Putting Guidelines Into Practice: The Audit of a Diabetic Retinopathy Classification Model

Luis Oala, Jana Fehr, Luca Gilli, Pradeep Balachandaran, Alixandro Werneck Leite, Saul Calderon-Ramirez, Danny Xie Li, Gabriel Nobis, Erick Alejandro Muñoz Alvarado, Giovanna Jaramillo-Gutierrez, Christian Matek, Arun Shroff, Ferath Kherif, Bruno Sanguinetti, Thomas Wiegand

SAIL 2022

12/20

A colorful smorgasbord of initiatives such as STARD-AI [1], CONSORT-AI [2], and the WHO/ITU FG-AI4H created guidelines for transparent assessment of ML4H performance. Integrating these guidelines into the ML development process to meet technical, ethical, and clinical requirements is challenging. While there appears to be no shortage in good practice guidelines on paper, the question on how well they can be adopted in practice remains unanswered. We applied the ITU/WHO FG-AI4H guidelines [3] (process depicted in figure 1) with the following quality assessment dimensions on three ML4H use cases.

	❶ Transmit	❷ Understand	❸ Audit	❹ Report
ITU/WHO FG-AI4H Reference Documents	Training and Test Data Specification (DEL 5.4) Data Requirements (DEL 5.1) Data Handling (DEL 5.5) Data Sharing (DEL 5.6)	Data Annotation Specification (DEL 5.3) Data Acquisition (DEL 5.2) Topic Description Document (TDD) (DEL 10.x) Model Questionnaire (J-038)	Ethics Consideration (DEL 1) Regulatory Considerations (DEL 2.2) Clinical Evaluation (DEL 7.4) Assessment Methods Reference (DEL 7.3)	Reporting Template (J-048)
Actors	Use Case Owner	Test Engineers	Test Engineers, Use Case Owner	Test Engineers

Fig 1: A flow chart of the FG-AI4H assessment process and its reference documents

ML4H use-case

Task	Classification type	Outcome
Diagnostic	binary	Diabetic Retinopathy yes/no

Quality Dimensions

- Transparent model reporting
- Bias and fairness (*aequitas*)
- Robustness under input perturbations
- Interpretability

Assessment results - a selection

06/21
SAIL (panned)

Diagnostic prediction of Diabetic Retinopathy

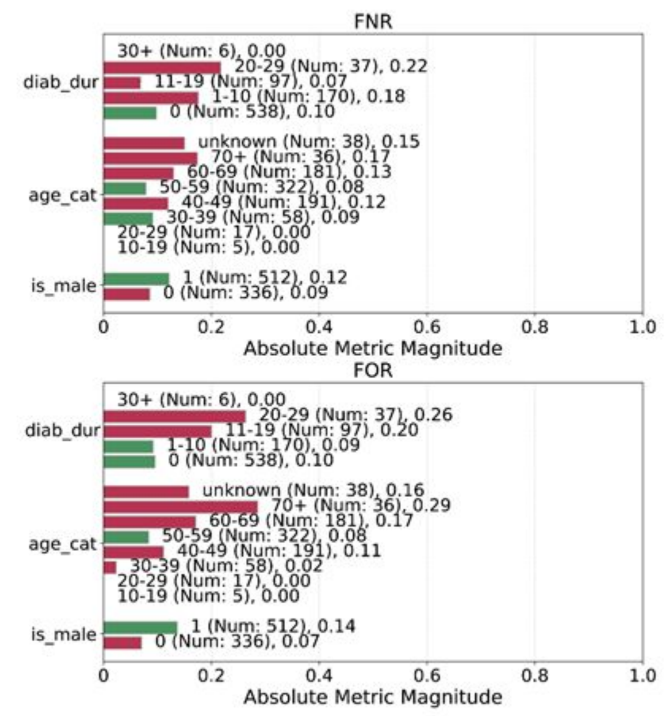


Fig 2: False Negative Rate (FNR) and False Omission Rate (FOR) stratified across diabetes duration, age and gender groups. Red marks significant disparities compared to reference

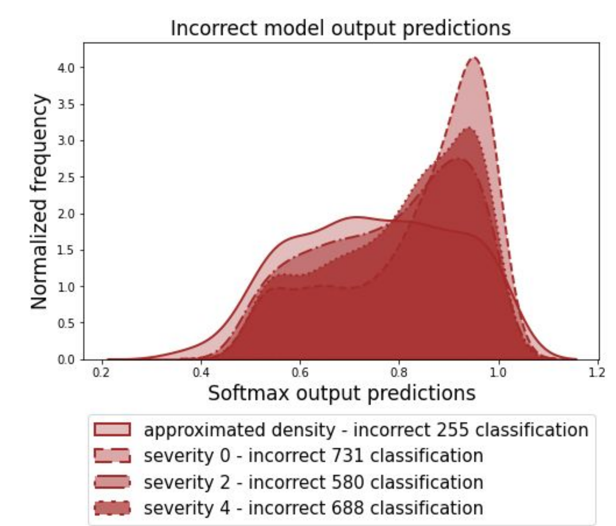


Fig 3: Model output frequency shift after perturbing the input image with jpeg compression

