ML4H Auditing: From Paper to Practice

Luis Oala luis.oala@hhi.fraunhofer.de
Machine Learning Group, Fraunhofer HHI, Germany

Jana Fehr jana.fehr@hpi.de
Machine Learning and Digital Health, Hasso-Plattner-Institute, Germany

Luca Gilli luca@clearbox.ai
Clearbox AI Solutions, Italy

Pradeep Balachandran pbn.tvm@gmail.com
Technical Consultant (Digital Health), India

Alixandro Werneck Leite alixandrowerneck@outlook.com
Machine Learning Laboratory in Finance and Organizations, Universidade de Brasília, Brazil

Saul Calderon-Ramirez sacalderon@itcr.ac.cr
Centre for Computational Intelligence, De Montfort University, United Kingdom

Danny Xie Li dxie@ic-itcr.ac.cr
Instituto Tecnológico de Costa Rica, Costa Rica

Gabriel Nobis gabriel.nobis@hhi.fraunhofer.de
Machine Learning Group, Fraunhofer HHI, Germany

Erick Alejandro Muñoz Alvarado erickmatec@estudiantec.cr
Instituto Tecnológico de Costa Rica, Costa Rica

Giovanna Jaramillo-Gutierrez gjgutierrez@protonmail.com
Milan and Associates SPRL, Spain

Christian Matek christian.matek@helmholtz-muenchen.de
Department of Medicine III, LMU Klinikum and Institute of Computational Biology, Helmholtz Zentrum München, Germany

Arun Shroff arunshroff@gmail.com
xtend.ai, U.S.

Ferath Kherif ferath.kherif@chuv.ch
Laboratory for Research in Neuroimaging, Lausanne University Hospital and University of Lausanne, Switzerland

Bruno Sanguinetti bruno.sanguinetti@dotphoton.com
Dotphoton AG, Switzerland

Thomas Wiegand thomas.wiegand@hhi.fraunhofer.de
TU Berlin and Fraunhofer HHI, Germany

Editors: Emily Alsentzer$^{\otimes}$, Matthew B. A. McDermott$^{\otimes}$, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy‡, Stephanie L. Hyland‡

Abstract

Healthcare systems are currently adapting to digital technologies, producing large quantities of novel data. Based on these data, machine-learning algorithms have been developed to support practitioners in labor-intensive workflows such as diagnosis, prognosis, triage or treatment of disease. However, their translation into medical practice is often hampered by a lack of careful evaluation in different settings. Efforts have started worldwide to establish guidelines for evaluating machine learning for health (ML4H) tools, highlighting the necessity to evaluate models for bias, interpretability, robustness, and possible failure modes. However, testing and adopting these guidelines in practice remains an open challenge. In this work, we target the paper-to-practice gap by applying an ML4H audit framework proposed by the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) to three use cases: diagnostic prediction of diabetic retinopathy, diagnostic prediction of Alzheimer's disease, and cytomorphologic classification for leukemia diagnostics. The assessment comprises dimensions such as bias, interpretability, and robustness. Our results highlight the importance of fine-grained and case-adapted quality assessment, provide support for incorporating proposed quality assessment considerations of ML4H during the entire development life cycle, and suggest improvements for future ML4H reference evaluation frameworks.

Keywords: Machine Learning, Health, Testing

Healthcare systems generate large amounts of data from medical imaging, sensors, or electronic health records. Modern machine learning technology has been developed to analyze big data for health, promising to reduce cost and labor for diagnostics and prognostics in different medical fields (Topol, 2019; Esteva et al., 2019). A sprawling ecosystem comprising academic, corporate, and institutional capital has produced numerous machine learning for health (ML4H) use cases such as detecting diabetic retinopathy in retina images (Gulshan et al., 2016) or predicting Alzheimer's disease from MRI images (Moradi et al., 2015). The exploratory excitement is matched by demands for a rigorous assessment of efficacy and safety as is standard protocol for any technological innovation in healthcare. Thus, a colorful smorgasbord of initiatives such as STARD-AI (Sounderajah et al., 2020), CONSORT-AI (Liu et al., 2020), SPIRIT-AI (Liu et al., 2020) and the World Health Organization (WHO)/International Telecommunication Union (ITU) Focus Group on Artificial Intelligence for Health (FG-AI4H) (Wiegand et al., 2019) created guidelines for transparent assessment of ML4H performance. Integrating these guidelines into the machine learning (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. In this paper, we report findings of applying the ITU/WHO FG-AI4H assessment guidelines to three ML4H use cases: 1) diagnostic prediction of Diabetic Retinopathy (DR), 2) diagnostic prediction of Alzheimer's disease (Alzheimer's Disease (AD)) and 3) cytomorphologic cell classification to aid leukemia diagnostics. We find that the protocol helps to uncover model weaknesses while also posing challenges during its application. We comment on these challenges and provide suggestions for further research directions in this field.

1. Related work and contributions

The aspiration towards reliable and safe ML4H products has led to multiple initiatives that provide guidance for quality assurance. Next to the initiatives mentioned in the introduction, these include the FDA (US-

FDA, 2019), International Medical Device Regulators Forum (IMDRF) (IMDRF, 2019), academic voices like He et al. (2019) and others summarized in (Wenzel and Wiegand, 2020). In addition, the academic ML community proposed standardized reporting for data resources (Gebru et al., 2018) and model development (Mitchell et al., 2019; Sendak et al., 2020; Raji et al., 2020). Furthermore, specialized software packages have emerged, often focusing on either a specific quality aspect, such as the AI Fairness 360 toolkit (Bellamy et al., 2018), a particular step in the ML life cycle, such as robust model training (Nicolae et al., 2018) or benchmarking Papernot et al. (2018); Hendrycks and Dietterich (2019).

However, to the best of our knowledge no work exists on applying a ML4H quality assessment guideline to existing use cases. This is what we call the paper-to-practice gap: abstract guidelines and specialized tools are available but they are not being applied. Decreasing the paper-to-practice gap is the main motivation of this work. Our team of test engineers collaborated with three clinical use case owners to carry out a trial audit following the FG-AI4H assessment guideline. We report implications of applying ML4H quality assessment guidelines in practice, audit results as well as positive and negative experiences with the FG-AI4H process. Our report can help other use case owners and stakeholders from ML4H quality assessment initiatives to improve their work.

2. Use Cases

In a call for participation, we chose three use cases based on the developers' capacity to share the model and testing data from different input modalities, that is, imaging and structured data. The use case models were developed by teams in the United States, India, Switzerland and Germany.

(1) Diagnostic prediction of diabetic retinopathy Diabetic retinopathy is a sight-threatening complication of diabetes in adults (Yau et al., 2012). Retinal scanning during routine diabetes care is crucial to detect the onset of DR at an early stage (Namperumalsamy et al., 2003). A company based in the United States and India has developed a deep neural network (DNN)-based algorithm to differentiate normal retina fundus images from those with DR using a Resnet-101. More information can be found in A.1.

(2) Diagnostic prediction of Alzheimer's disease AD is a neurodegenerative disease and the most common cause of dementia affecting millions of patients worldwide. AD causes progressive cognitive decline and ultimately death (Dubois et al., 2009a). Recently the combination of objective brain measures from MRI with ML has shown promising results with better diagnostic accuracy than the symptoms based approaches (Dubois et al., 2009b). A group of neuroscientists in Switzerland developed Gradient Boost and Naïve Bayes models on structured data containing Magnetic Resonance (MR)-extracted brain region volumes, age and gender for the classification of AD and cognitively normal older adults. More information in A.2.

(3) Cytomorphologic classification for leukemia diagnostics Morphological examination of leukocytes is a laborious and time-consuming process in the diagnostic routine for hematological disorders. A team based in Germany trained a DNN-based algorithm on two tasks (i) classifying single-cell images into a 15-category morphological scheme derived from routine diagnostics using a ResNeXt model (Xie et al., 2017) and (ii) answering clinically relevant binary questions on the blast character and atypicality of a given single-cell image (Matek et al., 2019). More information in Appendix A.3.

3. Method

For the audit, we chose the FG-AI4H guideline as evaluation framework. This choice was motivated by the following reasons. First, the

	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

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

FG-AI4H framework contains a rich documentation and is targeted specifically at ML4H. Second, the FG-AI4H reference documentation is open-source enabling any stakeholder to use and contribute to the guideline. Finally, due to the patronage of ITU and WHO the FG-AI4H is closely aligned with the Sustainable Development Goals. As summarized in Figure 1, the FG-AI4H process provides reference documentation along four steps of the assessment life cycle: data and model transmission (step 1), data and model understanding (step 2), audit (step 3), and reporting (step 4). In this process, use case owners and test engineers work together towards a final report for the use case assessment. We provide a more detailed introduction to individual reference documents in Appendix D. In the following, we focus on the reference documentation that featured most prominently in the audit simulation.

Questionnaire (J-038)[1]**.** The FG-AI4H process provides a structured questionnaire FG-AI4H (2020b) to elicit important information from developers about the intended use, data, and development process of the algorithms. This information is crucial to determine if the model output fits to the intended use and if sources of bias are present. An additional aim of the questionnaire is to facilitate communication between developers, medical domain experts and test engineers. The questionnaire is organized into seven sections, covering different aspects of model concerns: basic information, intended use, data used for development, legal aspects, ethical considerations, evaluation metrics, and caveats and recommendations. The questions for these sections were adapted from proposed reporting standards 'Modelcards' by Mitchell et al. (2019), 'Datasheets' by Gebru et al. (2018), and 'Model Facts' by Sendak et al. (2020). Each model developer completed this questionnaire.

Bias and fairness (DEL 7.3 - Sections 5.1 and 6.2)[1] According to FG-AI4H (2020a), we performed a quantitative bias analysis on test sets with *aequitas*, an open-source bias and fairness audit toolkit for *Python* (Saleiro et al., 2018). *Aequitas* quantifies bias through disparities of group metric values compared to a reference group. For each attribute, we selected the group with most samples as reference. Reference groups always have a disparity of 1. Disparities close to 1 indicate high similarity with the reference group. *Aequitas* defines fairness if disparities lie within the range of 0.8-1.25 and 'unfairness' otherwise. We performed the *aequitas* assessment on diagnostic use cases for

[1] The identifier in parentheses signifies the FG-AI4H document for easier reference

Alzheimer's disease and diabetic retinopathy. We calculated the group metrics false negative rate (FNR), false omission rate (FOR), negative predictive value (NPV), precision and predicted prevalence (PPREV) with aequitas and considered FNR and FOR as most relevant for diagnostic tasks where false-negatives should to be avoided. We report model performances on holdout sets with accuracy, sensitivity, specificity, positive predictive value (PPV), and NPV indicating 95% confidence intervals (CI). The analysis was not done for the leukemia use case as there were no available metadata to calculate metrics across groups.

Interpretability (DEL 7.3 - Section 6.1)[1] We applied a procedure recommended in FG-AI4H (2020a); Samek et al. (2019) and proposed in Lapuschkin et al. (2019) to obtain most representative model explanations. These can be analysed by model developers to identify undesired behaviour or data issues (Pfau et al., 2019). For each use case, we generated explanations for the test set and grouped them by type, that is their position in the confusion matrix. The explanations for the DR model were generated using the *Meaningful Perturbation* approach (Fong and Vedaldi, 2017). For the AD model, we generated explanations using SHAP values (Lundberg et al., 2020) and Anchors (Ribeiro et al., 2018). Finally the explanations for the cytomorphologic classification model were generated using the Grad-CAM method (Selvaraju et al., 2019). After obtaining model explanations, we applied DBSCAN clustering (Ester et al., 1996) to select most representative explanations. Together with model owners, we assessed if such explanations highlight any model issues. Details about this assessment can be found in C.3.

Robustness (DEL 7.3 - Sections 6.3 and 6.5)[1] Robustness refers to the ability of a ML model to maintain its performance under perturbations to the input (Li, 2018). It is important to assess if random variations that may occur in real-world applications affect the model performance. Different types of perturbations can be tested, but selecting the most realistic ones depend on the specific use case (Hendrycks and Dietterich, 2019). Following FG-AI4H (2020a), we determine for each use case the mean corruption error (mCE) defined in Hendrycks and Dietterich (2019) and report the variation of the probability scores due to several perturbation methods. Furthermore, we document the changes in approximated output score densities as a proxy for a model's sensitivity to the noise perturbations. Details on the robustness assessment can be found in C.4.

Reporting (J-048)[1] Regular exchange between the use case owners and test engineers facilitated the auditing process, giving test engineers the opportunity to obtain additional information for a particular use case, and providing use case owners feedback on potential model weaknesses. Finally, the results from the audit process are documented in a harmonized reporting template FG-AI4H (2020c) that is adapted from existing good practices (Gebru et al., 2018; Mitchell et al., 2019; Sendak et al., 2020). The template integrates information from the questionnaire, technical analysis, and interviews. It provides a common interface for the communication between use case owners and other stakeholders.

4. Experiments

Each use case went through the entire audit. This produced a rich repository of results and documentation. In the following, we present highlights, selected on the premise of exposing model weaknesses as well as insightful feedback from exchanges with the use case owners. Additional results are contained in Appendix B and Appendix C. All results are collected in a report card like Figure 5. The complete outputs from the audit

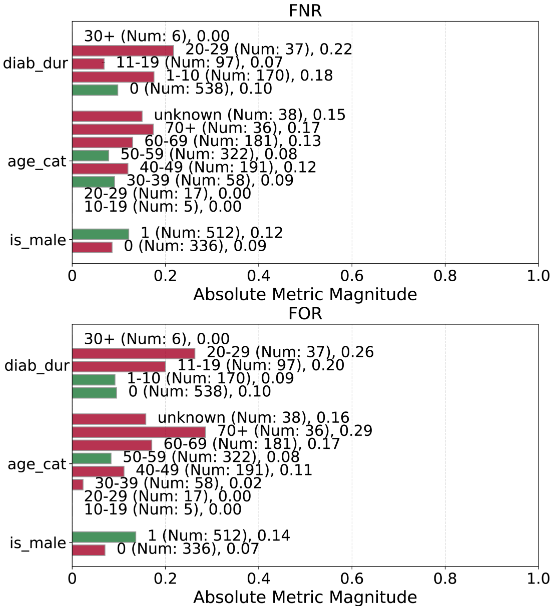


Figure 2: FNR and FOR Fairness of diabetic retinopathy predictions across the groups diabetes duration (diab_dur), age-groups (age_cat), and gender (is_male). 'Unfairness' and 'Fairness' within groups are marked in red and green, respectively.

were anonymously uploaded to the FG-AI4H server[2] under identifier J-049.

Use case 1: Diagnostic prediction of DR In the questionnaire, use case owners stated that the intended use of their model was to determine if Artificial Intelligence (AI) is capable of detecting DR at an early stage to assist in a primary care setting. From data reporting, it was unclear from which healthcare context (i.e. hospital, routine or diabetic care) the data was collected. Another unknown is the true diabetes status of individuals with a diabetes duration of zero years (representing the majority in the test set at 63.4%) and how many ophthalmologists labeled annotations. Non-gradable images stemming from operational errors were excluded from the test set. The model performance was reported to be comparable to a clinician, however details on the analysis are unknown. External validation was carried out but no details of the data source were given. More information elicited through the questionnaire is summarized in B. To the model owners, we pointed out a lack of metadata and cautioned that a binary classification task distinguishing normal vs. DR images does not fulfill the task of detecting DR at an early stage. We further recommended to record the healthcare context where patients presented and the types of cameras used at each site. It is likely that images contain artifacts generated by different cameras at different sites which can confound model development. Overall performance can be found in C.1. We stratified performance metrics calculated with *aequitas* across diabetes duration in years (0, 1-10, 11-19, 20-29, 30+), age (groups: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+), and gender (male, female). Figure 2 displays absolute FNR and FOR values, with disparities outside the fairness range marked as red and green otherwise. All metric values and disparities can be found in C.2. *Aequitas* indicated bias for all diabetes duration (diab_dur) groups (except the reference group of zero years of diabetes). Individuals with 30+ years of diabetes were underrepresented (n=6) and no false predictions were made in this group. Unfair FNR and FOR distributions were found for the age groups of 70, 60-69, 40-49 and unknown, with the highest metric values and disparity for 70+ years. This could be due to an under-representation of individuals who are 70 years or older and requires further analysis. FNR and FOR were higher for males than for females, however females represented only 34.2% of the data. FNR, FOR, NPV, precision and predicted prevalence metrics with disparities across groups can be found in table C.3. We recommend to apply strategies, such as reweighting from the AI Fairness 360 toolkit, (Bellamy et al., 2018) that can help to

[2] https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx

Table 1: Robustness metrics for the Diagnostic prediction of diabetic retinopathy using *mCE* Hendrycks and Dietterich (2019) for each corruption function applied, with a $E_{\text{clean}} = 0.2232$. The lower the more robust the model to the perturbation type tested.

