|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-B-014-R1 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | New York, 15-16 November 2018 | |
| **DOCUMENT** | | | | |
| **Source:** | | Institute of Pathology, Charité Universitätsmedizin Berlin & Berlin Institute of Health & Berlin Big Data Center & Berlin Center for Machine Learning | | |
| **Title:** | | Proposal: Machine learning-based profiling of tumor-infiltrating lymphocytes in breast cancer | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Prof. Dr. Frederick Klauschen Institute of Pathology, Charité Universitätsmedizin Berlin Berlin Institute of Health Berlin Big Data Center (BBDC) Berlin Center for Machine Learning (BZML) Germany | | Tel: +49-30-450536-053 E-mail: [frederick.klauschen@charite.de](mailto:frederick.klauschen@charite.de) |

|  |  |
| --- | --- |
| **Abstract:** | This document proposes a use case for the benchmarking of artificial intelligence methods in histopathological diagnostics with a focus on the identification of tumor cells and scoring of tumor-infiltrating lymphocytes in breast cancer.  Answers to the “proposal submission questionnaire” (FG-AI4H-B-006) are added as appendix in this revised document version. |

**Overview**

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 a reliable and reproducible 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.

**Impact**

Please explain the significance of the problem and describe the potential impact of the project. Please also provide a brief overview of existing work in the area of the project, and describe the current state of the art how the problem is currently addressed.

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).

**Data Availability**

Please describe what data sets would be available for the project. In particular, please describe if there are high quality open data sets for training purposes that are available, and / or if you would be able to contribute to an open data set for training purposes. Please also describe what (undisclosed) test data would be available for an evaluation. For any data set, please describe briefly if and how the data have been annotated.

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.

**Benchmarking**

Please describe what you expect participants in the benchmarking process to submit. Please also describe how the submissions should be evaluated, and why.

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).

**Organizer Details**

Please describe why your organization is interested in this project, and if you have run similar projects / benchmarks / challenges before.

The Institute of Pathology, Charité Universitätsmedizin Berlin 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 digital image analysis projects at the publicly funded Berlin Big Data Center and the Digital Health initiative 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.

