

ML4H Auditing: From Paper to Practice

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

Spotlight Presentation
Machine Learning for Health (ML4H) Workshop
Conference on Neural Information Processing Systems
December 11, 2020



Collaborators



Luis



Jana



Luca



Pradeep



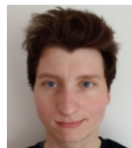
Alixandro



Saul



Danny



Gabriel



Erick



Giovanna



Christian



Arun



Ferath



Bruno



Thomas

Balancing act in ML4H

Exploratory excitement and rigorous assessment of efficacy and safety

A smorgasbord of guidelines

STARD-AI [Sou+20], CONSORT-AI [Liu+20], SPIRIT-AI [Liu+20], Focus Group on Artificial Intelligence for Health (FG-AI4H) [Wie+19], FDA [US-19], International Medical Device Regulators Forum (IMDRF) [IMD19], implications [He+19], data [Geb+18], model development [Mit+19]; [Sen+20], ...

Paper-to-practice gap

Abstract guidelines are available but they are not being applied routinely → lack of feedback and harmonization, process vacuum

	① Transmit	② Understand	③ Audit	④ Report →
ITU/WHO FG-AI4H Reference Documents	<p>Training and Test Data Specification (DEL 5.4)</p> <p>Data Requirements (DEL 5.1)</p> <p>Data Handling (DEL 5.5)</p> <p>Data Sharing (DEL 5.6)</p>	<p>Data Annotation Specification (DEL 5.3)</p> <p>Data Acquisition (DEL 5.2)</p> <p>Topic Description Document (TDD) (DEL 10.x)</p> <p>Model Questionnaire (J-038)</p>	<p>Ethics Consideration (DEL 1)</p> <p>Regulatory Considerations (DEL 2.2)</p> <p>Clinical Evaluation (DEL 7.4)</p> <p>Assessment Methods Reference (DEL 7.3)</p>	<p>Reporting Template (J-048)</p>
Actors	Use Case Owner	Test Engineers	Test Engineers, Use Case Owner	Test Engineers

Diabetic retinopathy



retina images
 224×224

CNN

binary classif.
(retinopathy yes/no)

Alzheimer's disease



structured data
 1×16

Gradient Boosting

binary classif.
(AD and CN)

Leukemia

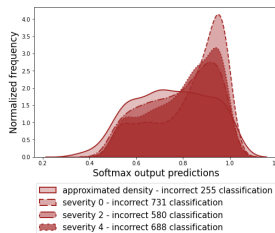
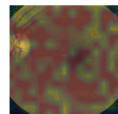
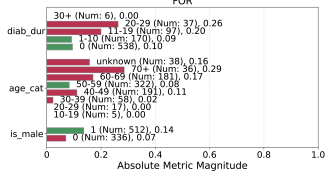
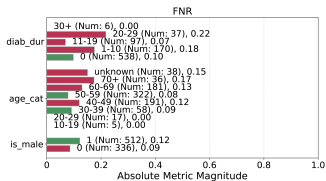


single-cell images
 400×400

CNN

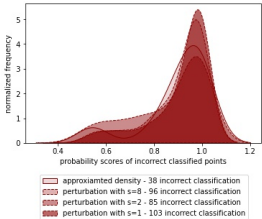
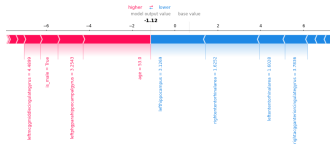
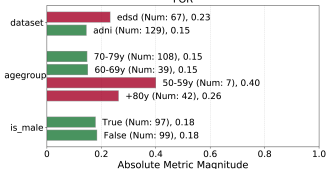
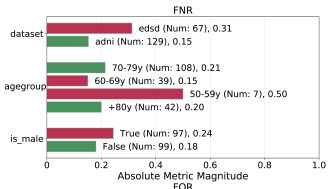
multi-class classif.
(15 morph. categories)

Results - Diabetic retinopathy



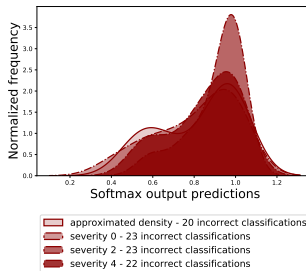
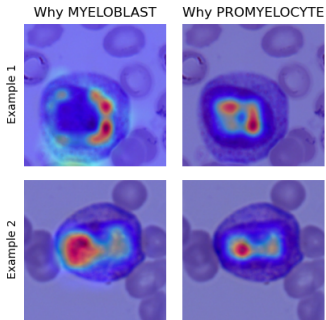
- ▶ Individuals with 30+ years of diabetes were underrepresented (n=6), unfair FNR and FOR distributions were found for the age groups of 70, 60-69, 40-49 and unknown
- ▶ Model does not focus on relevant image elements such as arteries
- ▶ Low entropy score distribution for misclassifications under JPEG compression

