Open Code Initiative
Evaluation Package

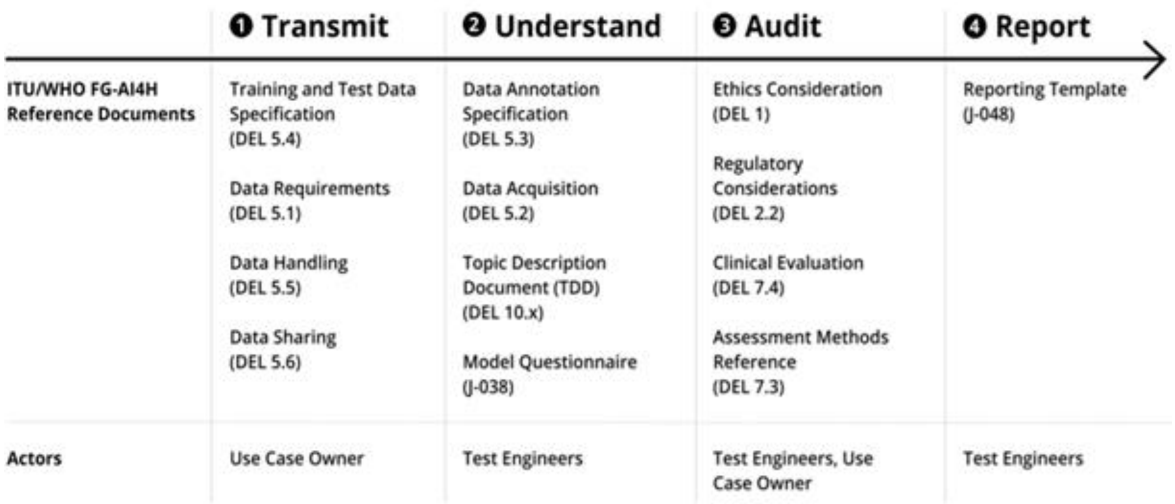
WG-DAISAM

Elora Schörverth, Fraunhofer HHI
Steffen Vogler, Bayer AG
Pradeep Balachandran, Ehealth Consultant
Alixandro Werneck Leite, LAMFO UnB
Danny Xie Li, Tecnológico de Costa Rica
Kamran Ali, Hasso Plattner Institute
Luis Oala, Fraunhofer HHI

WG-DAISAM, Evaluation and Reporting: From Paper to Practice - to Software

2019

Paper

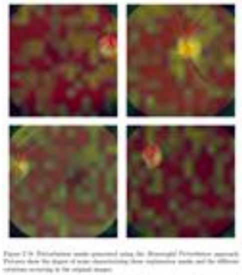


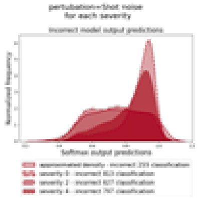
	❶ Transmit	❷ Understand	❸ Audit	❹ Report
ITU/WHO FG-AI4H Reference Documents	Training and Test Data Specification (DEL 5.4) Data Requirements (DEL 5.1) Data Handling (DEL 5.5) Data Sharing (DEL 5.6)	Data Annotation Specification (DEL 5.3) Data Acquisition (DEL 5.2) Topic Description Document (TDD) (DEL 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

2020

Practice

Oala, Luis, Jana Fehr, Luca Gilli, Pradeep Balachandran, Alixandro Werneck Leite, Saul Calderon-Ramirez, Danny Xie Li et al. "**ML4H Auditing: From Paper to Practice**." In *Machine Learning for Health*, pp. 280-317. PMLR, 2020.




pertubation=Shot noise
for each severity
incorrect model output predictions
Normalized frequency
Softmax output predictions
approximated density - incorrect 215 classification
severity 0 - incorrect 653 classification
severity 2 - incorrect 627 classification
severity 4 - incorrect 797 classification

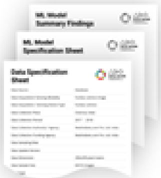


Figure 5: A snapshot of the FG-AI4H audit report cards for the retinopathy use case. Full version in Appendix B

2021

Software

Evaluation Package (EP)

Reporting Package (RP)

Evaluation Package: sub-streams

- Perturbed minds (research)
  - Plausible perturbations for robustness benchmarking in medical imaging
  - Fridays, 13.00 hrs Geneva time
  - Contact: Luis, Bruno
  - Outputs: Scientific publications, software
- Model reporting questionnaire (research)
  - A model reporting survey
  - Thursdays, 16.00 hrs Geneva time
  - Contact: Jana
  - More info today at 12.10 hrs
- Good practices (regulatory)
  - Tuesdays, 14.00 hrs Geneva time
  - Contact: Pradeep
  - Output: DEL 2.1/2.2

- “aiaudit.org”
  - Crowdsourced collection of SOTA AI quality assurance methods
  - Contact: Luis, Pat
  - Output: Web resource
- Regulatory ML4H “ICD-11”
  - An ontology to rule them all
  - Contact: Pradeep, Christian
  - Output: Software tool

Luis	luis.oala@hhi.fraunhofer.de
Bruno	bruno.sanguinetti@dotphoton.com
Jana	Jana.Fehr@hpi.de
Pradeep	pbn.tvm@gmail.com
Pat	pat.baird@philips.com
Christian	christian.johner@johner-institut.de