Perturbation c	$\sum_{s=1}^{3} E^f_{s,c}$	$\sum_{s=1}^{3} E^f_{s,c} - E^f_{\text{clean}}$	CE^f_c
Elastic transform	0.7252	0.0555	0.0765
Pixelate	1.0700	0.4003	0.3741
Brightness	1.4835	0.8137	0.5485
Motion blur	1.5337	0.8639	0.5633
Gaussian Blur	1.7860	1.1162	0.6250
Defocus blur	1.8454	1.1757	0.6371
Saturate	1.8732	1.2034	0.6425
Speckle noise	2.1427	1.4729	0.6874
Spatter	2.2464	1.5767	0.7019
JPEG compression	2.4730	1.8033	0.7292
Shot noise	2.6143	1.9445	0.7438
Contrast	2.6618	1.9921	0.7484
Gaussian Noise	2.7305	2.0608	0.7547
Impulse noise	2.8349	2.1651	0.7637
Fog	2.8996	2.2299	0.7690
Frost	2.9643	2.2946	0.7741

mitigate bias before classification. Model interpretations were generated by the *Meaningful Perturbation* approach (Fong and Vedaldi, 2017). The *Retinopathy* inputs have been explained by maximizing the *No Retinopathy* class activation. Explanation masks for four images classified as *Retinopathy* is shown in Figure C.9. These masks present a high degree of noise, indicating that the model does not focus on relevant image elements such as arteries. Such degree of noise made it difficult to obtain group explanations in clusters highlighting different classification behaviours. Further investigation on this matter was recommended to the model owner. The results for the 15 tested perturbations with 3 different perturbation levels, using the test dataset is displayed in Table 1. The *mCE* for the tested perturbations indicates high robustness for brightness, pixelate and motion blur transformations, and less robustness to contrast, Gaussian and impulse noise transformations which use case owners were made aware of. More details on this analysis can be found in C.4.

Use case 2: Diagnostic prediction of AD In the model reporting questionnaire, model owners stated the model's primary intended use was a clinical diagnostic prediction model for neurology that outputs disease class probabilities, leaving questions about the exact target. The model owners trained a binary classifier using Naïve Bayes and Gradient Boosting algorithms to distinguish cognitive normal (CN) individuals from cases with AD. Models were trained on the publicly available AD research data sets 'adni' (Mueller et al., 2005), and 'edsd' (Brueggen et al., 2017). The owners also included data from three European hospitals into training, which could not be shared for assessment due to data protection. Cognitive status labels are based on possibly subjective clinical assessment (i.e., cognitive testing and MR-imaging). The model owners excluded 1,472 (65.3%) entries from the total data set (n=2,254 including only adni and edsd), because their cognitive status was 'Mild cognitive impairment' (MCI) or 'Other.' The exclusion of these entries was not reported in the questionnaire. Models were trained with k-fold cross validation without keeping a holdout set. Gradient Boosting using brain volumes, age, and gender as predictors was the best performer, which we used for further evaluation. In the questionnaire, authors stated that they tested convolutional neural network (CNN) and support vector machine models, which was not available to us for evaluation. We pointed out that a binary classifier distinguishing individuals between CN and AD does not fulfil the intended use of clinically relevant prediction models. The model owners are working on including MCI classes, however these reduce the model performance significantly. We further requested to report test results on protected data from the 3 European hospitals. More information elicited through the questionnaire is summarized in B. As no holdout set was provided, we retrained the binary GradientBoost algorithm (n=586, 75%) and performed the bias assess-

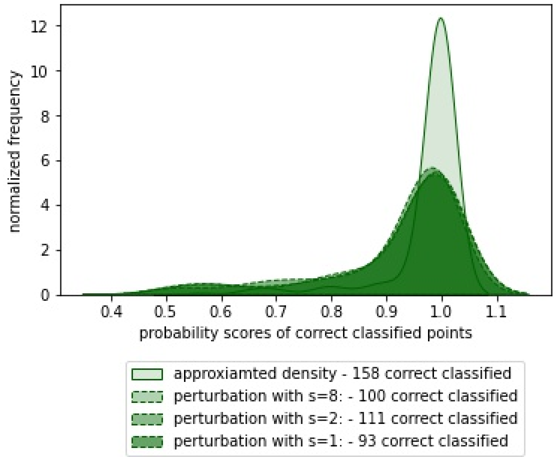


Figure 3: Approximated density functions of AD Gradient Boost scores under different perturbation levels with lognormal noise.

Table 2: FNR and FOR of predicting Alzheimer's disease across the groups dataset, age categories, and gender. Given are number of samples (#), absolute metric values with 95% confidence intervals (CI) and respective disparities δ. Disparities of 1.00 mark the reference group.

	Alzheimer's Disease				
	#	**FNR (CI)**	**FNR δ**	**FOR (CI)**	**FOR δ**
dataset: adni	129	0.15 (0.09-0.22)	1.00	0.15 (0.08-0.21)	1.00
dataset: edsd	67	0.31 (0.20-0.42)	2.05	0.23 (0.13-0.33)	1.60
age cat: 50-59	7	0.50 (0.13-0.87)	2.33	0.4 (0.04-0.76)	2.71
age cat: 60-69	39	0.15 (0.04-0.26)	0.70	0.15 (0.04-0.26)	1.02
age cat: 70-79	108	0.21 (0.14-0.29)	1.00	0.15 (0.08-0.21)	1.00
age cat: 80+	42	0.20 (0.08-0.32)	0.93	0.26 (0.13-0.40)	1.78
female	99	0.18 (0.10-0.26)	1.00	0.18 (0.11-0.26)	1.00
male	97	0.24 (0.16-0.33)	1.36	0.18 (0.10-0.26)	0.97

ment on the holdout test set (n=196, 25%). Overall performance can be found in C.1. We calculated performance metrics across data sets (either adni or edsd), age groups and gender. Absolute group metric values and disparities can be found in C.2. Unfairness was detected among all three groups. Table 2 lists FNR and FOR values with disparities for each group. FNR and FOR was higher for edsd samples compared to samples from adni. In the age group, FNR and FOR unfairness was indicated for 50-59 year old individuals. The age group of 50-59 years comprised only 7 individuals. In rare cases AD onset can start early in life (Mendez, 2017) and it should be evaluated if this type of AD can be detected by the algorithm. In the gender group, the FNR for males was higher than for females. FOR between gender groups were distributed equally. NPV, precision and predicted prevalence metrics across groups can be found in table C.3

We used TreeSHAP to generate explanations for the test set. Clustering selected 40 out of 196 points for human evaluation. For each of these points, an explanation based on Anchors was generated. All representative anchors contained either the brain feature *rightententorhinalarea* or *leftententorhinalarea*. Age or gender features were never anchoring conditions. We found that the SHAP value for age in the age groups 50-59y was pushing the prediction towards a positive prediction. This might indicate an undesired local behaviour of the decision boundary which could be improved by introducing monotonicity constraints with respect to age. More detailed results in C.3 While the model does not appear sensitive to perturbations of the gender feature (Figure C.20 and Figure C.21), the score densities shift more drastically when noise is applied to age and brain features (Figure 3).

Use case 3: Cytomorphologic classification to aid leukemia diagnostics The model's primary intended use was reported as a diagnostic decision aid for medical staff examining the leukocyte morphology in the diagnostics of Acute Myeloid Leukemia (AML). Data collection took place at the Laboratory of Leukemia Diagnostics at LMU Klinikum Munich. The model has not yet been validated on external data. There is no indication suggesting that sensitive demographic variables might influence model performance. The developers report that precision and sensitivity are correlated with the number of images in this class. Further investigations about model generalizability are planned. This may include cases with other haematological diagnoses, samples from

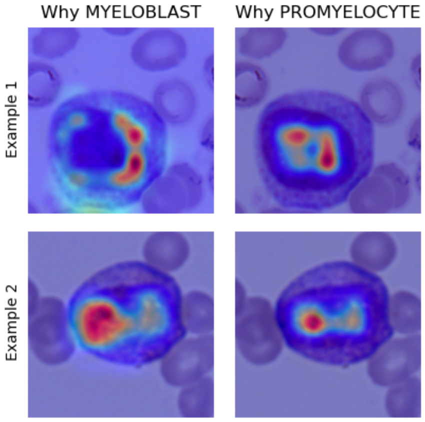


Figure 4: Two representative examples from use case 3. The columns contain Class Activation maps corresponding the predicted class (left) and correct class (right).

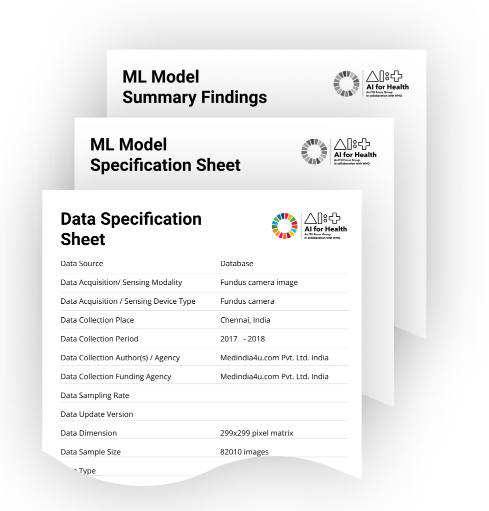
other centers, or differential sample processing. As published, the model can be used as a research-purpose decision aid only. More information from the questionnaire is summarized in A.3. Bias assessment with *aequitas* was not applicable to this use case, as no metadata was available. Model interpretations could only be obtained using the Grad-CAM method as the model architecture was built with the older *TensorFlow* 1.11 version, which is incompatible with other, more recent interpretability libraries. We grouped class activation map (CAM)s based on their softmax layer value and found 364 representative points out of the original 18,365 test images. Figure 4 shows two examples of misclassified images, and respective CAMs. Both examples indicate that the model learned to focus on the leukocyte's nucleus and cytoplasm, while ignoring erythrocytes and other background structures. Beyond this check, it remains difficult to explain the relative contribution of individual features to the classification decision in particular when highlighted CAM regions are distributed across extended regions of the input. To test the robustness of the cytomorphology classifier, 16 perturbations were applied to 73 input images. Predictions of the classifier turned out to be robust against most perturbations (cf. C.4), both for correct and incorrect classifications. Some perturbations increased output confidence of misclassifications, which may reflect the general calibration behaviour of the network types used in the study (Guo et al., 2017).

5. Discussion

Standardized assessment and reporting of ML4H quality are crucial to pave the way for translation into medical practice. First milestones in accommodating both technical and health considerations in a transparent and unified evaluation procedure have been reached by the FG-AI4H and tangential standardization efforts. It is due time to move these efforts from paper to practice to further improve them. We applied a selection of ML4H quality assessment methods from the framework suggested by the ITU/WHO working group FG-AI4H on three use cases. In the following, we discuss how each step contributed to the quality assessment, how results were perceived by model owners and make suggestions for further improvement and future research.

Questionnaire: The model developers reported that filling out the questionnaire required between 0.5-2 hours, but helped to review the development process critically considering weaknesses and further improvements. Transparent reporting through the questionnaire can help to perform further qualitative bias assessment with tools such as PROBAST (Wolff et al., 2019).

Bias and fairness: ML4H algorithms are prone to bias for multiple reasons. Due to an under-representation of certain patient groups, early disease stages or rare disease types in healthcare data, an algorithm could not provide the same benefit for everyone (Gianfrancesco et al., 2018). These potential sources of bias are included in ML4H quality assessment frameworks, but challenges remain as data sets often do not contain vari-



ML Model
Summary Findings

ML Model
Specification Sheet

Data Specification
Sheet

Data Source	Database
Data Acquisition/ Sensing Modality	Fundus camera image
Data Acquisition / Sensing Device Type	Fundus camera
Data Collection Place	Chennai, India
Data Collection Period	2017 - 2018
Data Collection Author(s) / Agency	Medindia4u.com Pvt. Ltd. India
Data Collection Funding Agency	Medindia4u.com Pvt. Ltd. India
Data Sampling Rate	
Data Update Version	
Data Dimension	299x299 pixel matrix
Data Sample Size	82010 images
Type	

Figure 5: A snapshot of the FG-AI4H audit report cards for the retinopathy use case. Full version in Appendix B

ables that allow group stratification and testing these sources. These variables are often either not recorded or cannot be shared due to data protection laws. More research and policies are needed to fill this gap for bias assessment in ML4H.

Interpretability: Interpretable model decisions can help to analyse mistakes and undesired model behaviour that can occur from spurious correlations, for example measurement artifacts, in the training data (Lapuschkin et al., 2019; Ribeiro et al., 2016; Pfau et al., 2019). One challenge for this assessment was that existing interpretability analysis tools are often implemented with external libraries that became incompatible with model requirements. Large test data can topple computational capacities of explanation methods. Interviews with model owners highlighted that although image interpretability maps can be useful to detect undesired behaviours, they still explain model mistakes to a limited extent. Especially for noisy heatmaps it is hard to identify systematic model mistakes. Interactive tools with UX/UI design would be helpful to navigate through different model behaviours.

Robustness: We assessed the robustness of the three use cases to identify when models fail, by perturbing the input data. One challenge with this approach is to generate perturbations that are meaningful in medical practice. Perturbations like Gaussian noise, produce blurry images that are realistic through operational errors. Other perturbations such as frost imitations are rather unsuitable for medical applications. It is not yet possible to generate specific artifacts that can be introduced during sample handling and preparation. While artifacts, such as morphologic changes in ageing blood samples, are well-known (Vives Corrons et al., 2004), others may present a challenge even to experienced examiners (Dalal and Brigden, 2002). Evaluating an algorithm on specific artifacts hence remains challenging, for which standard methods have yet to be established.

A limiting factor of the FG-AI4H audit framework is that it only features a selection of methods. We conducted analyses to cover the quality aspects bias, interpretability, and robustness, but did not yet perform analyses on generalization. An analysis on data quality is also recommended in the framework but could not be performed on these three use cases, as it requires raw imaging data, which were not available. Other than the provided clinical use case expertise, the performed assessment does not yet address clinical validation in itself, which requires clinical studies. However, we note that limiting the scope of ML4H auditing frameworks to ML specific problems and exploiting established processes for clinical assessment may be a viable way forward to close this gap, too.

References

Rachel K.E. Bellamy, Dey Kuntal, Michael Hind, Samuel C Hoffmann, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Natesan R Karthikeyan, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Mohinder Singh, Kush R Varshney, and Zhang Yunfeng. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *arXiv*, 2018. ISSN 21518556. doi: 10.1147/JRD.2019.2942287.

Katharina Brueggen, Michel J. Grothe, Martin Dyrba, Andreas Fellgiebel, Florian Fischer, Massimo Filippi, Federica Agosta, Peter Nestor, Eva Meisenzahl, Janusch Blautzik, Lutz Frölich, Lucrezia Hausner, Arun L.W. Bokde, Giovanni Frisoni, Michela Pievani, Stefan Klöppel, David Prvulovic, Frederik Barkhof, Petra J.W. Pouwels, Johannes Schröder, Harald Hampel, Karlheinz Hauenstein, and Stefan Teipel. The European DTI Study on Dementia — A multicenter DTI and MRI study on Alzheimer's disease and Mild Cognitive Impairment. *NeuroImage*, 144:305–308, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.03.067. URL http://dx.doi.org/10.1016/j.neuroimage.2016.03.067.

Bakul I. Dalal and Malcolm L. Brigden. Artifacts that may be present on a blood film. *Clinics in laboratory medicine*, 22(1): 81—100, vi, March 2002. ISSN 0272-2712. doi: 10.1016/s0272-2712(03)00068-4.

