

ML4H Audit: From Paper to Practice

The FG-AI4H Assessment Process

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.

	1 Transmit	2 Understand	3 Audit	4 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


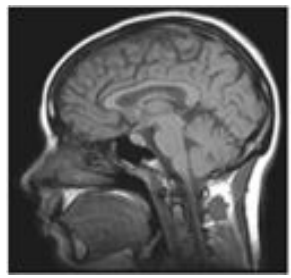
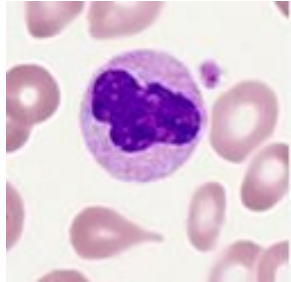
Fig 1: A flow chart of the FG-AI4H assessment process and its reference documents

ML4H use-cases

We assessed these three use cases with the following quality dimensions, based on the FG-AI4H framework

Quality Dimensions

- Transparent model reporting
- Bias and Fairness (aequitas)
- Robustness under perturbations
- Interpretability

Task	Classification type	Outcome	
Diagnostic	binary	Diabetic Retinopathy yes/no	
Diagnostic	binary	Alzheimer's Disease yes/no	
Cell characterization	multi-class	Cytomorphologic cell classification (15 classes)	

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

Assessment results - a selection

Diagnostic prediction of Diabetic Retinopathy

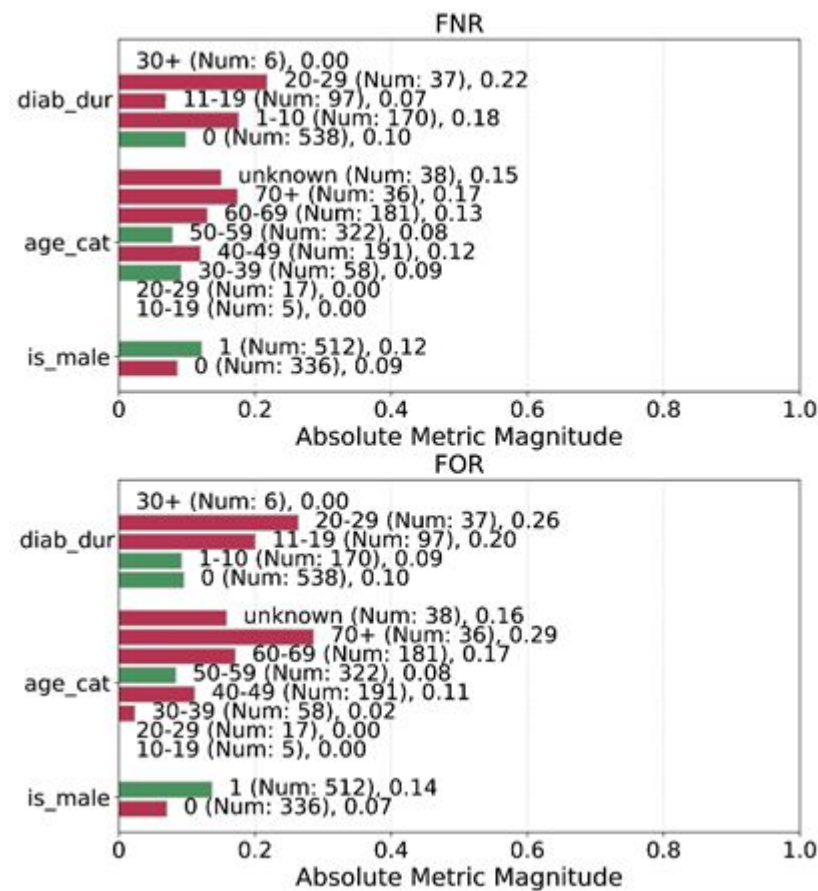


Fig 2: False Negative Rate (FNR) and False Omission Rate (FOR) stratified across diabetes duration, age and gender groups. Red marks significant disparities compared to reference