Ongoing

Upcoming

Scope

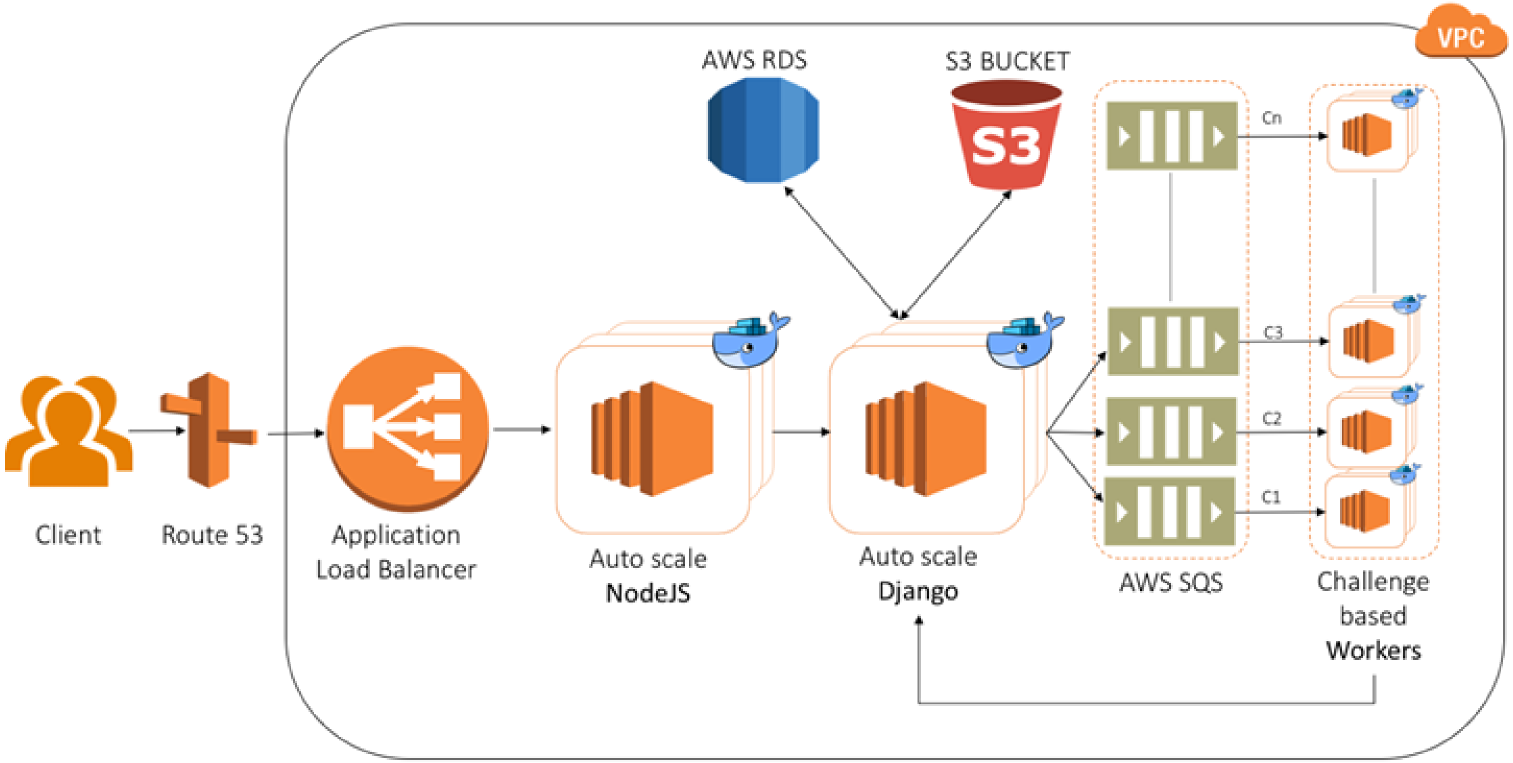
- Compare performance ML algorithms
- Customizable benchmarking
- Upload own datasets

AI for Health

EvalAi

- Open Source platform for evaluating ML algorithms
- Django-based
- Customizable benchmarking/evaluation as challenges
- Docker-based submissions by participants
- Computation in AWS