Döhner, Hartmut, Estey, Elihu, Grimwade, David, Amadori, Sergio, Appelbaum, Frederick R, Büchner, Thomas, Dombret, Hervé, Ebert, Benjamin L., Fenaux, Pierre, Larson, Richard A, Levine, Ross L, Lo-Coco, Francesco, Naoe, Tomoki, Niederwieser, Dietger, Ossenkoppele, Gert J, Sanz, Miguel, Sierra, Jeorge, Tallman, Martin S, Tien, Hwei-Fang, Wei, Andrew H, Löwenberg, Bob, Bloomfield, and Clara D. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, 129(4):424–447, 2017. URL https://doi.org/10.1182/blood-2016-08-733196.

Bruno Dubois, Howard H. Feldman, Claudia Jacova, Jeffrey L. Cummings, Steven T. Dekosky, Pascale Barberger-Gateau, André Delacourte, Giovanni Frisoni, Nick C Fox, Douglas Galasko, Serge Gauthier, Harald Hampel, Gregory A. Jicha, Kenichi Meguro, John O'Brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Salloway, Marie Sarazin, Leonardo C de Souza, Yaakov Stern, Pieter J Visser, and Philip Scheltens. Revising the definition of Alzheimer's disease: a new lexicon. *The Lancet Neurology*, 9:118–1127, 2009a.

Bruno Dubois, Harald Hampel, Howard H. Feldman, Philip Scheltens, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Karl Broich, Enrica Cavedo, Sebastian Crutch, Charles Duyckaerts, Giovanni B. Frisoni, Serge Gauthier, Alida A. Gouw, Marie-odile Habert, David M. Holtzman, Miia Kivipelto, Simone Lista, Gil D. Rabinovici, Christopher Rowe, Stephen Salloway, Lon S. Schneider, Reisa Sperling, Maria C. Carrillo, Jeffrey Cummings, and Cliff R Jack Jr. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *The Lancet Neurology*, 12(3):292–323, 2009b. doi: 10.1016/j.jalz.2016.02.002.Preclinical.

Simão Eduardo, Alfredo Nazábal, Christopher K. I. Williams, and Charles Sutton. Robust Variational Autoencoders for Outlier Detection and Repair of Mixed-Type Data. *arXiv*, 2019. URL http://arxiv.org/abs/1907.06671.

Martin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD Proceedings*, 1996. doi: 10.1016/B978-044452701-1.00067-3.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019. ISSN 1546170X. doi: 10.1038/s41591-018-0316-z. URL http://dx.doi.org/10.1038/s41591-018-0316-z.

FG-AI4H. Data and artificial intelligence assessment methods (daisam) reference. *Reference document DEL 7.3 on FG-AI4H server*, 2020a. URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx.

FG-AI4H. Model questionnaire. *Reference document J-038 on FG-AI4H server*, 2020b. URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx.

FG-AI4H. Daisam audit reporting template. *Reference document J-048 on FG-AI4H server*, 2020c. URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx.

Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:3449–3457, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.371.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé, and Kate Crawford. Datasheets for Datasets. pages 1–28, 2018. URL http://arxiv.org/abs/1803.09010.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in ML Algorithms Using EHR Data. *JAMA Intern. Med.*, 178(11):1544–1547, 2018. doi: 10.1001/jamainternmed.2018.3763.Potential.

Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22): 2402–2410, 12 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.17216. URL https://doi.org/10.1001/jama.2016.17216.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org, 2017.

Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

IMDRF. Artificial intelligence in healthcare opportunities and challenges, 2019. URL http://imdrf.org/docs/imdrf/final/meetings/imdrf-meet-190916-russia/yekaterinburg-14.pdf.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019. ISSN 20411723. doi: 10.1038/s41467-019-08987-4. URL http://dx.doi.org/10.1038/s41467-019-08987-4.

Jerry Zheng Li. *Principled approaches to robust machine learning and beyond*. PhD thesis, Massachusetts Institute of Technology, 2018.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie Calvert, Alastair K Denniston, The Spirit-ai, and Consort-ai Working Group. CONSORT-AI extension. *Nature Medicine*, 26(September):1364–1374, 2020. doi: 10.1038/s41591-020-1034-x.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

Christian Matek, Simone Schwarz, Carsten Marr, and Karsten Spiekermann. A single-cell morphological dataset of leukocytes from aml patients and non-malignant controls [data set]. the cancer imaging archive. https://doi.org/10.7937/tcia.2019.36f5o9ld.

Christian Matek, Simone Schwarz, Karsten Spiekermann, and Carsten Marr. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*, 1(11):538–544, 2019.

Mario F Mendez. Early-Onset Alzheimer's Disease. *Neurol. Clin.*, 35(2):263–281, 2017. doi: 10.1016/j.physbeh.2017.03.040.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, (Figure 2):220–229, 2019. doi: 10.1145/3287560.3287596.

Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Alzheimer's Disease Neuroimaging Initiative, et al. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. *Neuroimage*, 104:398–412, 2015.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am.*, 15(4):869–77, 2005. doi: doi:10.1016/j.nic.2005.09.008.

Perumalsamy Namperumalsamy, Praveen K. Nirmalan, and Kim Ramasamy. Developing a screening program to detect sight-threatening diabetic retinopathy in South India. *Diabetes Care*, 26(6):1831–1835, 2003. ISSN 01495992. doi: 10.2337/diacare.26.6.1831.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.

Font P., Loscertales J., and Benavente C. et al. Inter-observer variance with the diagnosis of myelodysplastic syndromes (MDS) following the 2008 WHO classification. *Ann Hematol*, 93:19–24, 2013. URL https://doi.org/10.1007/s00277-012-1565-4.

Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

Jacob Pfau, Albert T. Young, Maria L. Wei, and Michael J. Keiser. Global Saliency: Aggregating Saliency Maps to Assess Dataset Artefact Bias. *Machine Learning for Health (ML4H) at NeurIPS 2019*, pages 1–9, 2019. URL http://arxiv.org/abs/1910.07604.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:1135–1144, 2016. doi: 10.1145/2939672.2939778.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 1527–1535, 2018.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit. (2018), 2018. URL http://arxiv.org/abs/1811.05577.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128:336–359, 2019. doi: 10.1007/s11263-019-01228-7.

Mark P. Sendak, Michael Gao, Nathan Brajer, and Suresh Balu. Presenting machine learning model information to clinical end users with model facts labels. *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, Hutan Ashrafian, Ravi Aggarwal, Jeffrey De Fauw, Alastair K. Denniston, Felix Greaves, Alan Karthikesalingam, Dominic King, Xiaoxuan Liu, Sheraz R. Markar, Matthew D.F. McInnes, Trishan Panch, Jonathan Pearson-Stuttard, Daniel S.W. Ting, Robert M. Golub, David Moher, Patrick M. Bossuyt, and Ara Darzi. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nature Medicine*, 26(6):807–

808, 2020. ISSN 1546170X. doi: 10.1038/s41591-020-0941-1. URL http://dx.doi.org/10.1038/s41591-020-0941-1.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Eric J. Topol. human and artificial intelligence. *Nature Medicine*, 25(January), 2019. ISSN 1546-170X. doi: 10.1038/s41591-018-0300-7. URL http://dx.doi.org/10.1038/s41591-018-0300-7.

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.

Juan Lluis Vives Corrons, Stephanie Albarède, George Flandrin, Silke Heller, Katalin Horvath, Berend Houwen, Gunnar Nordin, Erika Sarkani, Milan Skitek, Marjan Van Blerk, and Jean Claude Libeer. Guidelines for blood smear preparation and staining procedure for setting up an external quality assessment scheme for blood smear interpretation. Part 1: Control material. *Clinical Chemistry and Laboratory Medicine*, 42(8):922–926, 2004. ISSN 14346621. doi: 10.1515/CCLM.2004.149.

Markus Wenzel and Thomas Wiegand. Toward global validation standards for health ai. *IEEE Communications Standards Magazine*, 4(3):64–69, 2020.

Thomas Wiegand, Ramesh Krishnamurthy, Monique Kuglitsch, N. Lee, Sameer Pujari, Marcel Salathé, Markus Wenzel, and Shan Xu. WHO and ITU establish benchmarking process for artificial intelligence in health. *The Lancet*, 394(10192):9–11, 2019. ISSN 1474547X. doi: 10.1016/S0140-6736(19)30762-7.

Robert F. Wolff, Karel G.M. Moons, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170 (1):51–58, 2019. ISSN 15393704. doi: 10.7326/M18-1376.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

Joanne W.Y. Yau, Sophie L. Rogers, Rho Kawasaki, Ecosse L. Lamoureux, Jonathan W. Kowalski, Toke Bek, Shih Jen Chen, Jacqueline M. Dekker, Astrid Fletcher, Jakob Grauslund, Steven Haffner, Richard F. Hamman, M. Kamran Ikram, Takamasa Kayama, Barbara E.K. Klein, Ronald Klein, Sannapaneni Krishnaiah, Korapat Mayurasakorn, Joseph P. O'Hare, Trevor J. Orchard, Massimo Porta, Mohan Rema, Monique S. Roy, Tarun Sharma, Jonathan Shaw, Hugh Taylor, James M. Tielsch, Rohit Varma, Jie Jin Wang, Ningli Wang, Sheila West, Liang Zu, Miho Yasuda, Xinzhi Zhang, Paul Mitchell, and Tien Y. Wong. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, 35(3):556–564, 2012. ISSN 01495992. doi: 10.2337/dc11-1909.

Appendix A. Detailed description of use cases

A.1. Diagnostic prediction of diabetic retinopathy

Use case description The initial goal was to develop an AI model that could be used for screening those at risk for diabetic retinopathy in primary care settings such as doctors offices and diabetes clinics in India for referable retinopathy.

Data The training data comprised 82,010 digital images of the retina from fundus cameras obtained from patients being screened for diabetic retinopathy from ophthalmologists in India. The images were in JPEG or PNG format from 2 to 5 megapixels in resolution. For training they were resized to 224×224 pixels and normalized using ImageNet standards. They were labeled into six categories: Nongradable, Normal, Mild, Moderate, Severe and Proliferative DR. For binary classification Normal & Mild were grouped as 'no retinopathy' and the rest as having retinopathy. The data is not publically available. We evaluated the algorithm on a provided holdout test set with 848 samples of which 424 (50.0%) were cases. The test set contained information about duration of diabetes in years, age, and gender. The diabetes duration ranged from 0 to 40 years with a mean of 3.6 years. The majority of individuals (n=538, 63.4%) was recorded with a diabetes duration of 0 years. Increasing diabetes duration were represented at lower numbers (1-10y: 170, 20.0%, 11-19y: 97, 11.4%, 20-29y: 37, 4.3%, 30+y: 6, 0.7%). The age distribution ranged from 13 to 85 years, with a median of 53. There were 38 (4.5%) entries with missing age. Men represented the majority (n=512, 60.8%).

Architectures Initial models trained Resnet-101 Convolutional Neural Network pretrained on ImageNet. Later models used EfficientNet-B5 architecture with pre-trained Imagenet weights. The model which is assessed here is a Resnet-101 for binary classification of diabetic retinopathy. Optimizers used were Adam, learning rate annealing, and FastAI's One cycle implementation of Cyclical learning rates. Model and data are proprietary.

A.2. Diagnostic prediction of Alzheimer's Disease

Use case description With increased life expectancy, the number of individuals who will potentially become demented is growing proportionally. Current estimates count world-wide over 48 million people suffering from dementia bringing the social cost of care to 1% of world's gross domestic product – GDP. These numbers led the World Health Organisation to classify neurocognitive disorders as a global public health priority. Our goal is to address previous limitations by using "real-world" imaging data obtained in the clinical routine, predictive ML algorithms, including benchmarking and cross-validation of the learned models. The intended output is to assign a level of probability to each of several possible diagnosis (Cognitively normal, MCI or AD) that is readily usable and interpretable by clinicians.

Data We used T1w brain imaging data from 2254 participants of two different multi-centric cohorts, the ADNI (Alzheimer‘s Disease Neuroimaging Initiative) and European DTI Study on Dementia (EDSD). From each 3D brain scan of each subject we calculated regional brain volumes. This procedure consisted of 3 steps. 1) We segmented each brain scan into 3 tissue classes: white matter, gray matter and CSF . 2) We spatially warped the 3D probabilistic Neuromorphometrics atlas of each individual brain scan to get an individualised brain atlas. 3) We then calculated the volumes of each of the 134 brain regions of the atlas. The gray matter volume is the sum of the gray matter values of the voxels of each atlas region in the segmented 3D image. We applied a Naïve Bayes and Gradient Boosting algorithm to the final data that included the regional brain volumes, age and gender to distinguish cognitive normal (CN) individuals from cases with AD.

A.3. Cytomorphologic classification to aid leukemia diagnostics

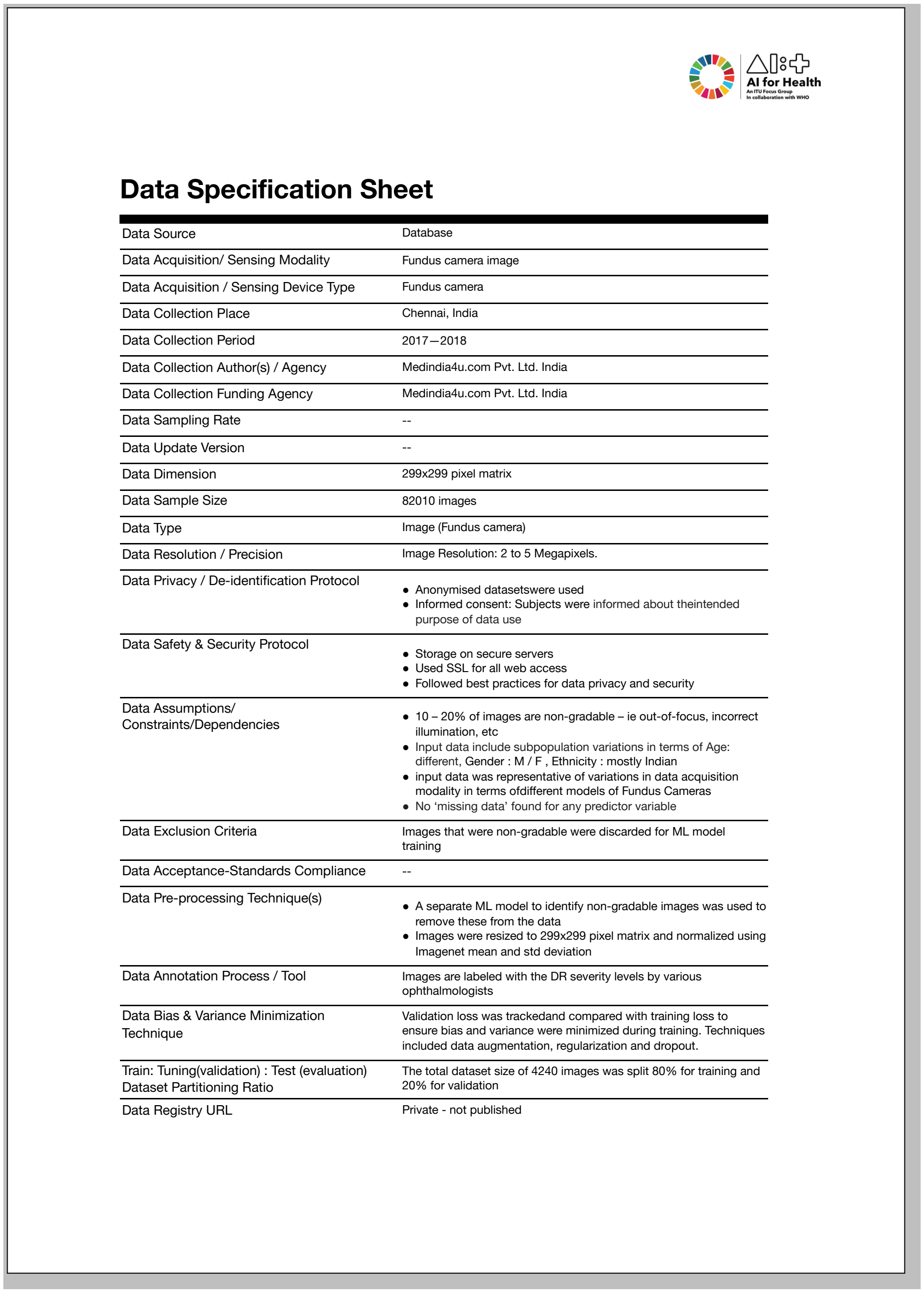
Use case description Examination of cells from the peripheral blood and bone marrow under a light microscope is a frequently used and long-established technique in the diagnosis of hematological disorders. For the diagnosis of leukemia in particular, morphological examination of leukocytes retains a key role in the diagnostic routine for its comparative technical simplicity and widespread availability. During this process, a trained examiner classifies cells on a slide into a scheme of morphological categories. According to current guidelines, at least 200 cells should be examined and classified in the case of bone marrow (Döhner et al., 2017). Still today, this classification and counting of leukocytes is widely performed manually, which is a laborious and time-consuming process and limits the number of examinations available overall. Furthermore, manual classification is potentially fraught with considerable intra- and inter-observer variability (P. et al., 2013). Finally, the intrinsically subjective nature of morphological examination makes it hard to combine this method with more recent diagnostic modalities that yield quantitative results. ML methods such as deep learning have the potential to bridge this gap and act as a quantitatively informed decision aid for the examiner.