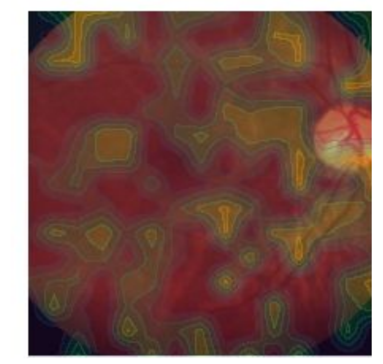


Fig 4: Interpreting model predictions through meaningful perturbations

Reporting caveats for model use

Diagnostic prediction of Diabetic Retinopathy (DR)

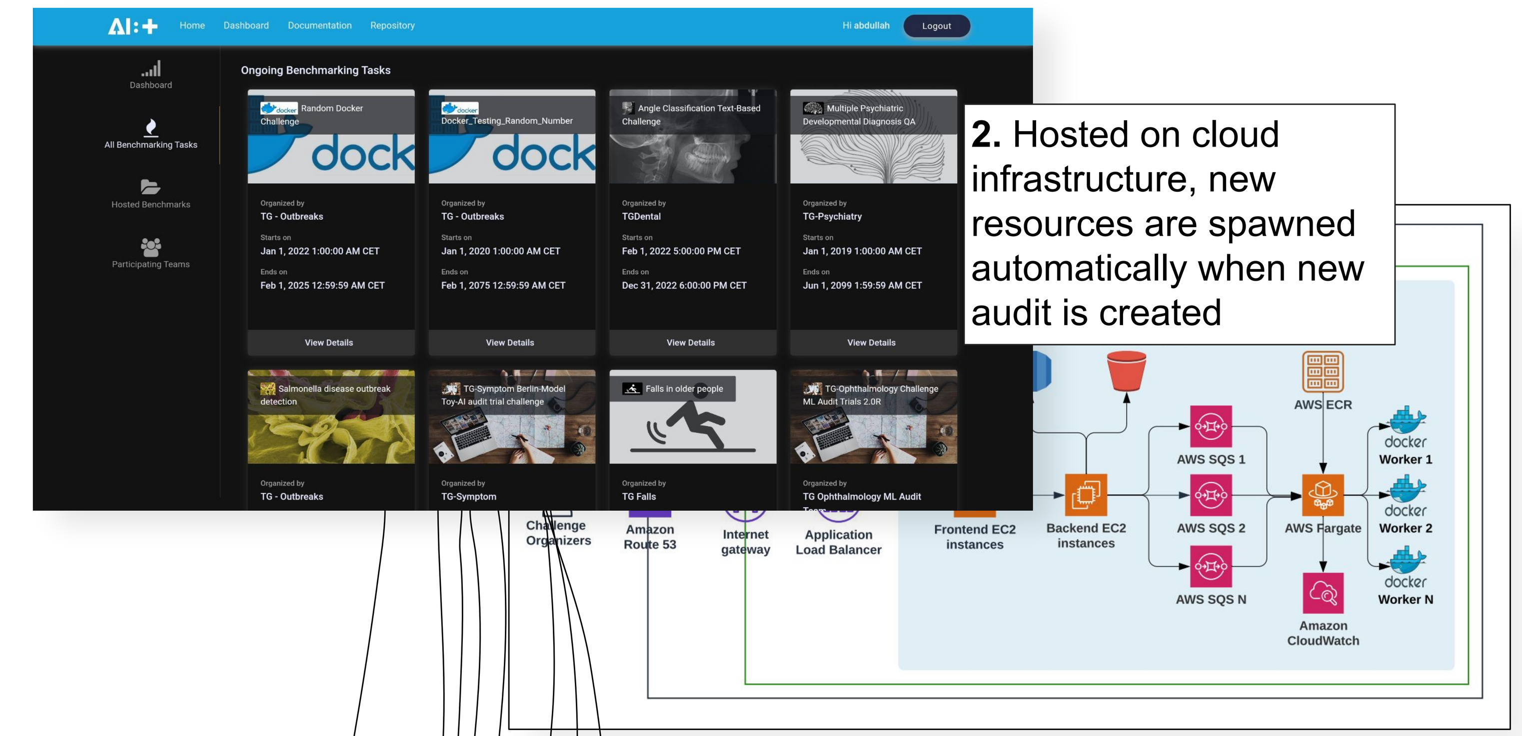
- Binary classification of DR vs. normal does not suffice the model's intended use of 'Detecting early signs of DR'
- Applicable healthcare context is unknown (hospital, routine care?)