Architecture

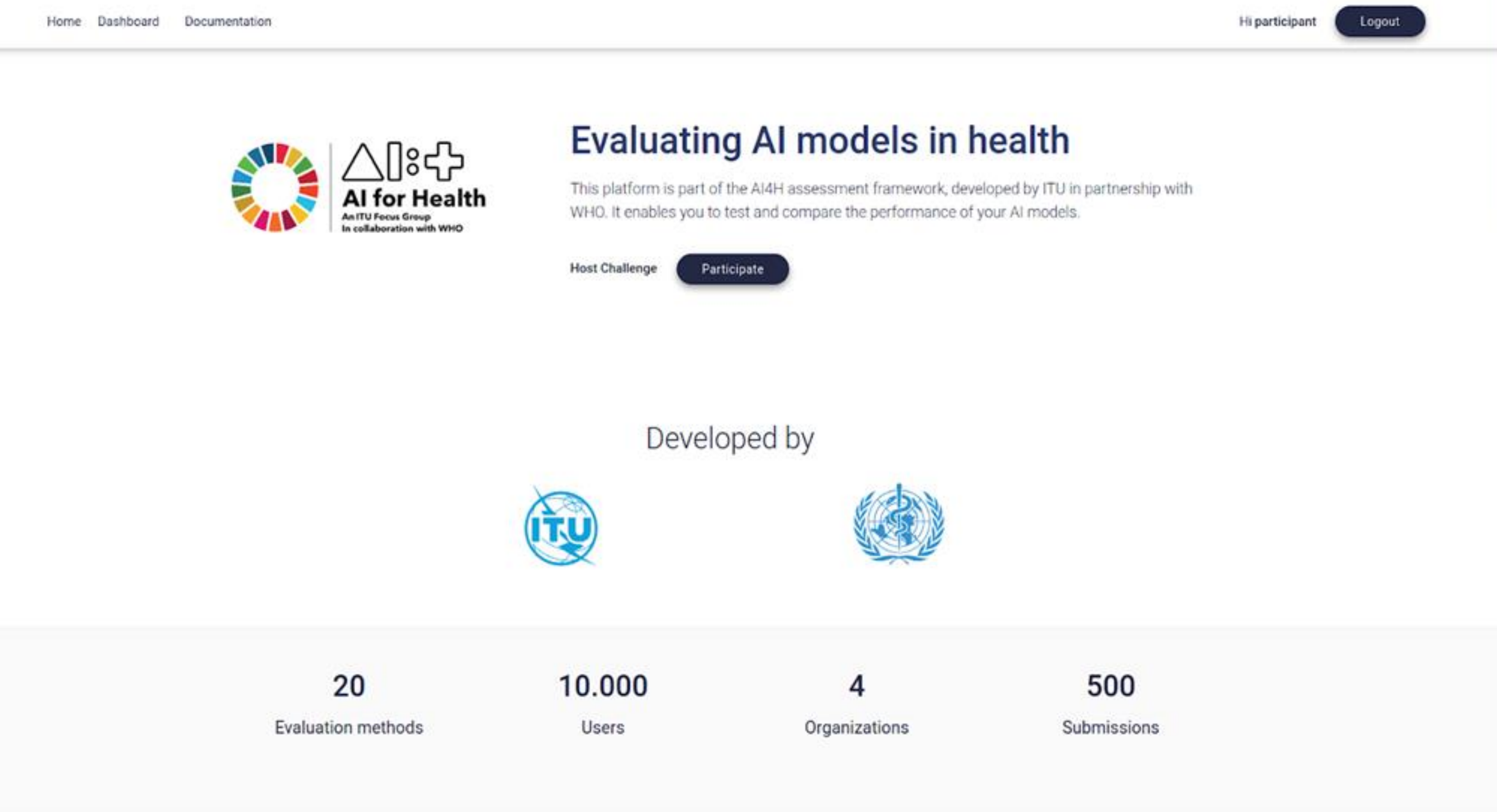


https://evalai.readthedocs.io/en/latest/architecture.htm
l

AI for Health

Current State

- Adapted Frontend
- Implemented questionnaire for qualitative evaluation
- Demo version running on AWS
- Submission of dockerized diabetic retinopathy model



Evaluating AI models in health

This platform is part of the AI4H assessment framework, developed by ITU in partnership with WHO. It enables you to test and compare the performance of your AI models.