Data The data comprises single-cell images from peripheral blood smears of 100 patients diagnosed with Acute Myeloid Leukemia (AML) at LMU Klinikum Munich between 2014 and 2017, and 100 non-malignant controls. For digitization of the monolayer, an M8 digital microscope (Precipoint GmbH, Freising/Germany) at 100x objective magnification and oil immersion was used. A cytologist then differentiated individual leukocytes on the scanned region in analogy with diagnostic routine. Around the annotated leukocytes, patches of the size 400 × 400 pixels were extracted without further modification to the image data output by the scanner. This yielded a set of 18,365 single-cell images classified into a scheme comprising 15 morphological categories derived from routine practice. To estimate intra- and inter-rater variability, a second, independent cytologist re-annotated the single cell images at two different times. Details of the data generation process are described in Matek et al. (2019). The full Munich AML Morphology Data Set is freely available on The Cancer Imaging Archive (Matek et al.).

Architectures Models were trained on the tasks of (i) classifying a given single-cell image into a 15-category morphological scheme derived from routine diagnostics and (ii) answering clinically relevant binary questions on the blast character and atypicality of a given single-cell image (Matek et al., 2019). Two distinct models were used in the training, namely the ResNeXt model developed by Xie et al. (2017) as well as a sequential model. Hyperparameters including the cardinality parameter were unchanged from the original setup of Xie et al. (2017) apart from adjusting the input and output channels. Both models were trained using the Adam optimizer. Stratified 5-fold cross-validation was used to estimate the variability of network predictions. Details on the architecture of the sequential model are provided by (Matek et al., 2019).

Appendix B. Audit report in FG-AI4H template

In the following we include audit results in the FG-AI4H reporting format. In order to avoid an indecent inflation of the appendices, the report cards for the other two use cases, along with all other outputs of the audit process, where submitted anonymously to the FG-AI4H document server of ITU under identifier FGAI4H-J-049: https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx.

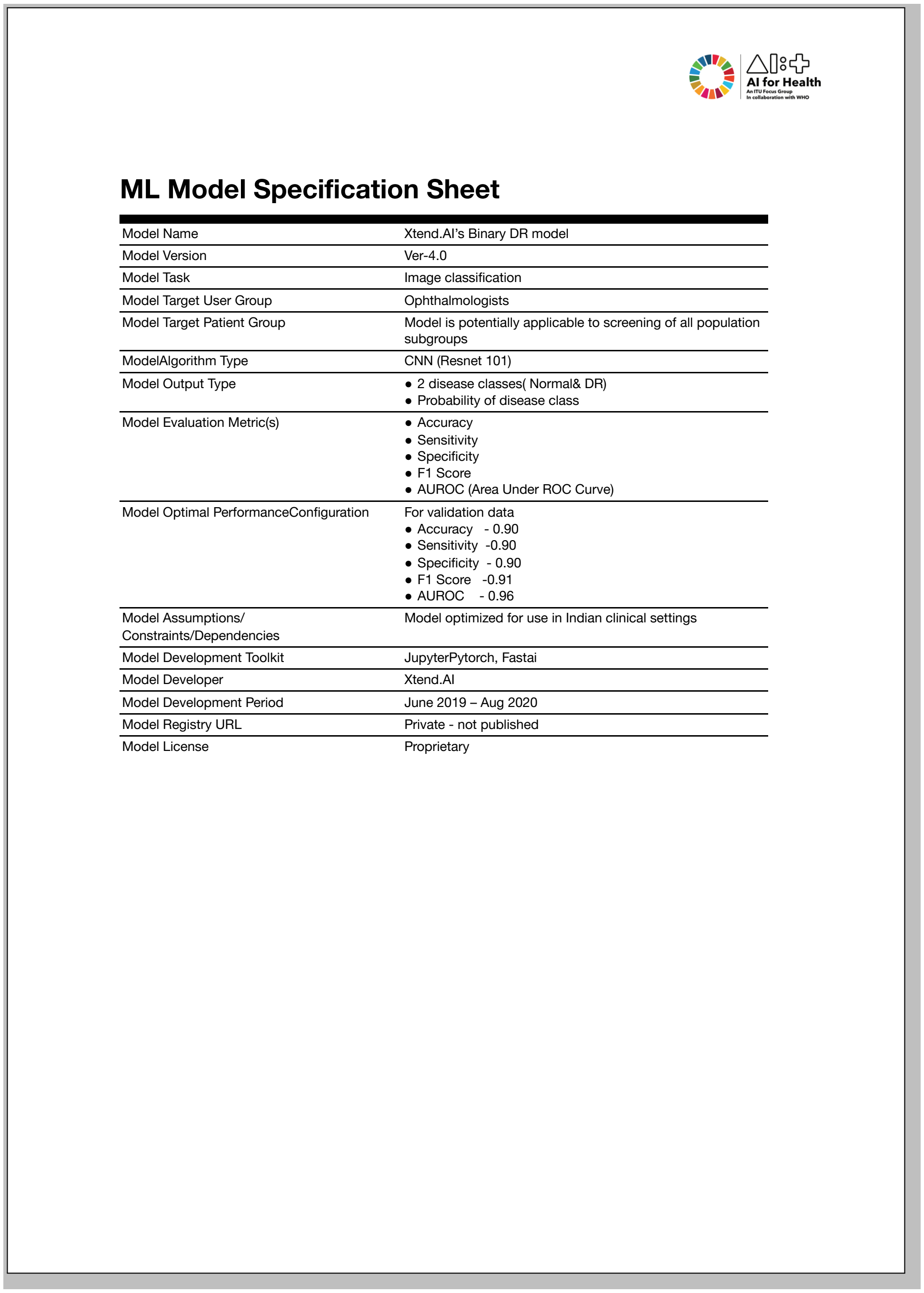


AI for Health
An ITU Focus Group
In collaboration with WHO

Data Specification Sheet

Data Source	Database
Data Acquisition/ Sensing Modality	Fundus camera image
Data Acquisition / Sensing Device Type	Fundus camera
Data Collection Place	Chennai, India
Data Collection Period	2017—2018
Data Collection Author(s) / Agency	Medindia4u.com Pvt. Ltd. India
Data Collection Funding Agency	Medindia4u.com Pvt. Ltd. India
Data Sampling Rate	--
Data Update Version	--
Data Dimension	299x299 pixel matrix
Data Sample Size	82010 images
Data Type	Image (Fundus camera)
Data Resolution / Precision	Image Resolution: 2 to 5 Megapixels.
Data Privacy / De-identification Protocol	● Anonymised datasetswere used ● Informed consent: Subjects were informed about theintended purpose of data use
Data Safety & Security Protocol	● Storage on secure servers ● Used SSL for all web access ● Followed best practices for data privacy and security
Data Assumptions/ Constraints/Dependencies	● 10 – 20% of images are non-gradable – ie out-of-focus, incorrect illumination, etc ● Input data include subpopulation variations in terms of Age: different, Gender : M / F , Ethnicity : mostly Indian ● input data was representative of variations in data acquisition modality in terms ofdifferent models of Fundus Cameras ● No 'missing data' found for any predictor variable
Data Exclusion Criteria	Images that were non-gradable were discarded for ML model training
Data Acceptance-Standards Compliance	--
Data Pre-processing Technique(s)	● A separate ML model to identify non-gradable images was used to remove these from the data ● Images were resized to 299x299 pixel matrix and normalized using Imagenet mean and std deviation
Data Annotation Process / Tool	Images are labeled with the DR severity levels by various ophthalmologists
Data Bias & Variance Minimization Technique	Validation loss was trackedand compared with training loss to ensure bias and variance were minimized during training. Techniques included data augmentation, regularization and dropout.
Train: Tuning(validation) : Test (evaluation) Dataset Partitioning Ratio	The total dataset size of 4240 images was split 80% for training and 20% for validation
Data Registry URL	Private - not published

Figure 6: The FG-AI4H data specification sheet for the retinopathy use case.

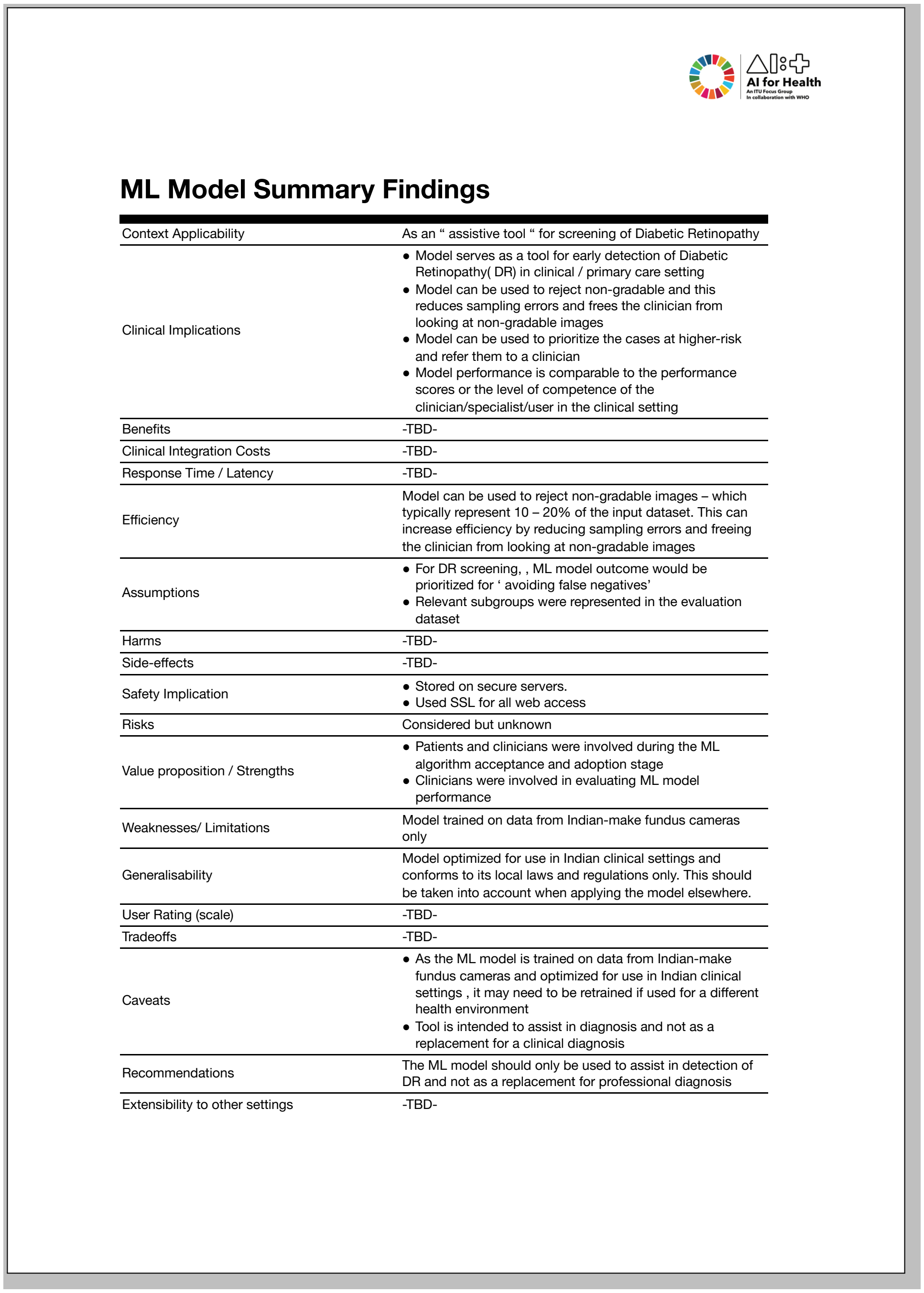


AI for Health
An ITU Focus Group
In collaboration with WHO

ML Model Specification Sheet

Model Name	Xtend.AI's Binary DR model
Model Version	Ver-4.0
Model Task	Image classification
Model Target User Group	Ophthalmologists
Model Target Patient Group	Model is potentially applicable to screening of all population subgroups
ModelAlgorithm Type	CNN (Resnet 101)
Model Output Type	• 2 disease classes(Normal& DR) • Probability of disease class
Model Evaluation Metric(s)	• Accuracy • Sensitivity • Specificity • F1 Score • AUROC (Area Under ROC Curve)
Model Optimal PerformanceConfiguration	For validation data • Accuracy - 0.90 • Sensitivity -0.90 • Specificity - 0.90 • F1 Score -0.91 • AUROC - 0.96
Model Assumptions/ Constraints/Dependencies	Model optimized for use in Indian clinical settings
Model Development Toolkit	JupyterPytorch, Fastai
Model Developer	Xtend.AI
Model Development Period	June 2019 – Aug 2020
Model Registry URL	Private - not published
Model License	Proprietary

Figure 7: The FG-AI4H model specification sheet for the retinopathy use case.



AI for Health
An ITU Focus Group
In collaboration with WHO

ML Model Summary Findings

Context Applicability	As an " assistive tool " for screening of Diabetic Retinopathy
Clinical Implications	● Model serves as a tool for early detection of Diabetic Retinopathy(DR) in clinical / primary care setting ● Model can be used to reject non-gradable and this reduces sampling errors and frees the clinician from looking at non-gradable images ● Model can be used to prioritize the cases at higher-risk and refer them to a clinician ● Model performance is comparable to the performance scores or the level of competence of the clinician/specialist/user in the clinical setting
Benefits	-TBD-
Clinical Integration Costs	-TBD-
Response Time / Latency	-TBD-
Efficiency	Model can be used to reject non-gradable images – which typically represent 10 – 20% of the input dataset. This can increase efficiency by reducing sampling errors and freeing the clinician from looking at non-gradable images
Assumptions	● For DR screening, , ML model outcome would be prioritized for ' avoiding false negatives' ● Relevant subgroups were represented in the evaluation dataset
Harms	-TBD-
Side-effects	-TBD-
Safety Implication	● Stored on secure servers. ● Used SSL for all web access
Risks	Considered but unknown
Value proposition / Strengths	● Patients and clinicians were involved during the ML algorithm acceptance and adoption stage ● Clinicians were involved in evaluating ML model performance
Weaknesses/ Limitations	Model trained on data from Indian-make fundus cameras only
Generalisability	Model optimized for use in Indian clinical settings and conforms to its local laws and regulations only. This should be taken into account when applying the model elsewhere.
User Rating (scale)	-TBD-
Tradeoffs	-TBD-
Caveats	● As the ML model is trained on data from Indian-make fundus cameras and optimized for use in Indian clinical settings , it may need to be retrained if used for a different health environment ● Tool is intended to assist in diagnosis and not as a replacement for a clinical diagnosis
Recommendations	The ML model should only be used to assist in detection of DR and not as a replacement for professional diagnosis
Extensibility to other settings	-TBD-

Figure 8: The FG-AI4H model summary findings sheet for the retinopathy use case.

Appendix C. Full quantitative results

C.1. Overall model performances

Diagnostic prediction of Diabetic retinopathy
The overall model performance on the test set (n=848) was 89.5% (CI: 87.4-91.6) accuracy, 89.2% (CI: 87.0-91.2) sensitivity, 89.9% (CI: 87.8-91.9) specificity, 89.8% (CI: 87.7-91.8) PPV, and 89.2% (CI: 87.1-91.3) NPV.