Assessment challenges

- No one-size fits all assessment framework
  Appropriate assessment methods need to be selected according to...
  - data
  - model tasks
  - model development libraries (e.g. tf version of model vs supported tf version of assessment tool)
- Bias and fairness analysis is often implicated by lack of data variables such as age, stage of disease, hospital specialization, … due to lack of collection or data protection
- Robustness analysis lacks realistic perturbations specific to the medical image domain
- Interpretability methods explain systematic model mistakes only limitedly

health.aiaudit.org

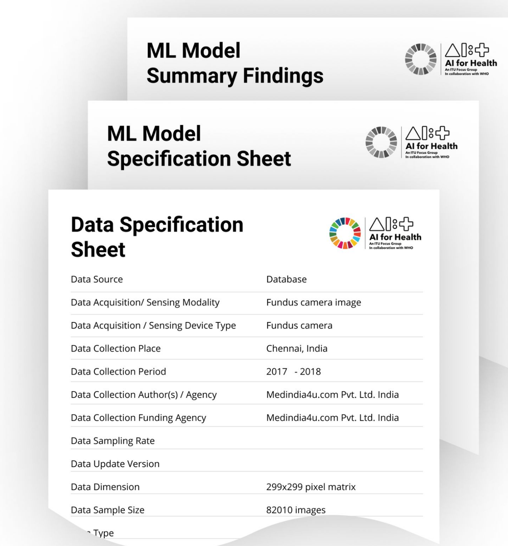
A platform for automated, standardized assessment and reporting of ML4H applications



2. Hosted on cloud infrastructure, new resources are spawned automatically when new audit is created

1. Currently eight interdisciplinary teams (ML, clinical, regulatory, ethics)

3. Any stakeholder can pull standardized report of the audit results



05/22
What happened since

Come contribute

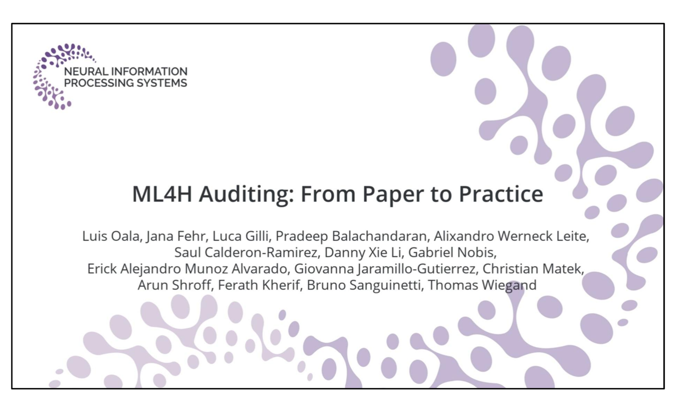
future

- We have cloud resources
- A small, full-time core developer team
- A collaborative and fun community around it all

t

luis.oala@hhi.fraunhofer.de

See our ML4H 2020 Spotlight for full results



Oala, L., Fehr, J., Gilli, L., Balachandran, P., Leite, A.W., Calderon-Ramirez, S., Li, D.X., Nobis, G., Alvarado, E.A.M., Jaramillo-Gutierrez, G. and Matek, C., 2020, November. Ml4h auditing: From paper to practice. In *Machine Learning for Health* (pp. 280-317). PMLR.

References

[1] Viknesh Sounderajah *et al.* (2020) Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. In: Nature Medicine 26, pp. 807–808

[2] Xiaoxuan Liu *et al.* (2020) CONSORT-AI extension. In: Nature Medicine 26, pp. 1364–1374.

[3] Wiegand, T. *et al.* (2019) WHO and ITU establish benchmarking process for artificial intelligence in health. In: Lancet 394, pp. 9–11.

Fraunhofer HHI · Digital Health Center by Hasso-Plattner-Institut · clearbox AI humans && machines · UnB · De Montfort University Leicester · TEC Tecnológico de Costa Rica · HelmholtzZentrum münchen Deutsches Forschungszentrum für Gesundheit und Umwelt · LMU · XTend.AI · UNIL | Université de Lausanne · DOT PHOTON