OpenEBench: advancing AI benchmarking

Codó L.[1,2], Redondo A.[1,2], Fernández J.M.[1,2], The OpenEBench Team ↔, Gelpí J.L.[1,2,3], Capella-Gutierrez S.[1,2]

[1] Spanish National Bioinformatics Institute (INB/ELIXIR-ES). [2] Barcelona Supercomputing Center (BSC). [3] Dept Biochemistry and Molecular Biomedicine, University of Barcelona (UB)

ELIXIR All Hands , 10-12 June 2024, Uppsala, Sweden

Establishing objective and transparent **benchmarking** methodologies is vital to ensure that **AI datasets, algorithms, and technologies** meet the standards for effectiveness and reliability demanded by the field.

OpenEBench is being extended to evaluate the **inference performance** of trained **AI models**. The diagram shows the flow being developed in the context of the **EuCanImage** community.

Preparatory phase

The community organizes the benchmarking event and pre-defines a set of assessment metrics

COMMUNITY · OEB TEAM

1 Design of the benchmarking strategy

2 Definition of assessment metrics

$$\text{Dice} = \frac{2\,X\,Area\,of\,overlap}{Total\,area}$$

TOOLS · METRICS · DATASETS · CHALLENGES · EVENTS · COMMUNITIES

3 Adoption of OpenEBench standardized benchmarking data model

4 Data Collection

5 Data Splitting Strategy

6 Benchmarking Datasets Preparation

Training dataset · Evaluation dataset · Gold Standard · Testing dataset

7 Technical Implementation of the Assessment Metrics

CSV participant predictions · Gold Standard · Other participant's assessments · Validation · Metrics computation · Consolidation · nextflow Standardized Benchmarking Workflow · git · JSON Assessment dataset · JSON Aggregation dataset

8 Long-term Archiving of Benchmarking Datasets

PERMURL

- DE-IDENTIFIED BIOIMAGES *(annotated and not annotated)* — EUROBIOIMAGING
- PHENOTYPIC DATA *(clinical, epidemiological...)* — EUROPEAN GENOME-PHENOME ARCHIVE
- ...

9 Participant Registration

10 Training Dataset Access

Training phase

The researcher participating at the benchmarking event trains its AI model

RESEARCHER · COMMUNITY

11 AI Model Training & Optimization

TRAINED MODEL *(versioned)*

```
├── LICENSE
├── README.md
├── main.nf
├── nextflow.config
├── model-container
│   ├── DockerFile
│   ├── src
│   │   └── model
│   │       └── inference.py
│   └── model
│       └── model.pkl
```

GitLab · scikit-learn · TensorFlow · PyTorch

Integration Phase

Containerized inference system for the trained AI model is loaded to OEB

RESEARCHER · COMMUNITY · Virtual Research Environment

12 On-hold datasets transference to OpenEBench premises

Gold Standard · Testing dataset · Verification subset

13 AI Model Inference Tool Containerization

14 Online Private Workspace Sign-up

Virtual Research Environment · AI Model Inference

15 AI Model Inference Container Registration

Back to 13

16 Inference Run on New AI Model for Technical Verification

Verification subset · AI Model Inference · Workflow Run RO-Crate · CSV valid predictions · nextflow · WfExS · Reproducible Inference Run

Testing Phase

Generate predictions on the testing dataset

RESEARCHER · Virtual Research Environment

17 Inference Run on Undisclosed Testing Dataset

Testing dataset · AI Model Inference · Workflow Run RO-Crate · CSV participant predictions · nextflow · WfExS · Reproducible Inference Run

Evaluation Phase

Predictions are assessed by the pre-defined metrics, implemented as a container-based workflow.

RESEARCHER · Virtual Research Environment

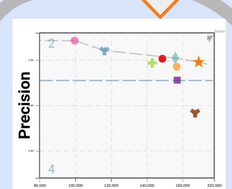
18 Assessment Metrics computation

Gold Standard · Other participant's assessments · Validation · Metrics computation · Consolidation · nextflow Standardized Benchmarking Workflow · JSON Aggregation dataset · JSON Assessment dataset

Back to 11 (limited No. of rounds)

Metrics Private Visual Analysis

19



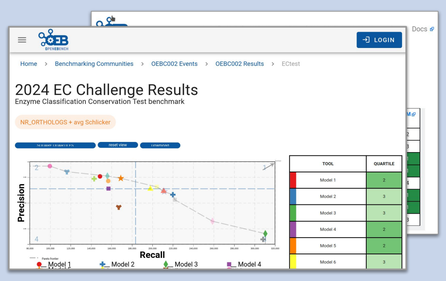
Publication Phase

Assessment metrics and dataset references are made publics

RESEARCHER · COMMUNITY · OEB TEAM

B2SHARE · EUDAT · doi

Data Publicated @ B2SHARE 23



Data Publicated @ OpenEBench Portal 22

Gold Standard · Testing dataset · AI Model Inference · JSON Aggregation dataset · JSON Assessment dataset · CSV participant predictions · Provenance · Metrics · PUBLICATION ASSETS

20 Publication Request

21 Publication Revision & Acceptance

COMMUNITY

EUCAN IMAGE

This effort is being driven by **EuCanImage**, a pan-European infrastructure dedicated to applying AI to cancer imaging. The AI algorithms for clinical diagnosis developed during this project will be validated through 8 different benchmarking events, using data to specific clinical use-cases. The tasks assessed will include **image segmentation, classification, and detection** (localization) on magnetic resonance imaging (MRI), computed tomography (CT) scans, and mammography (MG) images, focusing on various organs such as breast, liver, and colon.

REFERENCES

OpenEBench · Workflow RO-Crate · WfExS · Nextflow · euCanImage

Contact

Laia Codó Tarraubella
Barcelona Supercomputing Center (BSC)
laia.codo@bsc.es

↔ OpenEBench Team: https://openebench.bsc.es/team

Barcelona Supercomputing Center — Centro Nacional de Supercomputación
INB Spanish National Bioinformatics Institute · ELIXIR SPAIN

EUCAN IMAGE
EuCanImage has received funding from the European Union's Horizon 2020 programme under grant agreement No 952103

eosc ENTRUST
European Network of Trusted Research Environments
EOSC-ENTRUST project receives funding from the European Union's Horizon Europe Programme under grant agreement number 101131056.

Funded by the European Union

ELIXIR ALL HANDS 2024