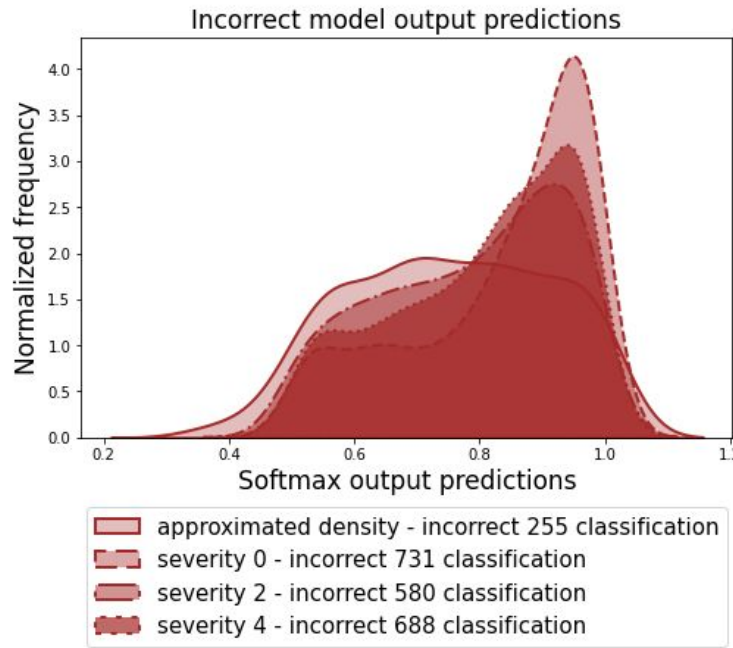


Fig 3: Model output frequency shift after perturbing the input image with jpeg compression