Diagnostic prediction of Alzheimer's Disease
Overall, the performance of the Gradient Boosting algorithm applied on the test set (n=196) to predict AD or CN had 80.6% (CI: 75.1-86.1) accuracy, 79.1% (CI: 73.4-84.8) sensitivity, 81.9% (CI: 76.5-87.2) specificity, 79.1% (CI: 73.4-84.8) PPV, and 81.9% (CI: 76.5-87.3) NPV.

Cytomorphologic classification to aid leukemia diagnostics
The model developers calculated classwise precision and sensitivity performances, instead of overall accuracy, as classes were represented at different frequencies in the data set. The model achieved sensitivity and precision above 0.9 for classes with more than 400 samples. Details can be found in Matek et al. (2019).

C.2. Bias and Fairness assessment with *aequitas*

We calculated the metric values FNR, FOR, NPV, precision, PPREV across groups. Absolute metric values and disparities for the prediction models for Alzheimer's disease and diabetic retinopathy are listed in table C.3.

C.3. Interpretability assessment

C.3.1. Diagnostic prediction of diabetic retinopathy

We generated interpretability maps by applying the *Meaningful Perturbation* approach which can be used to explain an input by generating a perturbation mask that minimizes the model's outputted probability in a constrained manner. This means generating a perturbed version of the original input in the following form

$$[\Phi(x_0; m)](u) = m(u)x_0(u) + (1 - m(u))\mu_0. \tag{1}$$

Where m is a perturbation mask used to explain the prediction, x_0 is the original image and μ_0 is a perturbation background (we used a black one). The mask m can be found by minimizing the following loss function

$$\begin{aligned}\min_{m\in[0,1]^\Lambda} \lambda_1 ||\mathbf{1} - m||_1 + \lambda_2 \sum_{u\in\Lambda} ||\nabla m(u)||_\beta^\beta \\ + \mathbb{E}_\tau[f_c(\Phi(x_0(\cdot - \tau), m))].\end{aligned} \tag{2}$$

The first two terms of the loss function are used to introduce constrains on the perturbation mask, in particular the second avoids finding adversarial examples. Since the model was written in *PyTorch* using a *Fast.ai* wrapper we used its *autograd* package to obtain the output derivative information with respect to a 15×15 pixels up-scaled perturbation mask. Once the derivative information is obtained the mask can be found by using a gradient descent approach to minimize the loss function defined above. Examples of perturbation

Table C.3: Metric values of predicting Alzheimer's disease and Diabetric retinopathy, statified across groups. Groups were defined as dataset, age categories (in years), and gender. Given are number of samples (#), absolute metric values of false negative rate (FNR), false omission rate (FOR), negative predicted value (NPV), precision and predicted prevalence (PPREV) with 95% confidence intervals (CI) and respective disparities to the reference as δ. Disparities of 1.00 mark the reference group.

						Diabetic Retinopathy					
	#	**FNR (CI)**	**FNR δ**	**FOR (CI)**	**FOR δ**	**NPV (CI)**	**NPV δ**	**precision (CI)**	**prec. δ**	**PPREV (CI)**	**PPREV δ**
diab dur: 0	538	0.10 (0.07-0.12)	1.00	0.10 (0.07-0.12)	1.00	0.91 (0.88-0.93)	1.00	0.91 (0.88-0.93)	1.00	0.49 (0.45-0.54)	1.00
diab dur: 1-10	170	0.18 (0.12-0.23)	1.79	0.09 (0.05-0.14)	0.96	0.91 (0.87-0.95)	1.00	0.77 (0.71-0.83)	0.85	0.36 (0.29-0.43)	0.73
diab dur: 11-19	97	0.07 (0.02-0.12)	0.70	0.20 (0.12-0.28)	2.10	0.80 (0.72-0.88)	0.88	0.94 (0.90-0.99)	1.04	0.74 (0.66-0.83)	1.51
diab dur: 20-29	37	0.22 (0.08-0.35)	2.22	0.26 (0.12-0.41)	2.76	0.74 (0.6-0.88)	0.81	1.00 (1.00-1.00)	1.10	0.49 (0.33-0.65)	0.99
diab dur: 30+	6	0.00 (0.00-0.00)	0.00	0.00 (0.00-0.00)	0.00	1.00 (1.00-1.00)	1.11	1.00 (1.00-1.00)	1.10	0.83 (0.54-1.0)	1.69
age cat: 10-19	5	0.00 (0.00-0.00)	NA	0.00 (0.00-0.00)	0.00	1.00 (1.00-1.00)	1.09	0.00 (0.00-0.00)	0.00	0.20 (0.00-0.55)	0.39
age cat: 20-29	17	0.00 (0.00-0.00)	0.00	0.00 (0.00-0.00)	0.00	1.00 (1.00-1.00)	1.09	0.57 (0.34-0.81)	0.62	0.41 (0.18-0.65)	0.80
age cat: 30-39	58	0.09 (0.02-0.17)	1.16	0.02 (0.00-0.06)	0.28	0.98 (0.94-1.00)	1.07	0.67 (0.55-0.79)	0.72	0.26 (0.15-0.37)	0.50
age cat: 40-49	191	0.12 (0.07-0.17)	1.53	0.11 (0.07-0.16)	1.33	0.89 (0.84-0.93)	0.97	0.88 (0.83-0.93)	0.96	0.48 (0.41-0.55)	0.93
age cat: 50-59	322	0.08 (0.05-0.11)	1.00	0.08 (0.05-0.11)	1.00	0.92 (0.89-0.95)	1.00	0.92 (0.89-0.95)	1.00	0.52 (0.46-0.57)	1.00
age cat: 60-69	181	0.13 (0.08-0.18)	1.66	0.17 (0.12-0.23)	2.05	0.83 (0.77-0.88)	0.90	0.95 (0.92-0.98)	1.03	0.55 (0.47-0.62)	1.06
age cat: 70+	36	0.17 (0.05-0.30)	2.22	0.29 (0.14-0.43)	3.43	0.71 (0.57-0.86)	0.78	0.86 (0.75-0.98)	0.94	0.61 (0.45-0.77)	1.19
age cat: unknown	38	0.15 (0.04-0.26)	1.92	0.16 (0.04-0.27)	1.89	0.84 (0.73-0.96)	0.92	0.90 (0.80-0.99)	0.97	0.50 (0.34-0.66)	0.97
female	336	0.09 (0.06-0.12)	0.70	0.07 (0.04-0.10)	0.52	0.93 (0.90-0.96)	1.08	0.92 (0.89-0.95)	1.04	0.45 (0.40-0.50)	0.85
male	512	0.12 (0.09-0.15)	1.00	0.14 (0.11-0.17)	1.00	0.86 (0.83-0.89)	1.00	0.89 (0.86-0.91)	1.00	0.53 (0.48-0.57)	1.00
						Alzheimer's Disease					
dataset: adni	129	0.15 (0.09-0.22)	1.00	0.15 (0.08-0.21)	1.00	0.86 (0.79-0.92)	1.00	0.75 (0.67-0.82)	1.00	0.52 (0.43-0.61)	1.00
dataset: edsd	67	0.31 (0.20-0.42)	2.05	0.23 (0.13-0.33)	1.60	0.77 (0.67-0.87)	0.90	0.91 (0.85-0.98)	1.23	0.36 (0.24-0.47)	0.69
age cat: 50-59	7	0.50 (0.13-0.87)	2.33	0.4 (0.04-0.76)	2.71	0.60 (0.24-0.96)	0.70	1.00 (1.00-1.00)	1.42	0.29 (0.00-0.62)	0.66
age cat: 60-69	39	0.15 (0.04-0.26)	0.70	0.15 (0.04-0.26)	1.02	0.85 (0.74-0.96)	1.00	0.90 (0.80-0.99)	1.27	0.49 (0.33-0.64)	1.12
age cat: 70-79	108	0.21 (0.14-0.29)	1.00	0.15 (0.08-0.21)	1.00	0.85 (0.79-0.92)	1.00	0.70 (0.62-0.79)	1.00	0.44 (0.34-0.53)	1.00
age cat: 80+	42	0.20 (0.08-0.32)	0.93	0.26 (0.13-0.40)	1.78	0.74 (0.60-0.87)	0.86	0.87 (0.77-0.97)	1.24	0.55 (0.40-0.70)	1.26
female	99	0.18 (0.10-0.26)	1.00	0.18 (0.11-0.26)	1.00	0.81 (0.74-0.89)	1.00	0.82 (0.74-0.90)	1.00	0.51 (0.41-0.60)	1.00
male	97	0.24 (0.16-0.33)	1.36	0.18 (0.10-0.26)	0.97	0.82 (0.75-0.90)	1.01	0.76 (0.67-0.84)	0.92	0.42 (0.32-0.52)	0.84

masks are shown in Figure C.9. These masks are characterised by a high degree of noise which was also observed using other interpretability techniques such as Saliency maps and SHAP. Because of this noise it was not possible to find any useful clustering approach.

C.3.2. Diagnostic prediction of AD

The SHAP method was implemented using TreeSHAP, an implementation of the method designed to efficiently calculate a local additive explanation for tree based models. We used the *tree_path_dependent* option, which does not require any fitting on the training set. The method generates an additive explanation in the following form

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j. \tag{3}$$

Where g is the local explanation, $\boldsymbol{z}$ is a vector containing binary values corresponding to the presence of simplified explanation variables and ϕ_j is the attribution to each of these variables. The explanations generated by SHAP have been combined with anchors, which are local rules that "anchor" the prediction to its output value (i.e. the features that should not be changed to keep the prediction the same). The definition of an anchor A is the following

$$\mathbb{E}_{D_x(z|A)}[\mathbf{1}_{f(x)=f(z)}] \leq \tau, A(x) = 1. \tag{4}$$

Where $D(z|A)$ is the distribution of neighbours which follow the anchoring rule A. Anchors and SHAP values can provide the model owner with complementary information. SHAP values, for example, can quantify the effect of a feature on the prediction but do not provide

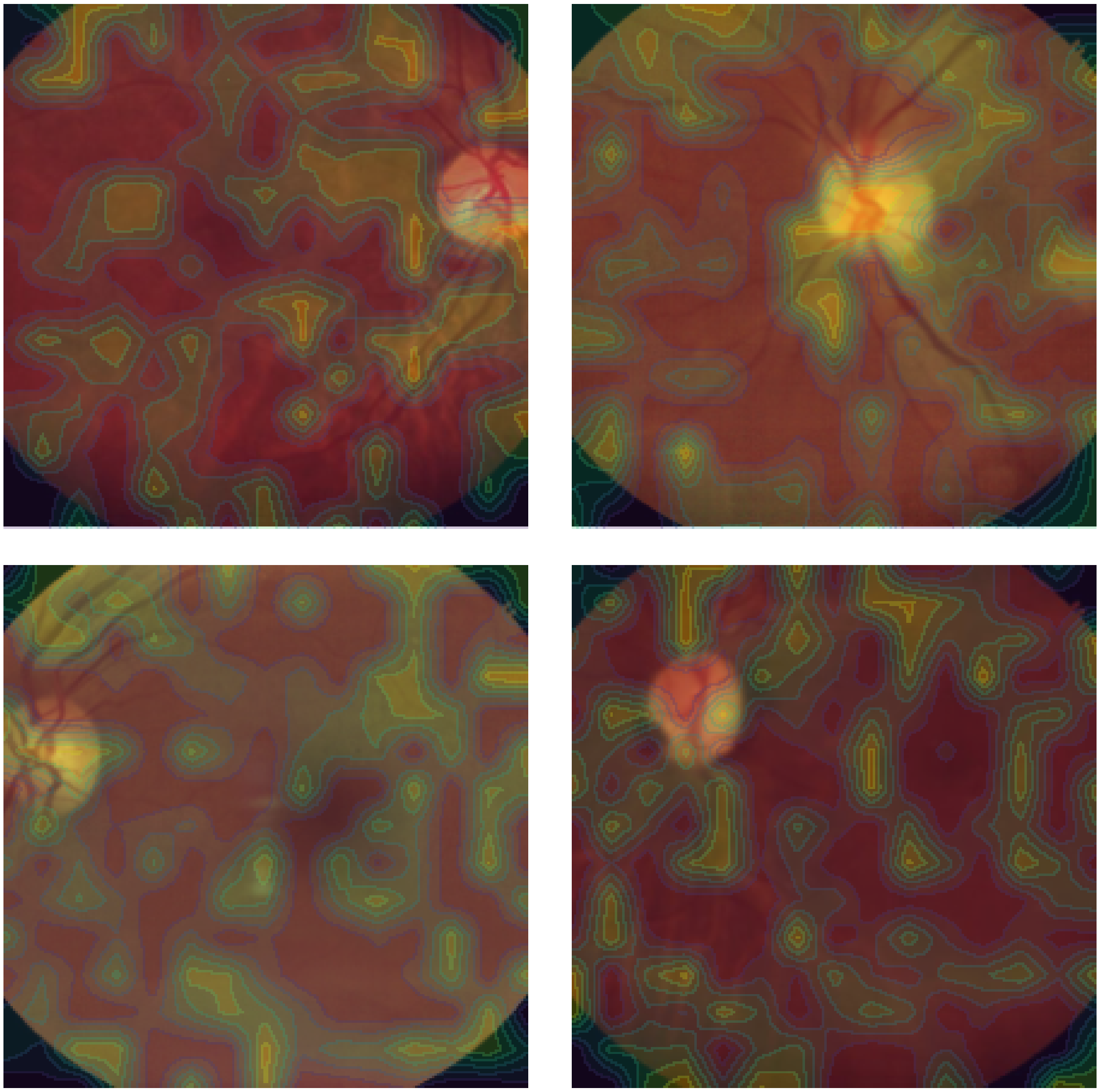


Figure C.9: Perturbation masks generated using the *Meaningful Perturbation* approach. Pictures show the degree of noise characterizing these explanation masks and the different rotations occurring in the original images.

information about local dependencies, which can be understood by using anchoring rules. We generated anchors using the method implemented in the original paper, by specifying a maximum rule length equal to 6 and a precision threshold of 95%. The selection of the most representative explanations was done by first grouping the assessment sets according to their position in the confusion matrix and to the subgroups identified during the bias analysis. Once each subset of points was defined we applied a DBSCAN clustering algorithm using a cosine similarity metric and a cluster radius of $\epsilon = 0.3$. Figures C.10,C.11 and C.12 show

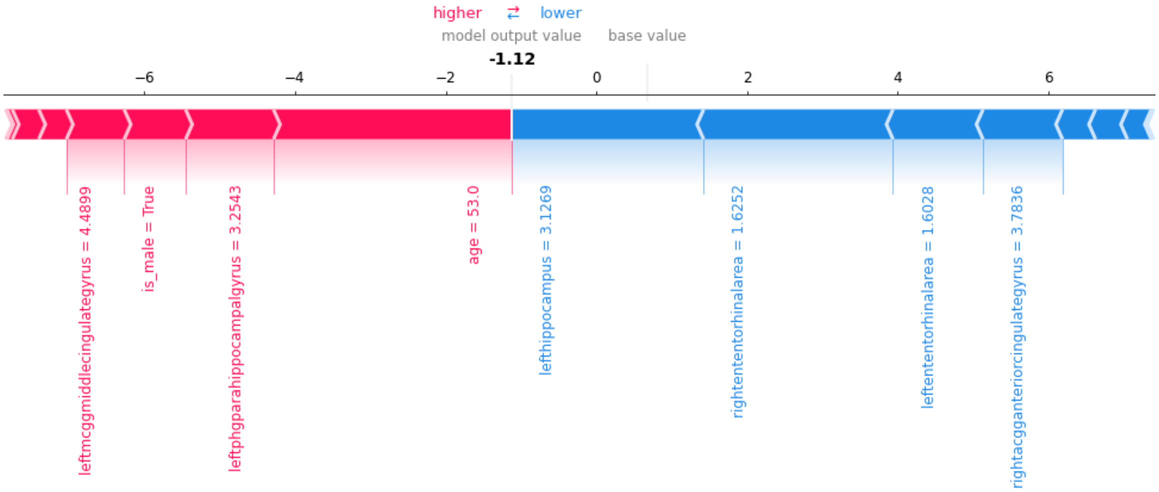


Figure C.10: Representative SHAP explanation for a False Negative prediction in the 50-59 age group. Corresponding anchoring rule is:
IF (rightententorhinalarea > 1.62) & (leftententorhinalarea > 1.59) & (rightacgganteriorcingulategyrus <= 3.83) THEN PREDICTION IS FALSE