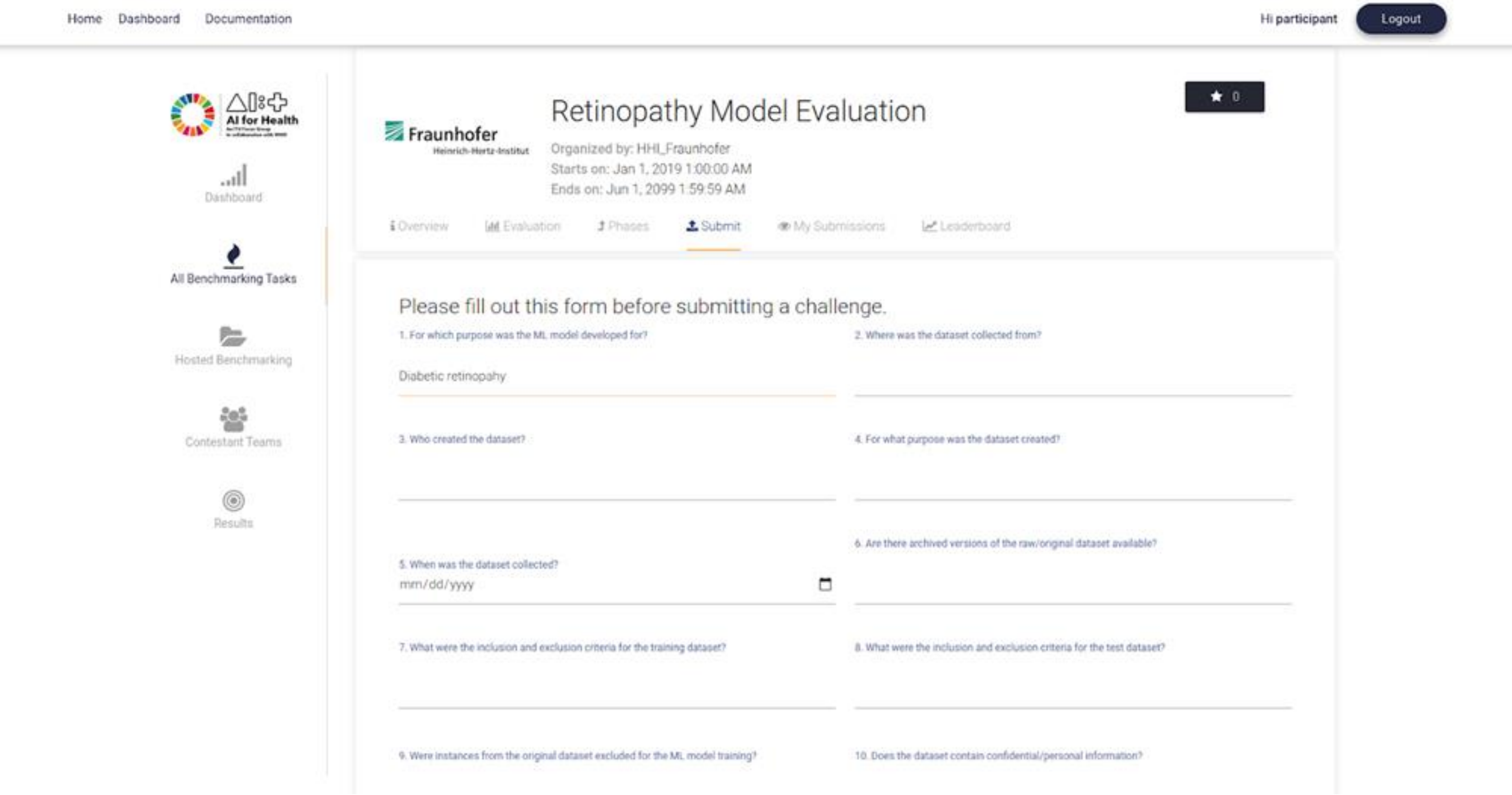
Developed by

20
Evaluation methods

10.000
Users

4
Organizations

500
Submissions



Retinopathy Model Evaluation

Organized by: HHI_Fraunhofer
Starts on: Jan 1, 2019 1:00:00 AM
Ends on: Jun 1, 2099 1:59:59 AM

Please fill out this form before submitting a challenge.

1. For which purpose was the ML model developed for?

Diabetic retinopahy

2. Where was the dataset collected from?

3. Who created the dataset?

4. For what purpose was the dataset created?

5. When was the dataset collected?

mm/dd/yyyy

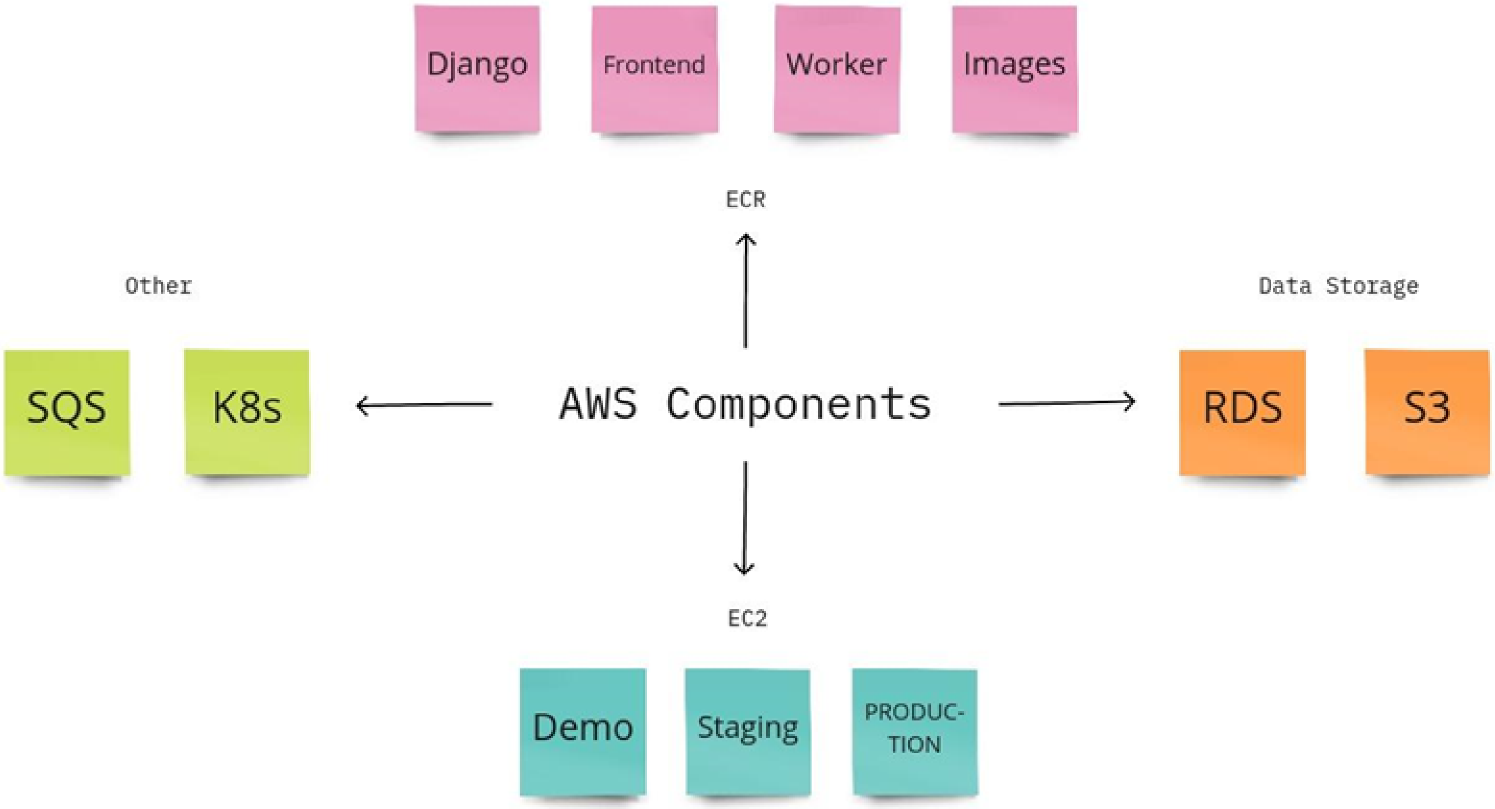
6. Are there archived versions of the raw/original dataset available?

