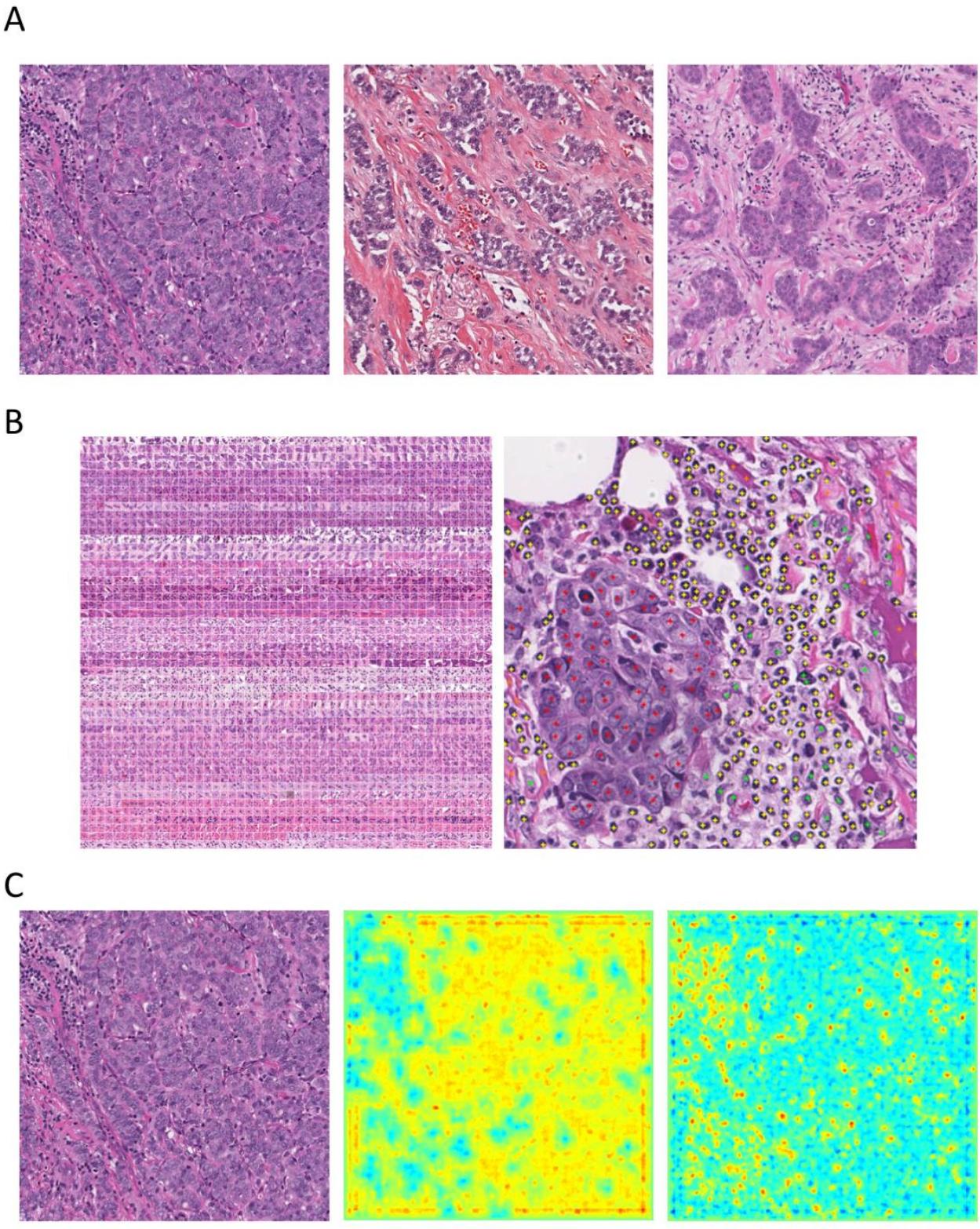


*Figure 1: A) Example breast cancer images used in the benchmark data set showing variable morphology and degree/distribution of tumor-infiltrating lymphocytes. B) Manually annotated tumor cells (red) and stromal cells (green) and lymphocytes (yellow). C) Example of machine learning-based identification of tumor cells (mid panel) and lymphocytes (right) in H&E histology (left) with visual feedback through explanatory heatmaps.*