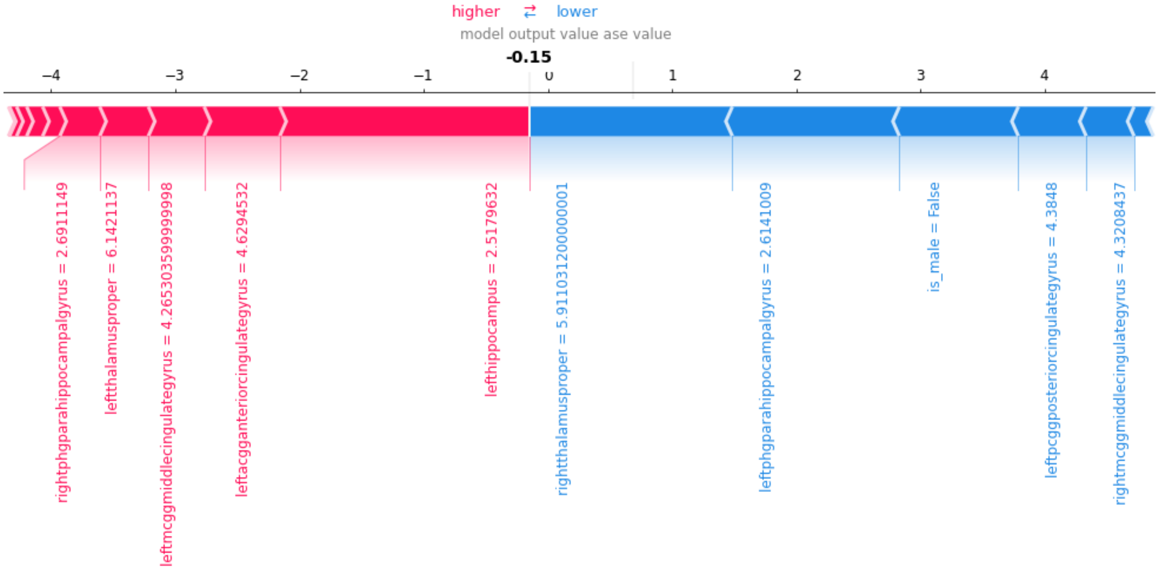


Figure C.11: Representative SHAP explanation for a False Negative prediction in the 60-69 age group. Corresponding anchoring rule is:
IF (rightententorhinalarea) > 1.46 & (leftententorhinalarea > 1.35) & (leftphgparahippocampalgyrus <= 2.73) & (rightthalamusproper <= 6.52) THEN PREDICTION IS FALSE

the SHAP force plots for the most representative false negatives subdivide according to their age group.

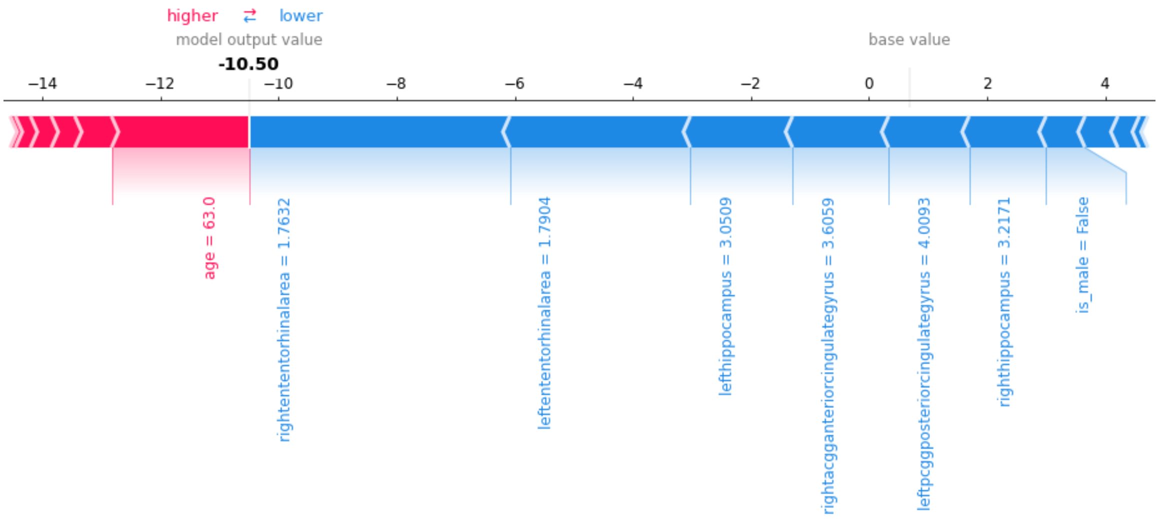


Figure C.12: Representative SHAP explanation for a False Negative prediction in the 60-69 age group. Corresponding anchoring rule is:
IF (rightententorhinalarea) > 1.72 & (leftententorhinalarea > 1.69) & (righthippocampus <= 3.22) THEN PREDICTION IS FALSE

C.3.3. Cytomorphologic classification to aid leukemia diagnostics

The explanations for this model were generated using the Grad-CAM method. Grad-CAM produces a coarse localization map to highlight region importance for a given activation target based on the last convolutional layer of a CNN. We implemented it by using the the *keras-vis* library. The clustering on the assessment set has been performed by first grouping images based on their position in the confusion matrix then by clustering each group according to their softmax outputs. We performed the clustering using DBSCAN with a cosine similarity metric and an $\epsilon = 0.05$.

C.4. Robustness assessment

In the context of machine learning for medical diagnosis we aim for a model f that is robust with respect to (random) variations, naturally occurring during data acquisition. These natural variations ought to be defined for medical data. For classical image recognition a set of corruption functions C and a set of perturbation functions $\mathcal{E}$ is provided in Hendrycks and Dietterich (2019), to define corruption robustness as the expected probability that $f(c(x)) = y$ holds true for label y and corruption c, whereas perturbation robustness is defined as the expected probability that $f(\varepsilon(x)) = f(x)$ holds true for the perturbation ε. Since neither C nor $\mathcal{E}$ represent typical variations occurring in medical data we use the corruption functions in C to analyse both: corruption robustness and perturbation robustness. Moreover, we use the term perturbation to describe these two modes of robustness. The key aspect of this analysis is not the specific choice of perturbation methods, but to describe a general procedure after perturbation methods are fixed (potentially by some domain experts). With this in mind we apply in a first step different perturbations with three severity levels to our model, record the perturbed model output and compare it to the clean model output in three different ways: visualize the (approximated) distribution density of the probability scores of the correct classified points. Visualize the (approximated) distribution density of the probability scores of the incorrect classified points. Visualize the (approximated) distribution density of all probability scores. In a second step we measure the effect of these perturbations by the mean corruption error (mCE) proposed in Hendrycks and Dietterich (2019). This is an empirical metric, unlike more theoretical approaches to calculate robustness bounds as the Lipschitz constant (Szegedy et al., 2013), which can be too loose or hard to calculate for complex models. In Hendrycks and Dietterich (2019) the mean corruption error is calculated as follows: For a trained classifier f calculate the top-1 error rate for each perturbation type c, with a varying level of severity s, written as $E^f_{s,c}$. The model's error rate tested with the unperturbed observations is defined as E^f_{clean}. Authors in Hendrycks and Dietterich (2019) recommend a normalization using a baseline to account for the different degrees of difficulty. In this work, as we test one completely black box model with access only to test data, we normalize it against the error rate of the tested model with the perturbations and define the corruption error according to:

$$\text{CE}^f_c = \left(\textstyle\sum_{s=1}^S E^f_{s,c} - E^f_{\text{clean}}\right) / \left(\textstyle\sum_{s=1}^S E^f_{s,c}\right). \tag{5}$$

The mean corruption error mCE is calculated by averaging over the corruption errors.

C.4.1. Diabetic Retinopathy model

The retinopathic detection problem is a classification problem based on unstructured data input. These unstructured data are composed of images from the ophthalmology domain, treated as a classical image pattern recognition problem. Unstructured data comes from different scopes, which the quality of this can be very low. Also consider that there may be disturbances that occur in daily life of the domain that can alter the decision of the model. For this reason, IA engineers incorporate these corruptions into transformations in the preprocessing data augmentation process in order to simulate these corruptions, to improve the inference of the model. In this use case, the owner of the model, use in total seven different types of transformations to augment the data set, the outlines transformations are flip, rotate, zoom, contrast, brightness, lighting and sketch, without damaging the image semantic domain. For this reason in this section we outline perturbation methods on image data based on Hendrycks and Dietterich (2019) to evaluate the model robustness. For image data we have a three channel images (red, green, blue). The model input consists of this color images.

Perturbation of images We apply perturbation to images, where the feature space of the normal images are from 0 to 255 in the three channels of color (RGB), we have to make sure that this perturbed images are from the same color feature space. In the experiments we apply in total 16 perturbations on four different background to the images using the functions from the Github repository of Hendrycks and Dietterich (2019): brightness, contrast, elastic transform, fog, frost, Gaussian blur, jpeg compression, pixelate, saturate, speckle noise, spatter, Gaussian noise, shot noise, impulse noise, defocus blur and motion blur. For each perturbation we apply three severity level: 0, 2, 4 on all 843 images in the test set spread along all the classes, where the data was provided by the use case owner. On the first experiment we expose the elastic transform perturbation to the input data C.13.

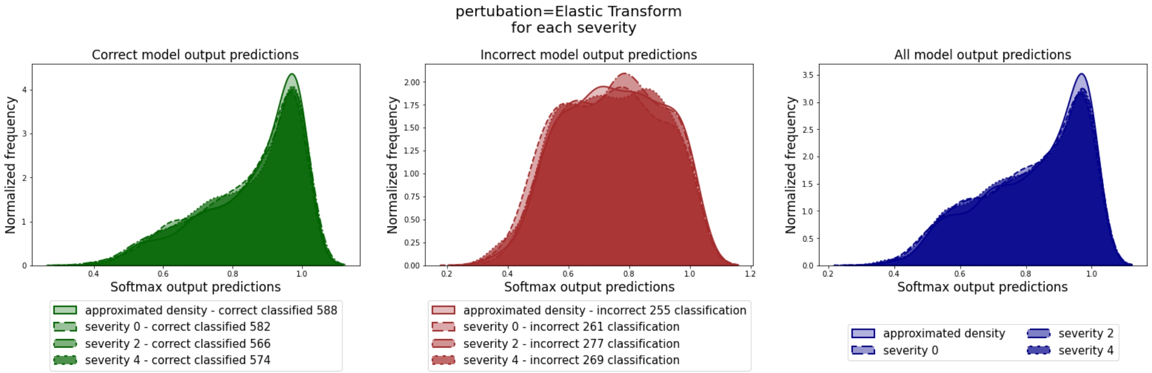


Figure C.13: Perturbation: Elastic transform. For Diabetic Retinopathy model. The densities of the probability scores of the perturbed model differ only slightly compared to the clean versions, along all three plots.

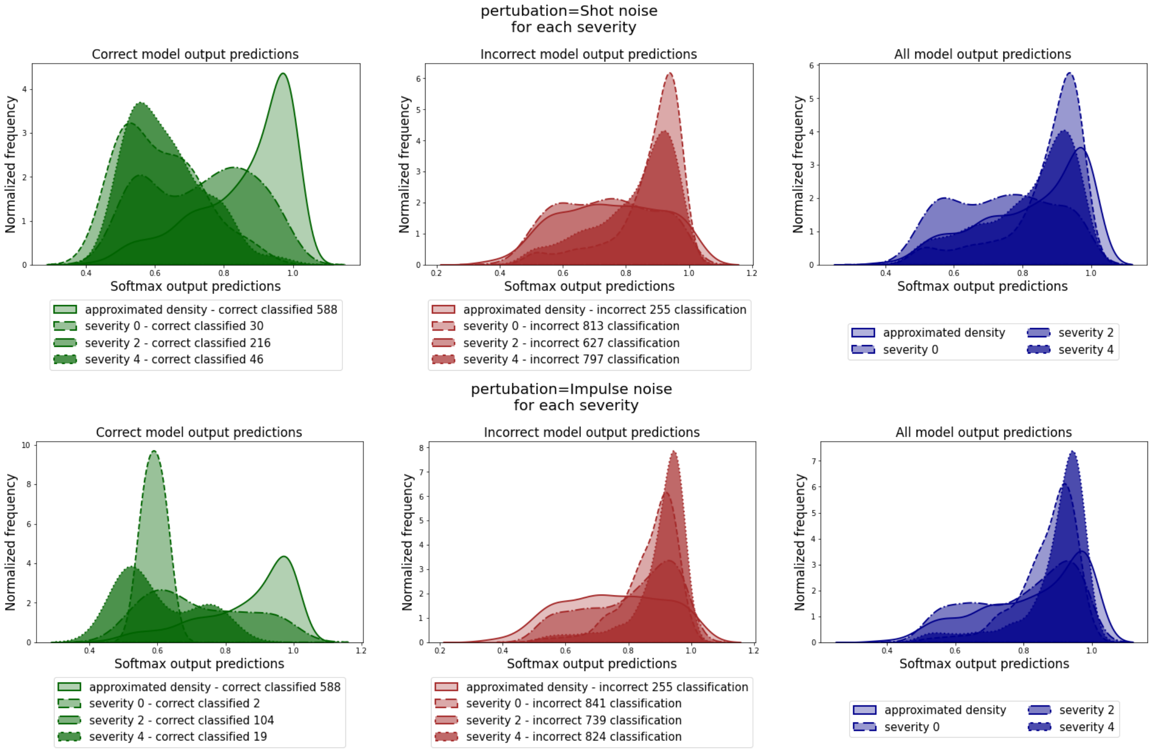


Figure C.14: Noise corruptions. For Diabetic Retinopathy model. The densities of the incorrect classified points included the non gradable class for each severity tends to shift to higher values. The correct classified points shift to lower values or flatting thought the noises.

We call the model robust regarding to the perturbed features with the three severity, we can observe the model only predicts six wrong outputs for severity 0, setting the accuracy to drop 1%. Has a lightly effect on the model output. For the experiments using noise

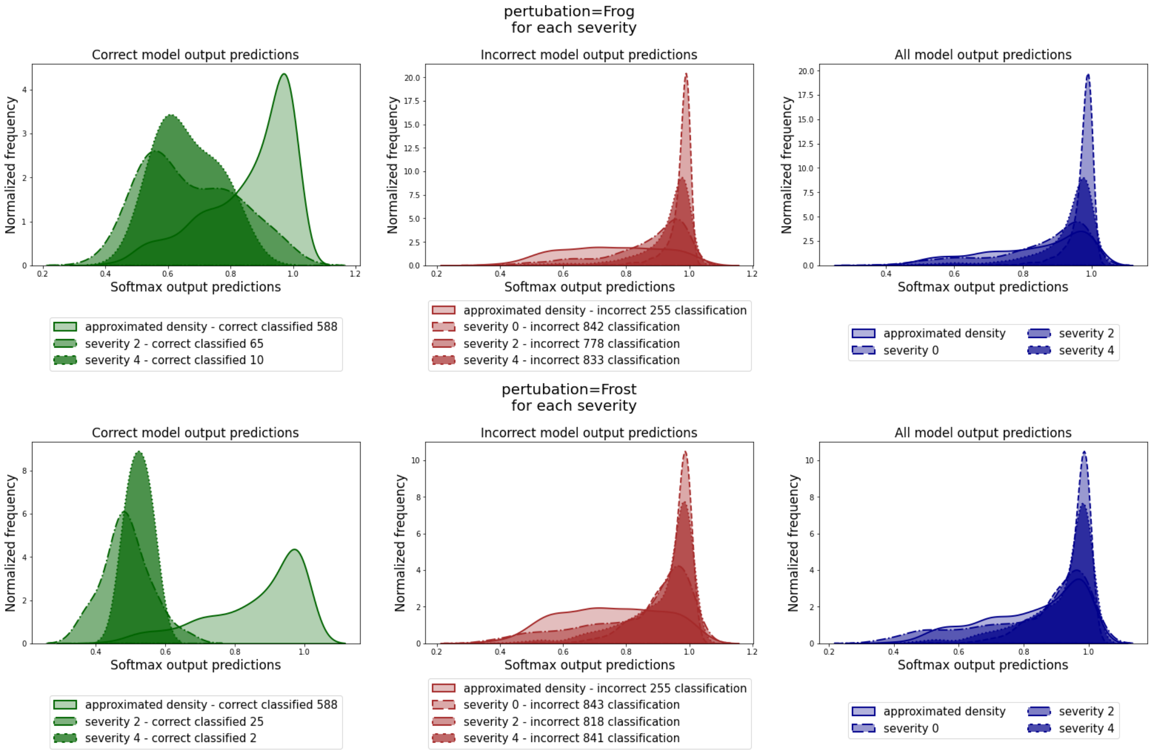


Figure C.15: Weather corruptions. For Diabetic Retinopathy model. There are a few points correct when the data is exposed to the weather corruptions where the incorrect classified points shift highest values for incorrect predictions outputs. The use case owner addressed that this kind of perturbations is unseen in ophthalmology domain.

perturbations as in figure C.14, the effect has an immense impact on the model outputs predictions, setting the model to behave insecure and classified the outputs as a non gradable (image quality not sufficient for diagnosis), in this kind of situation we record this outputs into incorrect classified point, consider that the image quality may produced wrong prediction outputs of the model even for an expert in this domain field. Some of the corruptions exposed, there are some perturbations that are atypical in the ophthalmology domain, the use case owner listed some of them: fog, frost addressed in the figure C.15, high level of contrast and brightness, Jpeg compression in the figure C.16 can damage the semantic domain of the data, where not even an expert of this domain has a good diagnosis of the image. For some corruptions of noise, the effect can be reproduce of the dust in the camera where is taken the image for diagnosis as can see it on the figure C.14, also for the blur corruption cause by the unfocused or the movement of the camera when the picture is taking, the results are addressed in the figure C.17. We also labeled some extra corruption function to evaluate the model robustness, we condensed this results in figure C.18.