7. What were the inclusion and exclusion criteria for the training dataset?

8. What were the inclusion and exclusion criteria for the test dataset?

9. Were instances from the original dataset excluded for the ML model training?

10. Does the dataset contain confidential/personal information?



Django Frontend Worker Images

ECR

Other

SQS K8s

AWS Components

Data Storage

RDS S3

EC2

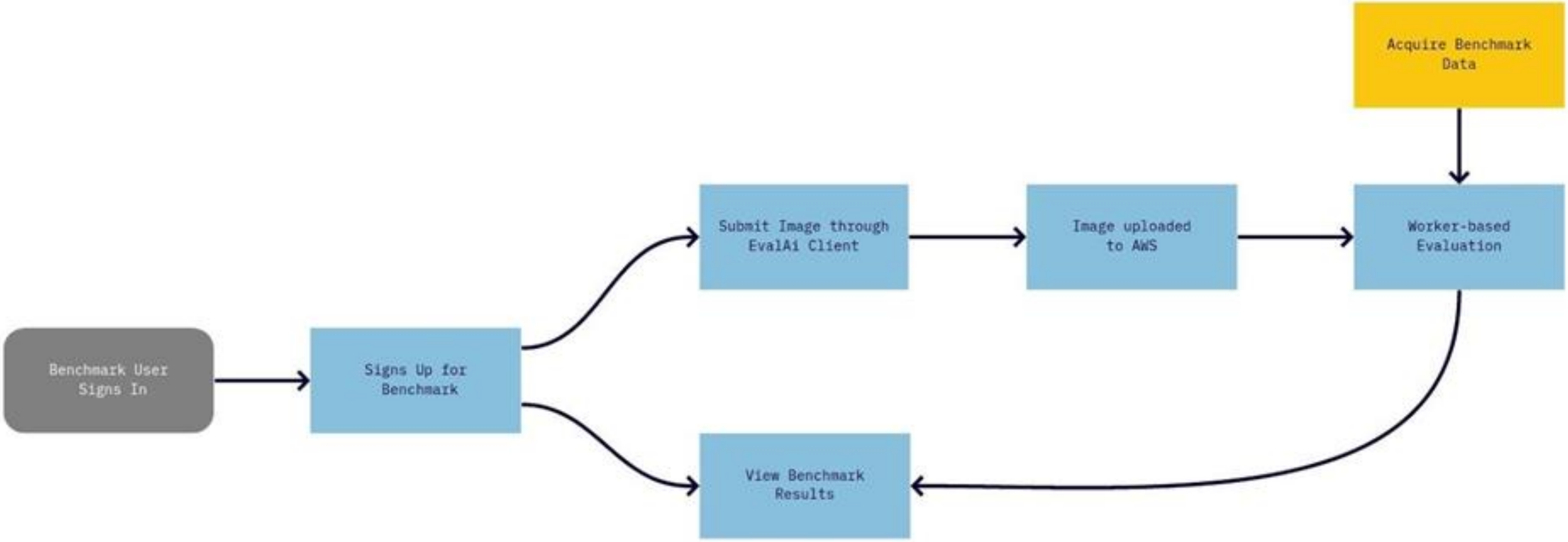
Demo Staging PRODUC-TION

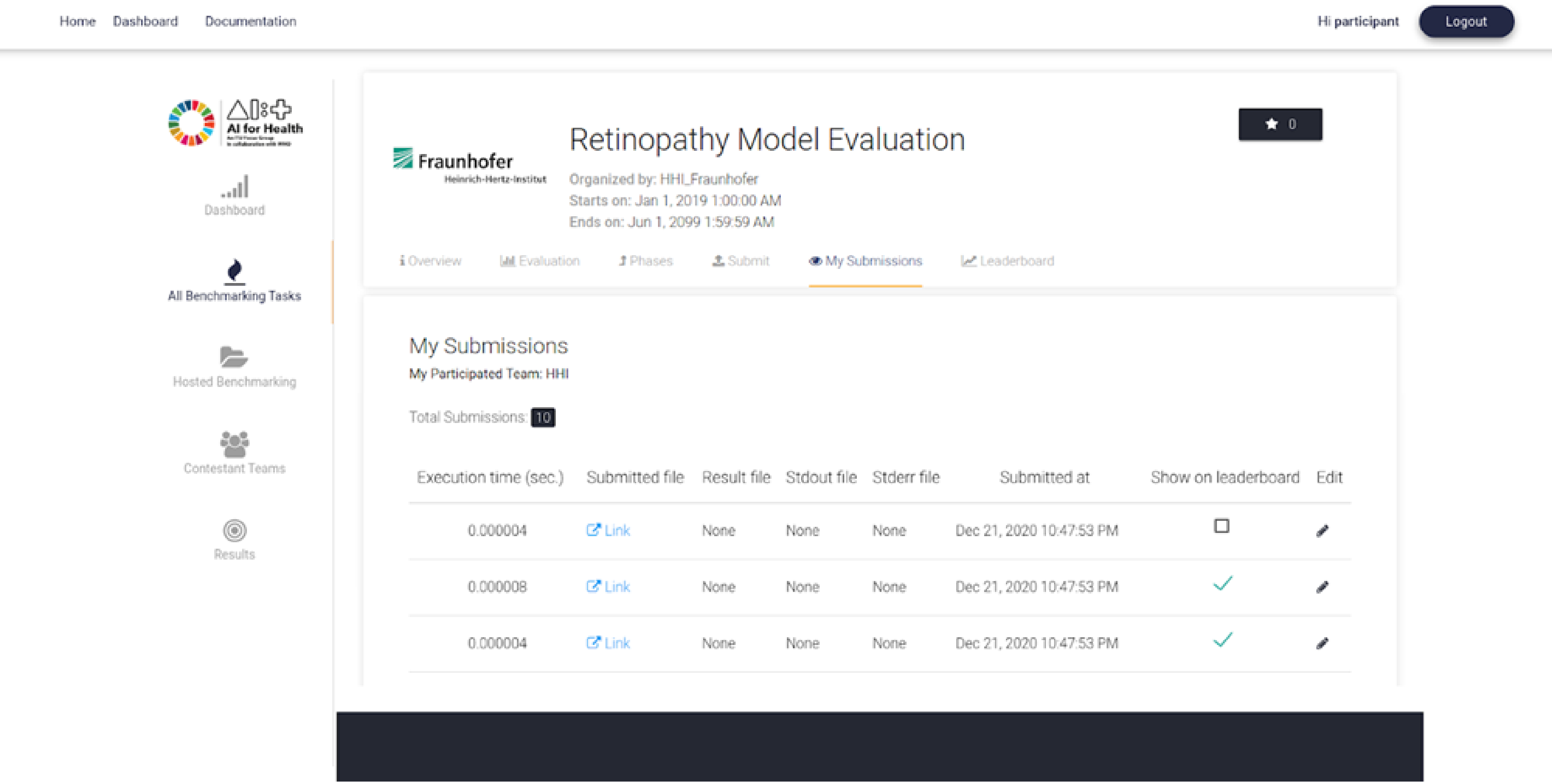
Next Steps

- Writing evaluation script for diabetic retinopathy
- Execute full benchmarking cycle
- Adapt the questionnaire
- Interested to join and work with an interesting tech stack (Django, AWS, Docker...)? Contact
  - Elora

    elora.schoerverth@hhi.fraunhofer.de
  - Alixandro

    alixandrowerneck@outlook.com


Acquire Benchmark Data
Benchmark User Signs In
Signs Up for Benchmark