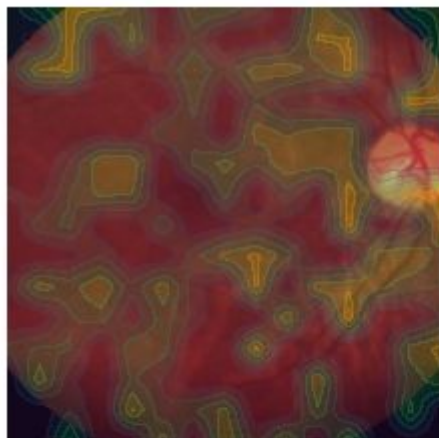


Fig 4: Interpreting model predictions through meaningful perturbations

Diagnostic prediction of Alzheimer's Disease

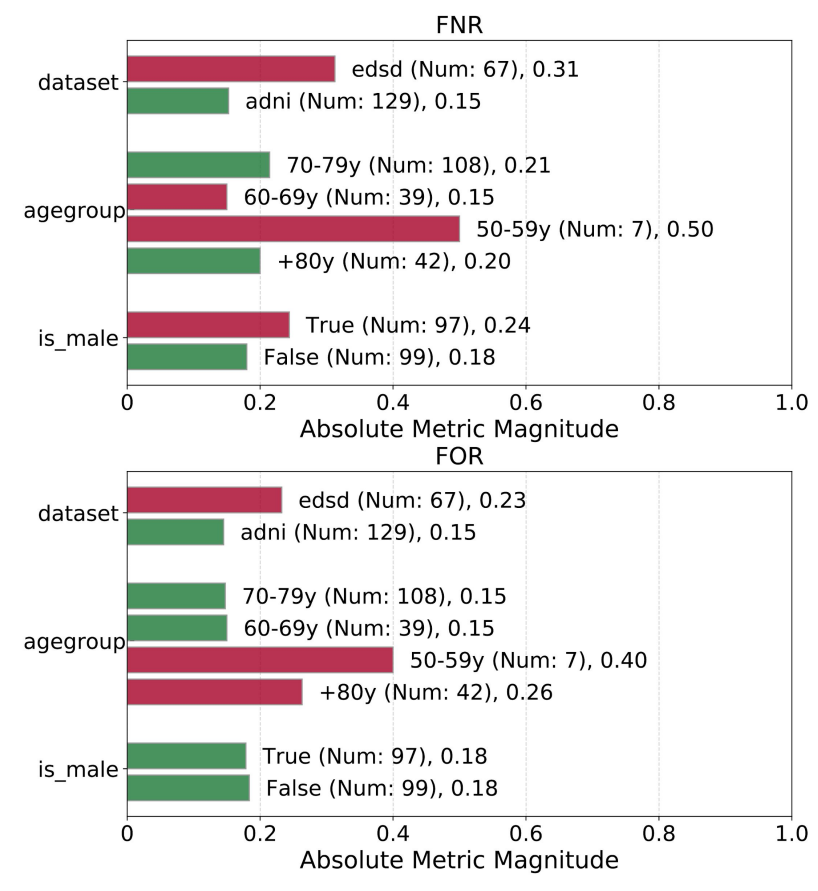


Fig 5: False Negative Rate (FNR) and False Omission Rate (FOR) stratified across dataset, age and gender groups. Red marks significant disparities compared to reference

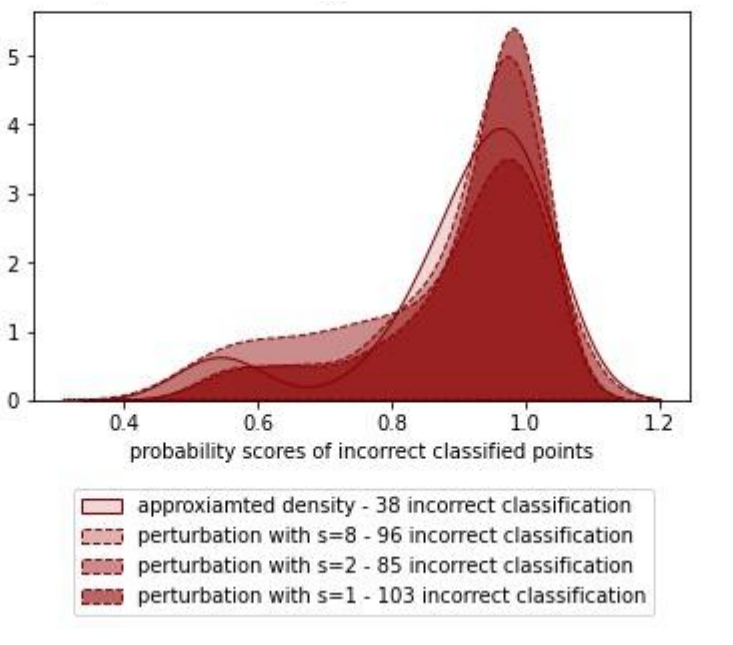


Fig 6: Model output frequency shift after perturbing age and brain volume features with lognormal noise

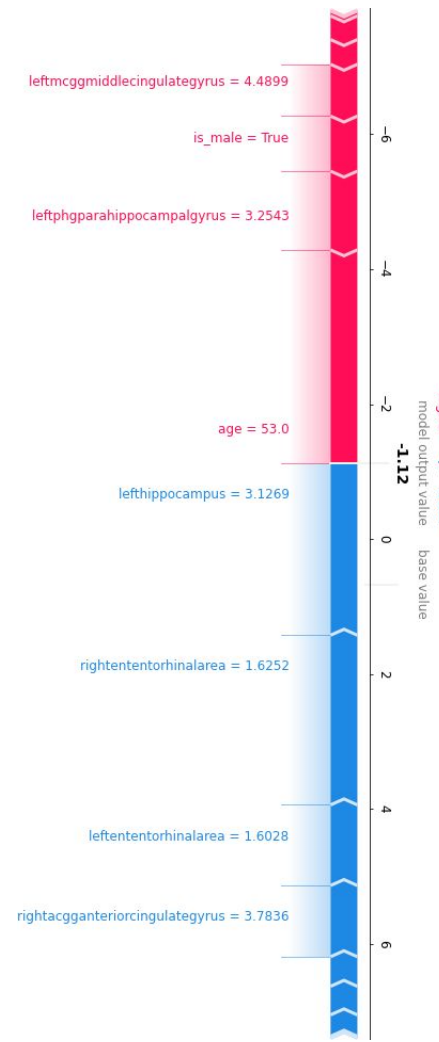


Fig 7: Representative SHAP explanation for a False Negative prediction in the 50-59 age group