Results - Alzheimer's disease

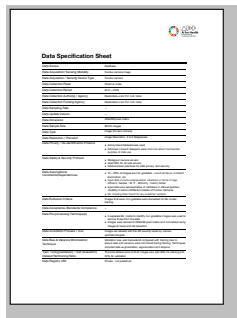


- ▶ FNR and FOR unfairness was indicated for 50-59 year old individuals, the age group of 50-59 years comprised only 7 individuals
- ▶ Age or gender features were never anchoring conditions, SHAP value for age in the group 50-59 years was pushing the prediction towards a positive prediction
- ▶ Low entropy score distribution for misclassifications under lognormal noise, challenge of meaningful perturbations

Results - Leukemia

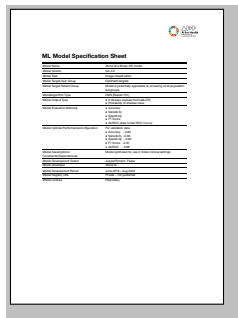


- ▶ No bias assessment with *aequitas* possible as no metadata was available
- ▶ Model learned to focus on the leukocyte's nucleus and cytoplasm, while ignoring erythrocytes and other background structures
- ▶ JPEG compression increased output confidence of misclassifications, which may reflect the general calibration behaviour of the network type [Guo+17]



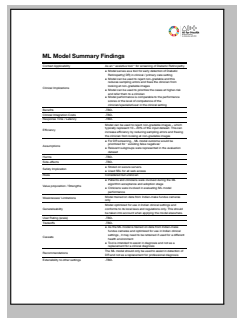
Data Specification Sheet

This form contains various fields for data specification, including sections for Data Source, Data Format, Data Access, and Data Security. It includes checkboxes and text input areas for detailed reporting.



ML Model Specification Sheet

This form contains various fields for ML model specification, including sections for Model Description, Model Performance, and Model Security. It includes checkboxes and text input areas for detailed reporting.



ML Model Summary Findings

This form contains various fields for ML model summary findings, including sections for Model Summary, Model Findings, and Model Recommendations. It includes checkboxes and text input areas for detailed reporting.

Full report cards can be accessed under identifier
FGAI4H-J-049-A01 at

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx>

What we need

- ▶ Keep assessment frameworks flexible to accommodate dependencies wrt data, model and task
- ▶ Maintain meta-information for datasets to enable bias, fairness and robustness analysis
- ▶ Domain-specific synthesis of meaningful robustness testing data

Next steps

- ▶ Towards in-silico testing of clinical endpoints
- ▶ Expanding audits to more use cases
- ▶ Automation of audit process through an assessment platform

Call for Collaboration



<http://www.itu.int/go/fgai4h>

(1) Have your use case audited

- ▶ FG-AI4H already has 20+ use case groups

(2) Become an auditor

- ▶ Write papers with us analyzing vulnerabilities of ML4H models
- ▶ Join one of the specialized research projects (e.g. development of automated assessment platform, measurement specific robustness benchmarks, OOD data and generalization)

Contact me ☺

luis.oala@hhi.fraunhofer.de

<https://forms.gle/BwdGP98hvKSRYBxy8>

[luisoala.github.io](https://github.com/luisoala)



Timnit Gebru et al. “Datasheets for Datasets”. In: (2018), pp. 1–28. arXiv: 1803.09010. URL: <http://arxiv.org/abs/1803.09010>.



Chuan Guo et al. “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1321–1330.



Jianxing He et al. “The practical implementation of artificial intelligence technologies in medicine”. In: *Nature medicine* 25.1 (2019), pp. 30–36.



IMDRF. *Artificial Intelligence in Healthcare Opportunities and Challenges*. 2019. URL: <http://www.imdrf.org/docs/imdrf/final/meetings/imdrf-meet-190916-russia-yekaterinburg-14.pdf>.



Xiaoxuan Liu et al. “CONSORT-AI extension”. In: *Nature Medicine* 26.September (2020), pp. 1364–1374. DOI: 10.1038/s41591-020-1034-x.



Margaret Mitchell et al. “Model cards for model reporting”. In: *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* Figure 2 (2019), pp. 220–229. DOI: 10.1145/3287560.3287596. arXiv: 1810.03993.



Mark P. Sendak et al. “Presenting machine learning model information to clinical end users with model facts labels”. In: *npj Digital Medicine* 3.1 (2020). ISSN: 2398-6352. DOI: 10.1038/s41746-020-0253-3. URL: <http://dx.doi.org/10.1038/s41746-020-0253-3>.



Viknesh Sounderajah et al. “Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group”. In: *Nature Medicine* 26.6 (2020), pp. 807–808. ISSN: 1546170X. DOI: 10.1038/s41591-020-0941-1. URL: <http://dx.doi.org/10.1038/s41591-020-0941-1>.



US-FDA. *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SAMD)—discussion paper and request for feedback.* 2019. 2019.



Thomas Wiegand et al. “WHO and ITU establish benchmarking process for artificial intelligence in health”. In: *The Lancet* 394.10192 (2019), pp. 9–11. ISSN: 1474547X. DOI: 10.1016/S0140-6736(19)30762-7.