Submit Image through EvalAi Client
Image uploaded to AWS
Worker-based Evaluation
View Benchmark Results



Retinopathy Model Evaluation

Organized by: HHI_Fraunhofer
Starts on: Jan 1, 2019 1:00:00 AM
Ends on: Jun 1, 2099 1:59:59 AM

My Submissions

My Participated Team: HHI

Total Submissions: 10

Execution time (sec.)	Submitted file	Result file	Stdout file	Stderr file	Submitted at	Show on leaderboard	Edit
0.000004	Link	None	None	None	Dec 21, 2020 10:47:53 PM	□	✎
0.000008	Link	None	None	None	Dec 21, 2020 10:47:53 PM	✓	✎
0.000004	Link	None	None	None	Dec 21, 2020 10:47:53 PM	✓	✎

Open Code Initiative
Reporting Package

WG-DAISAM

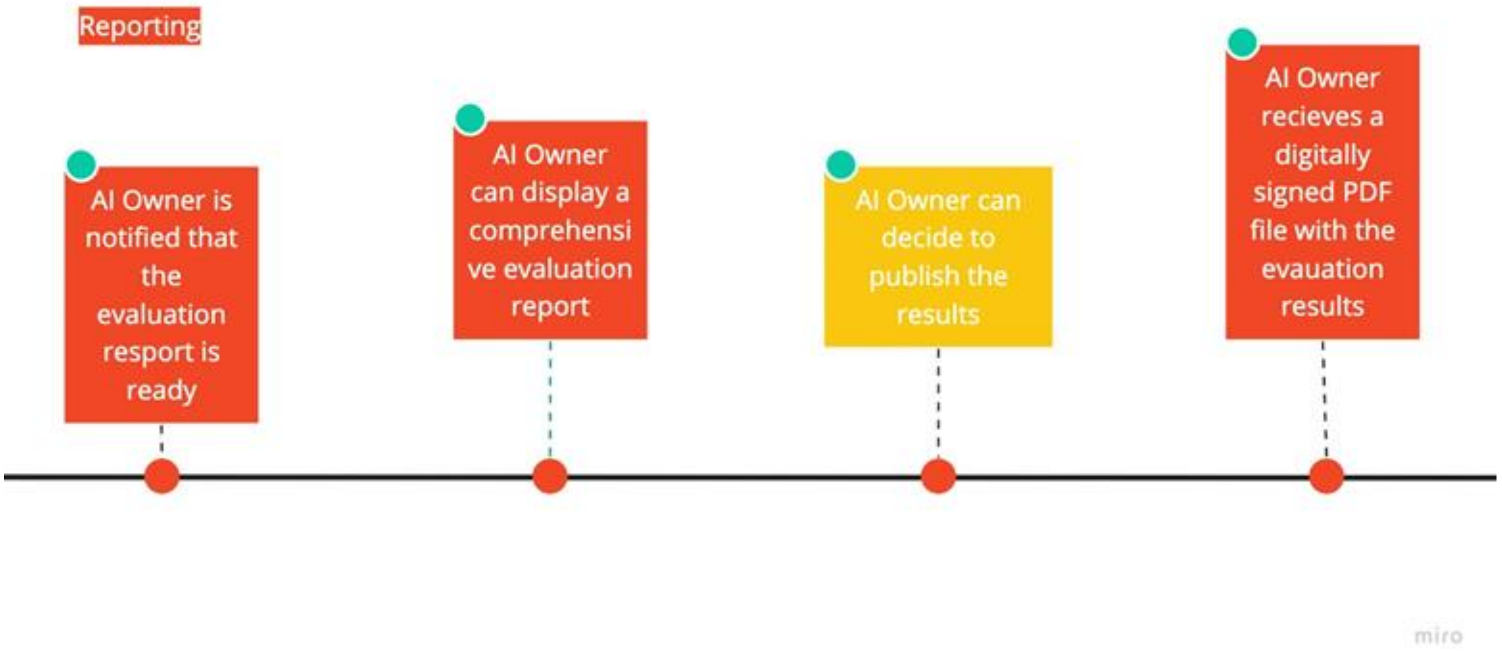
Pradeep Balachandran, Alixandro Werneck, Andrea Garcia, Danny Xie Li, Dominik Schneider, Elora Schörverth, Joachim Krois, Kamran Ali, Marc Lecoultre, Shobha Iyer, Shruti Choudhary, Steffen Vogler, Luis Oala