Mean corruption error The metric to evaluate the diabetic retinopathy model robustness is defined in formula 5. We apply in total 16 perturbation functions for severity $s \in \{0, 2, 4\}$ with $E^f_{\text{clean}} = 0.2232$ and record the results of the experiment in ascending order in 1. The mean corruption error along all perturbations is $mCE = 0.6336$.

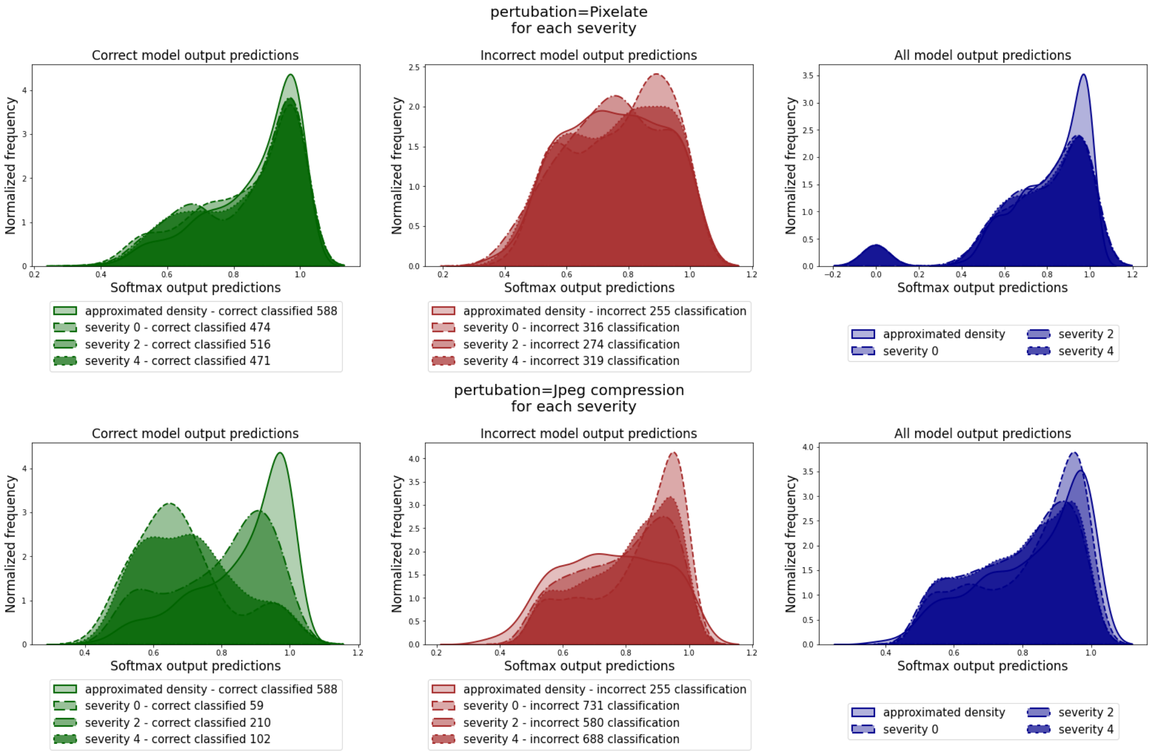


Figure C.16: Digital corruptions. For Diabetic Retinopathy model. Some digital corruptions are possible in ophthalmology domain, the image compression, quality. For example the pixelate and jpeg compression, where the pixelate corruption have a small impact in the model decision against the jpeg compression. The jpeg compression have a strong effect on the model decision outputs where the correct classified points tends to sketch to the left or right side.

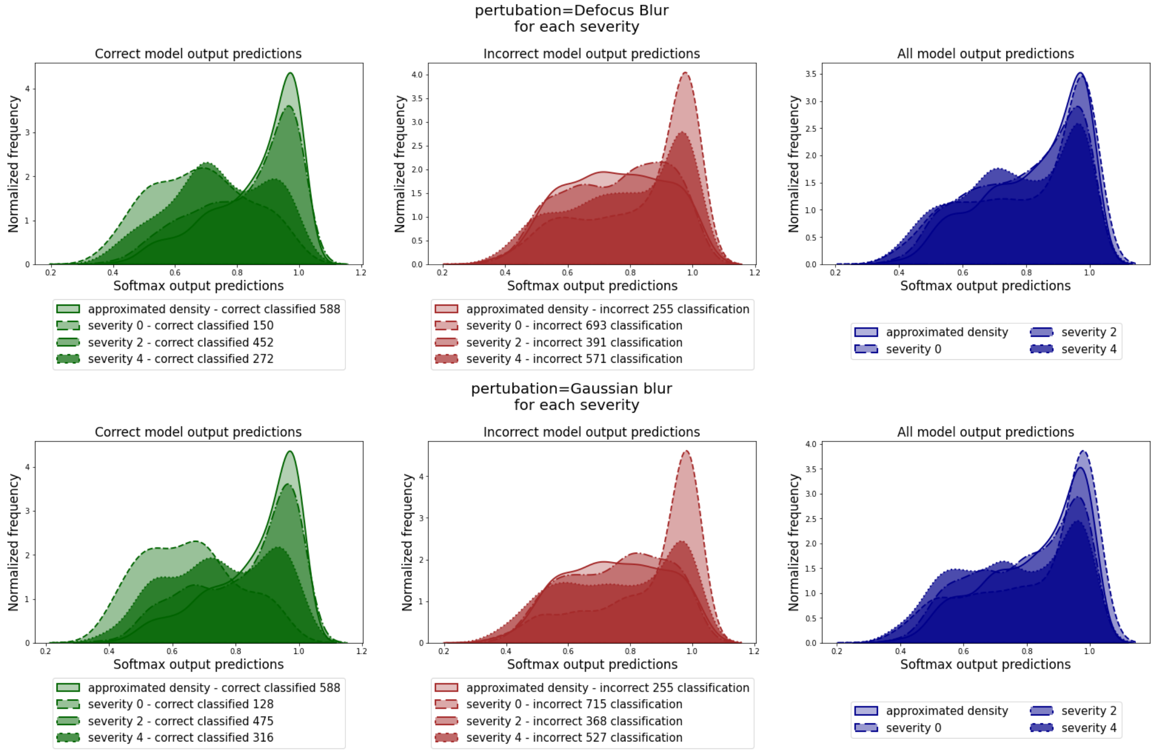


Figure C.17: Blur corruptions. For Diabetic Retinopathy model. The bur corruption has strong influence on the model correct classified points, the distribution density flatter to lower values for each severity levels.

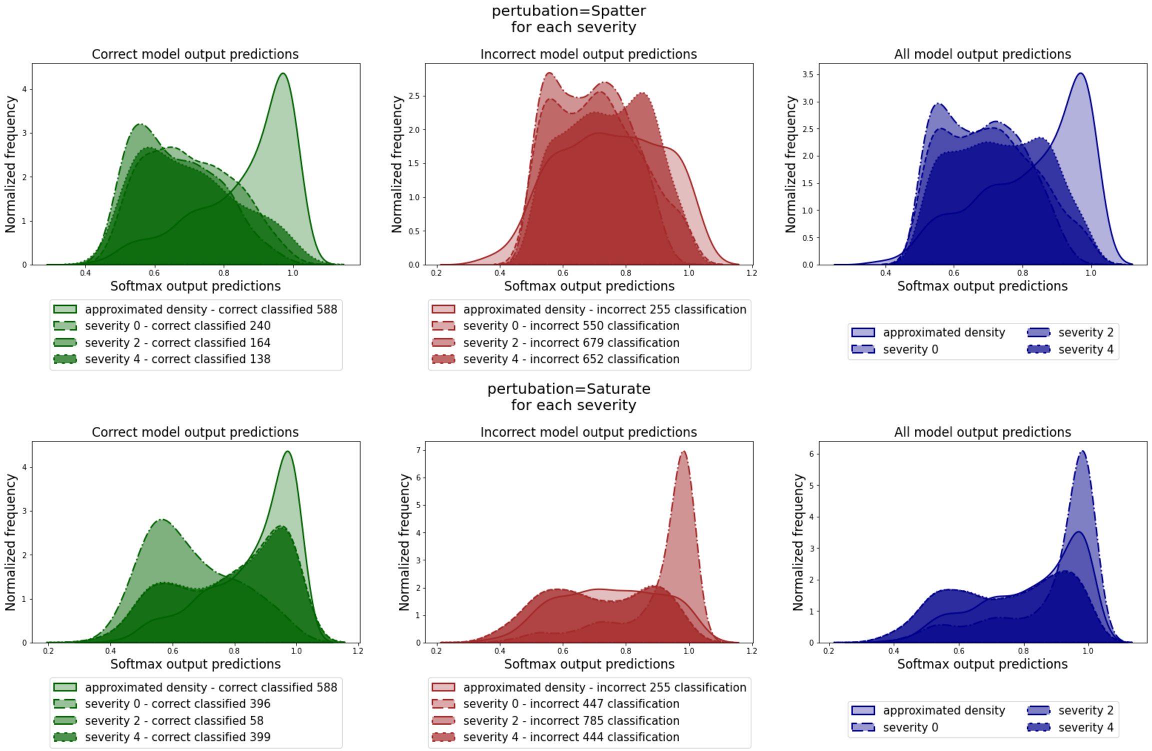


Figure C.18: Extra corruptions. For Diabetic Retinopathy model. These corruptions have a strong influence, sketching the correct classified points distribution density values to lower values. In the other hand the incorrect classified points are flatter in the space. We might call the model not robust against this extra perturbations, where the correct classified points start to drop.

C.4.2. AD model

In contrast to the Diabetic Retinopathy model and the cytomorphology model, the input for the AD model is structured data. Since this structured data is extracted from MR-images, the perturbations should be applied directly to these images rather than to the extracted features. Nevertheless, test engineers do not always have access to these images. For that reason we outline in this section perturbation methods for tabular data based on Eduardo et al. (2019). For tabular data we distinguish between categorical features (that can only take a finite number of values) and continuous features (that can take uncountable many values). The AD model input consists of fifteen continuous features: age and fourteen brain features and one categorical feature: gender.

Perturbations of continuous features If we apply perturbations to images, we need to ensure that the perturbed pixels are still in the used color range. With continuous features in tabular data we face the same problem, but in general, we do not know the ranges of the features. To ensure that the perturbed data is still in the feature space, we choose the range for each feature according to the maximum and minimum value of this feature in the test set. Assuming continuity, we ensure that the perturbed features end up in the feature space, by cutting of higher and lower values after sampling. We apply three different types of noise to the continuous features: Gaussian noise (Gn), Laplace noise (Ln) and log-normal noise (lln) with three different severity levels. In order to determine the variance of these distributions, we calculate the standard deviation $\hat{\sigma}^k$ of feature k in the test set and define the variance as $\sigma^k := s \cdot \hat{\sigma}^k$ with respect to some scaling s. Figure C.19 shows the effect of the continuous perturbation, applied to all brain features.

Perturbations of the categorical feature gender There is only one categorical feature in the model: gender. The authors in Eduardo et al. (2019) propose the following perturbation method for categorical features: Replace the categorical feature of data point x by a discrete random variable over its C possible values. The probability to draw value $c \leq C$ is given by $\frac{p_c^\beta}{\sum_{c=1}^C p_c^\beta}$ with scaling factor $\beta \in [0, 1]$, where p_c is the relative frequency of the value c in the test set. Note that for $\beta = 0$ we sample from a uniform distribution and for $\beta = 1$ we sample according to the relative frequency. So β can be seen as a scaling factor on how much we take the relative frequency into account. Further, the authors propose to leave out the true value of the feature in x, so that each categorical feature is perturbed. In our case this approach leads to a deterministic perturbation, where we replace each gender by its counterpart. We define the first categorical perturbation for the feature gender according to the approach above, but consider the true value of each feature as well during sampling. The effect of this perturbation is shown in figure C.20.

In the second perturbation method we first assign male to each feature, second assign female to each feature and third replace each gender by its counterpart.The effect of this perturbation is visualized in figure C.21.

For both categorical perturbations, the perturbed model differs very mildly in prediction and in overall model output from the clean model. This observation holds among all applied severity levels. For that reason we might call the AD model robust with respect to the feature gender.