Cytomorphologic cell classification for Leukemia diagnostics

\* Dataset and patient variable stratification not applicable for this use case

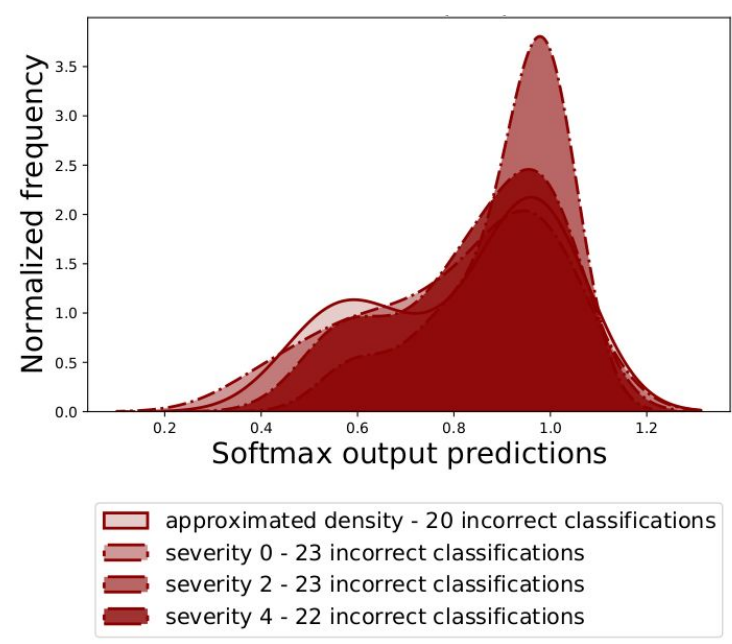


Fig 8: Model output frequency shift after perturbing the input image with jpeg compression

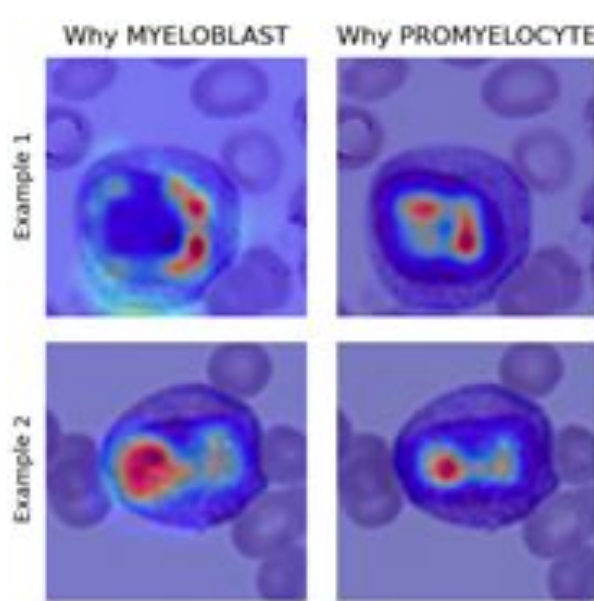


Fig 9: Grad-Class Activation map to interpret how the model obtained predictions



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?)

Diagnostic prediction of Alzheimer's Disease (AD)

- Binary classification of AD vs. normal does not suffice the model's intended use of 'Detecting early signs of AD'
- Not evaluated in populations with different AD prevalence

Cytomorphologic cell classification

- Generalizability unknown as model was trained on data from one hospital only
- Model performance depends on class-frequency

Assessment challenges

- No one-size fits all assessment framework  
Appropriate assessment methods need to be selected according to...
  - data
  - model tasks
  - Model development libraries
- 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

Interested?

luis.oala@hhi.fraunhofer.de

<https://forms.gle/BwdGP98hvKSRYBxy8>

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.