Scope

- Prepares and presents AI4H model evaluation results generated by the Evaluation Package (EP)

- Provides a customizable reporting interface to support ease of comparison, classification and reproducibility of different types of AI4H model evaluation results

Reporting service-workflow



Reporting

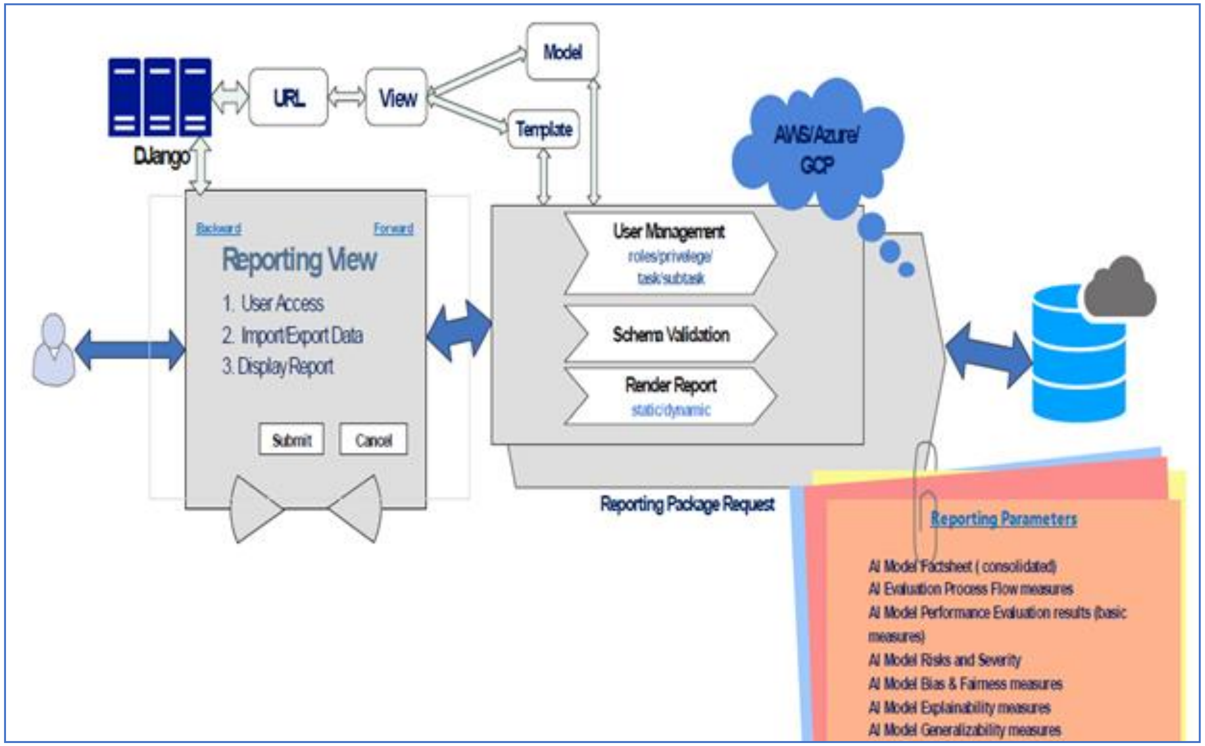
AI Owner is notified that the evaluation resport is ready

AI Owner can display a comprehensive evaluation report

AI Owner can decide to publish the results

AI Owner recieves a digitally signed PDF file with the evauation results

Architecture


DJango
URL
View
Model
Template
AWS/Azure/
GCP
Backward
Forward
Reporting View
1. User Access
2. Import/Export Data
3. Display Report
Submit
Cancel
User Management
roles/privelege/
task/subtask
Schema Validation
Render Report
static/dynamic
Reporting Package Request
Reporting Parameters
AI Model Factsheet (consolidated)
AI Evaluation Process Flow measures
AI Model Performance Evaluation results (basic measures)
AI Model Risks and Severity
AI Model Bias & Fairness measures
AI Model Explainability measures
AI Model Generalizability measures

Module	Implementation Technology
Backend module	Django-VIEW (Python)
Database module	Django-MODEL (SQLIte DB)
Frontend module	Django-TEMPLATE file (HTML, CSS)

Current Status

Prototype implementation done using Django framework

Backend

✓ Web API to fetch the Diabetic Retinopathy model evaluation result data from the Evaluation Service