**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>

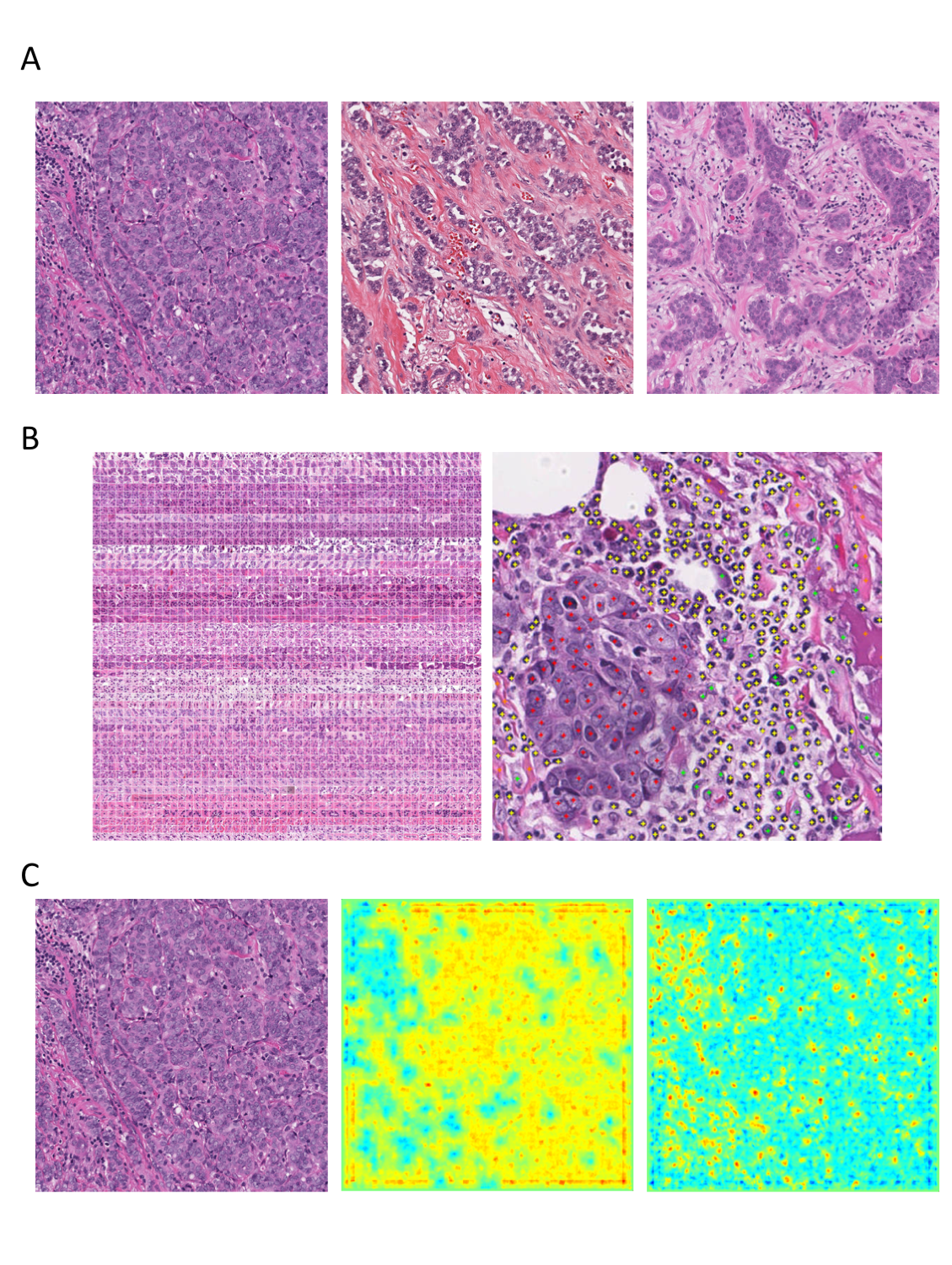


Figure 1: A) Example breast cancer images used in the benchmark data set showing variable morphology and degree/distribution of tumor-infiltrating lymphocytes. B) Manually annotated tumor cells (red) and stromal cells (green) and lymphocytes (yellow). C) Example of machine learning-based identification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

Appendix  
Answers to questionnaire

The “proposal submission questionnaire” (FG-AI4H-B-006) is answered in the following:

1. *Relevance - How relevant is the health problem to be addressed?*

Solid tumors such as breast cancer, lung cancer and melanoma constitute a serious, live-threatening health problem, affecting a large number of people around the globe, and TILs can serve as biomarker for these tumors, as pointed out in detail in the “Overview” and the “Impact” sections.

1. *Impact - What level of impact will a benchmark in the context of the proposed project have?*

Benchmarking of AI-based solutions for TIL scoring will help pathologists to decide if a particular AI-based solution is suited to support them in their work, to compare the performance of different solutions and to learn about their limits.

1. *Existing work - Does the project start from scratch, or are there preliminary experiences?*

The project is based on years of experience in diagnostic and digital pathology. The organizer is part of the “International Immuno-Oncology Biomarker Working Group on Breast Cancer” and has experience in collaborating with machine learning researchers (cf. sections on “Organizer details” and “References”).

1. *Feasibility - Is the project feasible, based on the current state of the art?*

Feasibility has been demonstrated, e.g., in the following state-of-the-art publications:

* 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>
* Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., ... & Klauschen, F. (2018). Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. arXiv preprint arXiv:1805.11178. <https://arxiv.org/abs/1805.11178>

1. *Data Availability - Is there sufficient data available? How much of it can be openly available? How much of it as part of the non-disclosed data set?*

This question is answered in the section on “Data Availability”.

1. *Data Quality - Is the available data of high quality?*

Data are acquired in excellent quality in a state-of-the-art routine diagnostics laboratory by professional and experienced staff.

1. *Annotation / Label Quality - Are the annotations / labels of the data of high quality?*

Annotation and labelling is performed by trained/board-certified pathologists and is therefore of excellent quality.

1. *Data Provenance - Has the data been obtained in a professional and ethically correct way?*

Data are obtained in one of Europe’s largest university hospitals obligated to the highest professional and ethical standards.

1. *Benchmarking - Do the applicants have a clear proposal about what exactly should be evaluated / measured?*

Yes, details of what exactly should be evaluated and measured are explained in the section on “Benchmarking”.

1. *Organizers - Can the Focus Group work with the applicants, and do they have the time / resources to work with the Focus Group on the problem?*

Yes, the Focus Group can work with the organizer on the problem because the organizer has first-hand access to the histological samples, because he is running a routine histopathological diagnostics laboratory, and has professional laboratory personnel and trained pathologists in his team available to work on the project.

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