Mean corruption error To quantify the effect of the perturbations we calculate the mean corruption error over all eight perturbations: the three continuous perturbations

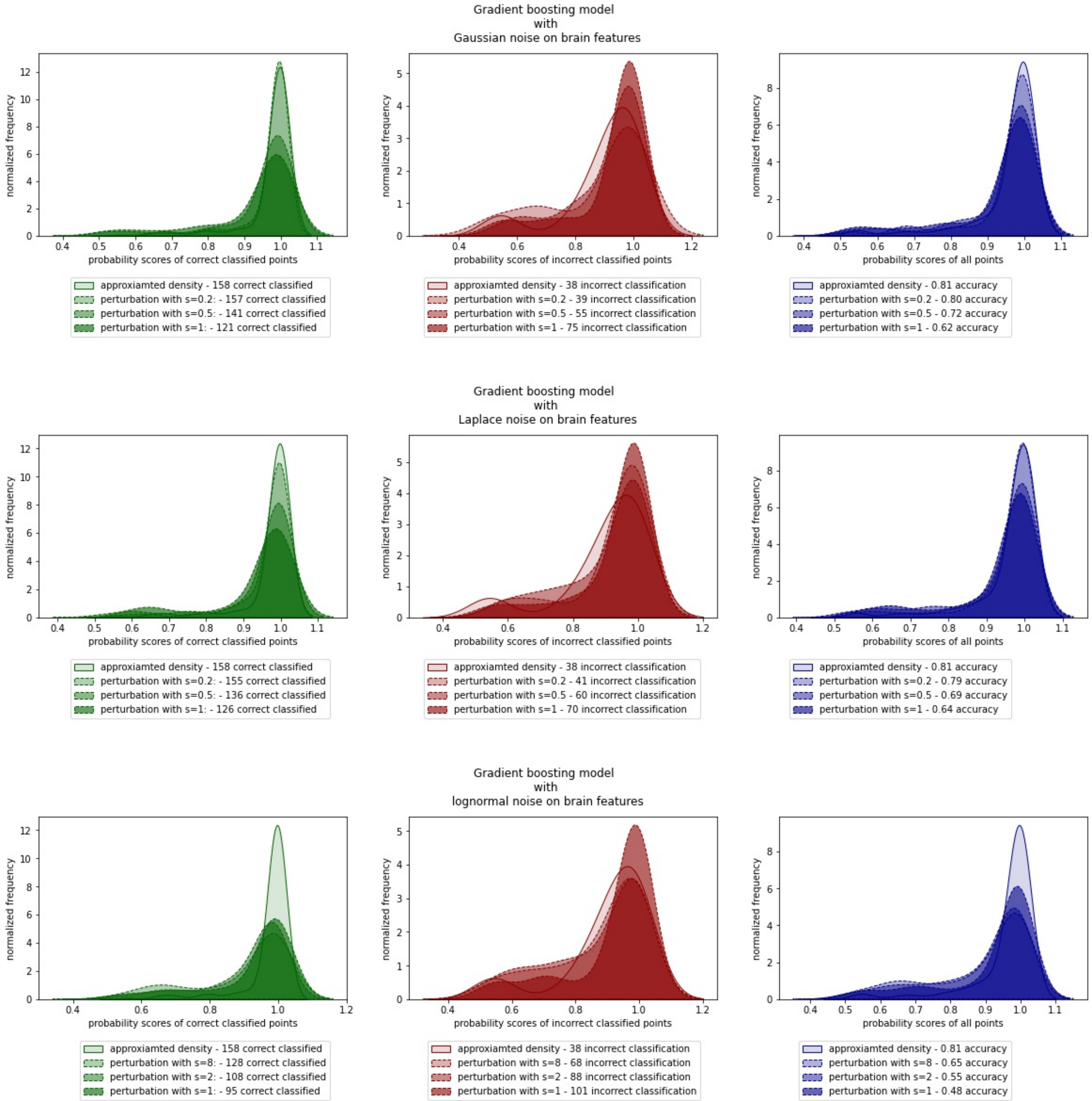


Figure C.19: The distribution densities of the probability scores flatten throughout all noise types. The scores of the correct classified points shift to lower values. For Gaussian noise and Laplace noise of scale 0.2 we might call the model robust, since accuracy changes only by one point. For all other noise types and scaling levels, we observe a decrease of accuracy at least by 9%. The perturbation of lognormal noise with $s \in \{1, 2\}$ diminishes the accuracy of the model close to random guessing and worse.

specified above, applied to age and brain features with $s \in \{0.2, 0.5, 1\}$ for Gn and Ln and $s \in \{1, 2, 8\}$ for lln. The categorical perturbation with $\beta \in \{0, 0.5, 0.8\}$ and the deterministic perturbation. The corruption errors of these perturbations calculated according to (5) are listed in C.4 in ascending order. The mean corruption error differs (slightly) from

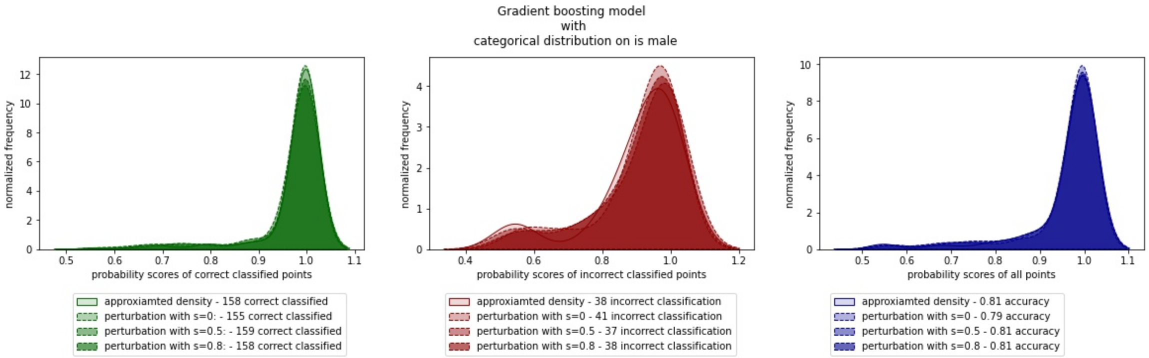


Figure C.20: The effect of the categorical perturbation is slightly visible in the distribution densities of the probability scores. The perturbed prediction differs at most by three points from the clean prediction.

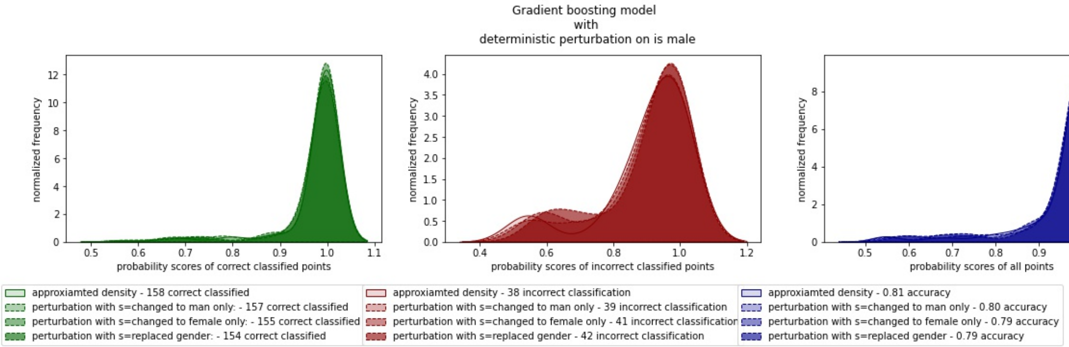


Figure C.21: Setting the gender for all medical representatives first to male and second to female leads only to a decrease of incorrect classified points by two and three. Replacing the gender by its counterpart leads to four more incorrect classified points.

trial to trial due to the stochasticity of the perturbations. The calculations in C.4 can be reproduced with the plots and statistics provided in the folder `AD_robustness_analysis` at https://github.com/luisoala/daisam. Averaging over all these errors leads to a mean corruption error of $mCE = 0.1766$.

Table C.4: Robustness metrics for the AD model with $E^f_{\text{clean}} = 0.1939$

Perturbation c	$\sum_{s=1}^{3} E^f_{s,c}$	$\sum_{s=1}^{3} E^f_{s,c} - E^f_{\text{clean}}$	CE^f_c
Gn on age	0.5765	-0.0051	-0.0088
Ln on age	0.5816	1.1E-16	1.9E-16
lln on age	0.6020	0.0204	0.0339
categorical perturbation	0.6071	0.0255	0.0420
deterministic perturbation	0.6224	0.0408	0.0656
Gn on brain features	0.9031	0.3214	0.3559
Ln on brain features	0.9031	0.3214	0.3559
lln on brain features	1.3469	0.7653	0.5682

C.4.3. Cytomorphology model

To run the robustness experiments a ResNeXt model was provided by the use case owner, details on the training process are explained extensively in Matek et al. (2019). The model was used to classify RGB images of size 400 × 400 into 15 different classes. The test set consisted of 73 images distributed along all the classes, the data was provided by the use case owner.

Image perturbation The perturbations applied to this model test set, where the ones proposed by Hendrycks and Dietterich (2019), the images on the test set where RGBA images so in order to apply the perturbations and the inference, the alpha channel was ignored for the perturbation process and also the inference. To simplify the analysis, three perturbation severity levels where used, 0, 2 and 4. Variation on the effect of the perturbations, exposed the need of feedback from the use case owner in order to select transformations that modeled real case scenarios to which the model should be robust to, this led to ignore the results of the contrast, fog, frost, motion blur and saturate perturbations.

Digital transformations like pixelation, jpeg compression and different brightness levels were selected as the most plausible perturbations, for the pixelation and jpeg compression the model accuracy and softmax output values were very similar as seen in C.22, but the brightness perturbation produced a spike on the incorrect classifications and its confidence. This shows that the model is susceptible to variations in the brightness of the images and this kind of perturbation may produce overconfident results.

Mean corruption error To calculate the metric, all 16 perturbations were evaluated on its three degrees of severity, the results are presented in table C.5.

Table C.5: Robustness metrics for the Cytomorphology model with an $E^f_{\text{clean}} = 0.2740$

Perturbation c	$\sum_{s=1}^{3} E^f_{s,c}$	$\sum_{s=1}^{3} E^f_{s,c} - E^f_{\text{clean}}$	CE^f_c
Elastic transform	0,7671	-0,0548	-0,0714
Pixelate	0,8630	0,0411	0,0476
JPEG compression	0,9315	0,1096	0,1176
Motion blur	1,1507	0,3288	0,2857
Defocus blur	1,1781	0,3562	0,3023
Gaussian blur	1,2055	0,3836	0,3182
Spatter	2,1233	1,3014	0,6129
Gaussian noise	2,1918	1,3699	0,6250
Frost	2,4110	1,5890	0,6591
Speckle noise	2,5068	1,6849	0,6721
Shot noise	2,534	1,7123	0,6757
Impulse noise	2,5342	1,7123	0,6757
Brightness	2,6712	1,8493	0,6923
Contrast	2,6986	1,8767	0,6954
Saturate	2,7808	1,9589	0,7044
Fog	2,8767	2,0548	0,7143

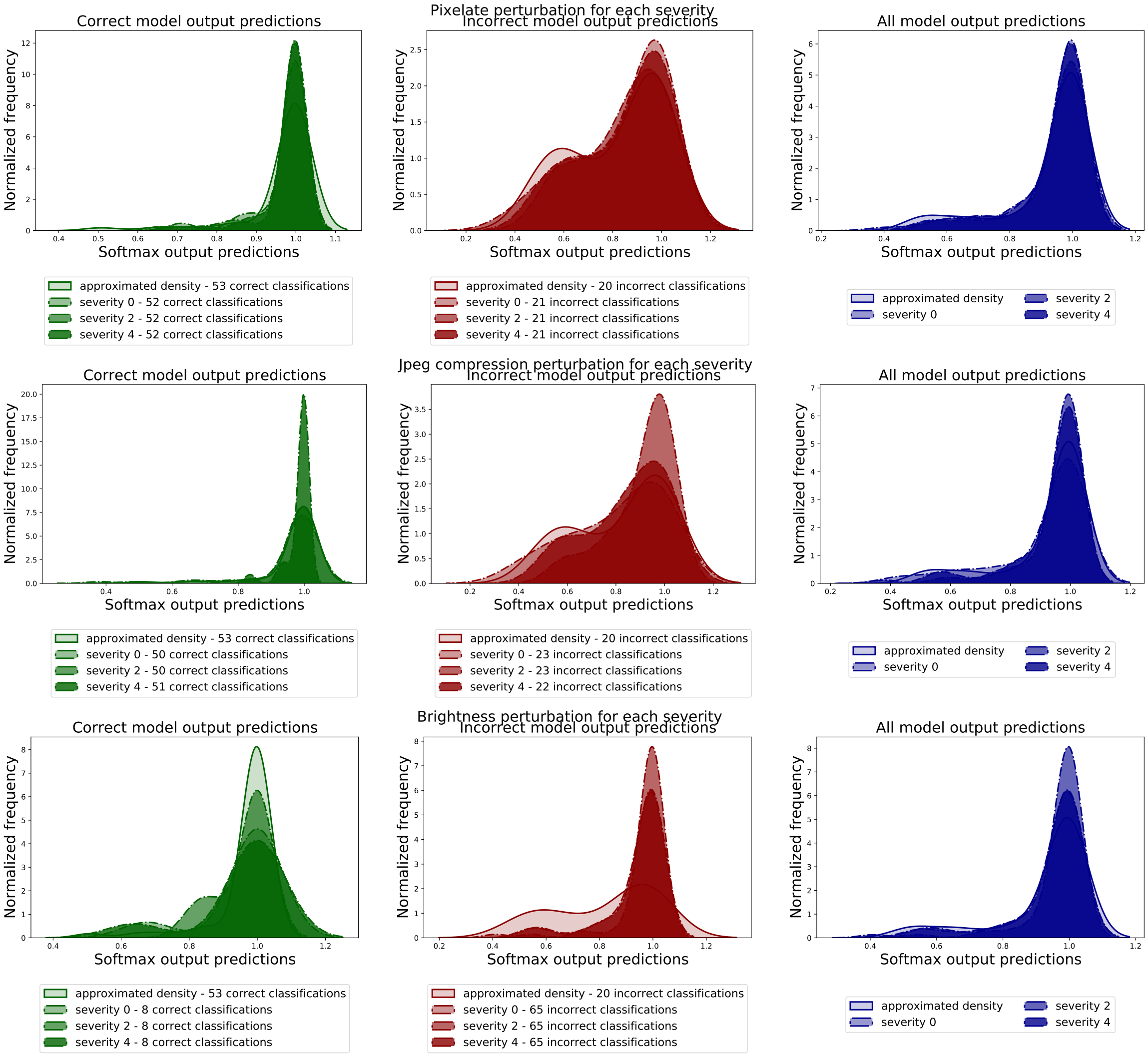


Figure C.22: Digital corruptions on the cytomorphology dataset.

Appendix D. FG-AI4H Overview

The FG-AI4H "works in partnership with the World Health Organization (WHO) to establish a standardized assessment framework for the evaluation of AI-based methods for health, diagnosis, triage or treatment decisions" according to the initiative's website. To that end FG-AI4H has produced reference documentation for the assessment process which is summarized in Figure 1 of the main paper. In the following, we provide short summaries of what each of these documents aims to achieve. The full documents can be accessed via the ITU collaboration environment at https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/SitePages/Home.aspx[3].

D.1. ITU/WHO FG-AI4H reference documents

AI4H ethics considerations (DEL 01) This document discusses the ethical issues and challenges posed by digital technologies including AI/ML technologies and tries to provide harmonised ethics guidance for the design and implementation of AI in global health especially for the benefit of how low- and middle-income countries.

Good practices for health applications of machine learning: Considerations for manufacturers and regulators (DEL 2.2) This document presents a set of good machine learning practice guidelines intended to educate the developers and manufacturers of healthcare AI solutions on conducting a comprehensive requirements analysis and streamlining conformity assessment procedures for continual product improvement in an iterative and adaptive manner in conformance to the appropriate standards and regulations.

Data requirements (DEL 5.1) This document presents the acceptance criteria for data submitted to the FG-AI4H and states the governing principles and rules. These principles are crucial because the core of the benchmarking framework for AI for health methods will be an undisclosed test data set – per use case of each topic area to be defined – that will not be made accessible to the AI developers. combines a set of four deliverables as umbrella.

Data acquisition (DEL 5.2) This document presents a framework for public healthcare data acquisition and management model based on standard protocol for its easy adoption by any country or international health organizations.

Data annotation specification (DEL 5.3) This document provides general guideline of data annotation specification, including definition, background and goals, framework, standard operating procedure, scenario classifications and corresponding criteria, as well as recommended metadata, etc.

Training and test data specification (DEL 5.4) This document provides guidelines on the systematic way of preparing technical requirements specification for datasets used in training and testing of ML models and discusses the best practices of data quality assurance aimed at minimizing the data error risks during the training and test data preparation phase of machine learning process lifecycle.

Data handling (DEL 5.5) This document provides information on the data handling policies, the need for a factual framework in compliance with data protection laws and regulations dealing with the use of personal health data. It also outlines how data will be handled, once they are accepted.

[3] An ITU user account is necessary for access which we obtained free of charge like so https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/registrationsteps.pdf

Data sharing (DEL 5.6) This document outlines the established data sharing methods and novel methods based on distributed and federated environments for privacy preserving AI/ML models. The scope of this document includes a description of all the necessary steps and requirements to enable secure data sharing and the specifications of the role of the data providers, data processors and the data receivers.

Topic Description Document(TDD) (DEL 10) This document specifies the requirements of a standardized benchmarking process for the specific AI-based use case of the respective Topic Group within the AI4H Focus Group. It covers all scientific, technical, and administrative aspects relevant for setting up this benchmarking.

Data and artificial intelligence assessment methods (DAISAM) reference (DEL 7.3) This document provides a comprehensive overview and discusses in detail, the data and ML model quality assessment methods for evaluating the bias, interpretability, and robustness for different use cases represented by the Topic Groups.

Clinical evaluation of AI for health (DEL 7.4) The document provides guidelines for the evaluation of AI in health for use by researchers, clinicians/patients, developers, and policy makers with a framework to understand whether the models are safe, effective and cost- effective, and also to compare model performance with current standards of care, and between each other to facilitate clinically meaningful improvements in a complex clinical environment.