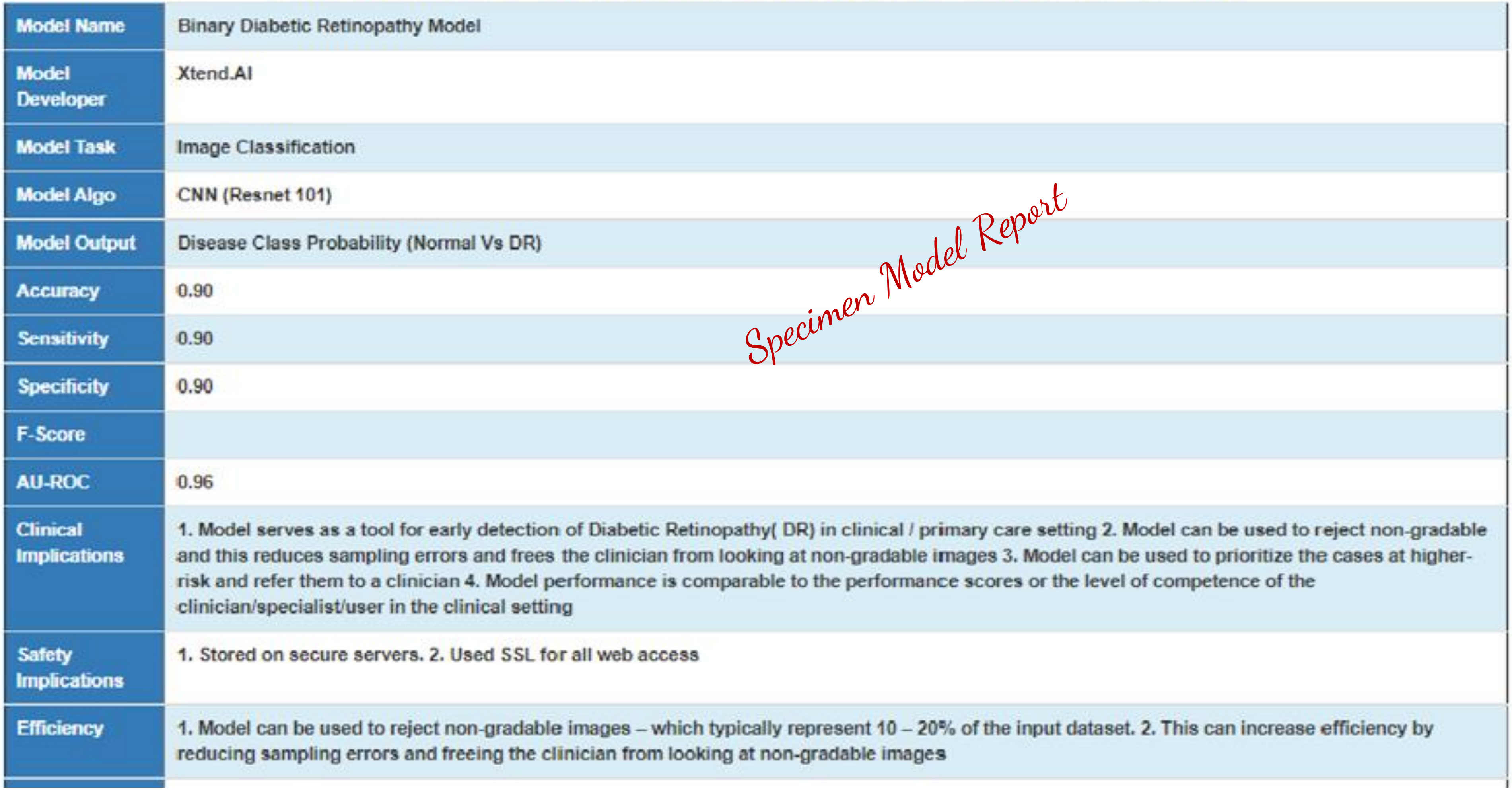
Frontend

✓ Web based GUI for model report ID submission

✓ Web based rendering template to display Diabetic Retinopathy model report

✓ Utility to enable downloading the model report in PDF format

Binary Diabetic Retinopathy Model-Summary



Model Name	Binary Diabetic Retinopathy Model
Model Developer	Xtend.AI
Model Task	Image Classification
Model Algo	CNN (Resnet 101)
Model Output	Disease Class Probability (Normal Vs DR)
Accuracy	0.90
Sensitivity	0.90
Specificity	0.90
F-Score	
AU-ROC	0.96
Clinical Implications	1. Model serves as a tool for early detection of Diabetic Retinopathy(DR) in clinical / primary care setting 2. Model can be used to reject non-gradable and this reduces sampling errors and frees the clinician from looking at non-gradable images 3. Model can be used to prioritize the cases at higher-risk and refer them to a clinician 4. Model performance is comparable to the performance scores or the level of competence of the clinician/specialist/user in the clinical setting
Safety Implications	1. Stored on secure servers. 2. Used SSL for all web access
Efficiency	1. Model can be used to reject non-gradable images – which typically represent 10 – 20% of the input dataset. 2. This can increase efficiency by reducing sampling errors and freeing the clinician from looking at non-gradable images

Prototype Demo →

Next Steps

Module integration with Evaluation Package

- Report data schema validation with Evaluation Service

- Data-READ Web API validation with Evaluation Service
- Interested to join? Contact
    - Pradeep pbn.tvm@gmail.com
    - Alixandro alixandrowerneck@outlook.com