Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

Call identifier: DIGITAL-2022-CLOUD-AI-02

D6.2. EUCAIM Benchmarking platform

Responsible partner: BSC

Author(s): Carles Hernandez-Ferrer (BSC), Pedro Miguel Martínez-Gironés (HULAFE), José Guilherme de Almeida (CF), Josep Lluís Gelpí (UB), Salvador Capella (BSC)

Contributor(s): Carina Soler-Pons (HULAFE), María Beser-Robles (HULAFE), Miriam Groeneveld (radboudumc).

Reviewer(s): Michał Kosno (GUMED), Yiannis Roussakis (GOC)

Date of delivery: March 26, 2025

Version: 1.0

Table of contents

1. Introduction

1.1 Document Purpose

This document outlines the technical capabilities of the benchmarking platform designed as part of the European Cancer Imaging Initiative (EUCAIM). It provides a detailed description of the platform's (future) features and functionalities, demonstrated through a proof-of-concept (PoC) on scientific benchmarking. The PoC serves as an initial step toward validating the platform's ability to support diverse benchmarking events across various aspects of EUCAIM.

The benchmarking platform will serve as a comprehensive tool to host multiple benchmarking events focusing on two primary domains: technical benchmarking and scientific benchmarking. These events will assess and compare AI models, preprocessing tools, analytical tools, and datasets to ensure they meet the initiative's high standards for performance and reproducibility.

By enabling consistent evaluation across different areas, the platform will play a critical role in advancing the goals of EUCAIM and fostering collaboration within the cancer imaging research community.

1.2 Document Scope

This document defines the benchmarking strategy within the EUCAIM initiative, focusing on its role in evaluating AI models, data preprocessing tools, analytical software, and datasets. It provides a comprehensive overview of the benchmarking platform's - being OpenEBench[1] and Grand Challenge[2] - as well as methodologies. The document outlines the key benchmarking domains—scientific and technical—detailing their respective goals, challenges, and expected outcomes within the EUCAIM initiative. Finally, it describes the Proof of Concept (PoC) conducted to validate the platform's capabilities and the integration of benchmarking results into OpenEBench and Grand Challenge.

2. Benchmarking

Benchmarking in the EUCAIM project is a cornerstone for ensuring that the platform's software tools, models, and infrastructure meet the highest standards of **performance**, **scalability**, and **usability**. This section provides a comprehensive overview of the **scientific** and **technical benchmarking** processes that drive the integration and optimization of artificial intelligence (AI) models, data preprocessing tools, analytical software, and platform architecture components.

2.1 Scientific benchmarking

Scientific benchmarking focuses on the evaluation of AI models, data analysis tools, and datasets to ensure they are valid, reliable, and applicable to clinical and research contexts. The key elements of this effort include:

- **AI Models for Inference**. Benchmarked against established reference datasets to evaluate metrics such as accuracy, sensitivity, specificity, and robustness. Particular emphasis is placed on assessing model performance across diverse clinical scenarios and data modalities to ensure generalizability and reliability.

[1] https://openebench.bsc.es/
[2] https://grand-challenge.org/

- **Data Preprocessing Software**. Evaluated for their ability to handle diverse data modalities (e.g., medical imaging, demographic data) and perform key preprocessing tasks such as data cleaning, harmonization, and transformation. Benchmarking ensures compatibility with the federated architecture and scalability across different data sources.
- **Analytical Software**. Tested against reference datasets to verify their capability to generate meaningful insights for varied clinical and research use cases. This may includes assessing the software's efficiency in handling high-dimensional data and its potential to support exploratory and hypothesis-driven analysis
- **Datasets**. Benchmarking efforts include identifying hidden biases within datasets, assessing their representativeness, and comparing them using standardized metrics. This process helps uncover potential internalized structures that could influence model training and outputs, thereby promoting fairness and interpretability in AI-driven analyses.

2.2 Technical benchmarking

Technical benchmarking ensures the seamless integration, robustness, and scalability of the platform's components within real-world operational environments. Core activities include:

- **Integration and Compatibility Testing (Central Services)**. Verifies that all software components function cohesively within the central EUCAIM infrastructure. This includes testing interoperability between central services such as data registries, AI model repositories, and user interfaces.
- **Integration and Compatibility Testing (Federated Nodes)**. Ensures that preprocessing tools, analytical frameworks, and other software components operate harmoniously within the federated architecture. This involves evaluating performance and compatibility at the local premises of the federated nodes, ensuring consistent operation across decentralized environments.
- **AI Model Deployment**. Benchmarked for efficient inference and adaptability across various hardware configurations, from cloud environments to edge devices. Testing includes evaluating latency, throughput, and resource utilization to identify optimal deployment scenarios and ensure responsiveness in clinical workflows.
- **Platform Scalability and Resilience.** The EUCAIM platform undergoes stress testing to assess its scalability under varying loads, including the simultaneous processing of large datasets across multiple federated nodes. Resilience benchmarks ensure that the platform maintains functionality and recovers efficiently in the event of disruptions.

3. Plan and infrastructure

The EUCAIM proof-of-concept (PoC) focuses on benchmarking two AI models according to their task, to evaluate their performance, interoperability, and suitability for integration into the platform. Additionally, the PoC aimed to explore the dataset aspects of EUCAIM to identify potential data biases—such as sex/gender biases, race/ethnic biases, and other biases of medical and biomedical interest relevant to AI training. However, due to the lack of accessibility to EUCAIM's datasets, the PoC concentrated primarily on evaluating AI models.

At the platform level, the PoC also tested the publication of benchmarking results in both OpenEBench and Grand Challenge, ensuring compatibility and dissemination within these established frameworks. This approach not only validates the technical capabilities of the platform but also paves the way for future integration and scalability in benchmarking activities.

3.1 Strategy

The PoC adopts a twofold approach to benchmarking AI models:

3.1.1 Benchmarking two segmentation models performing the same task on the same dataset

- **Objective:** Evaluate and compare 3 segmentation models performing the same task on the same dataset.
- **Analysis:** Analyze differences in segmentation accuracy and usability between 2 specialized models and a general-purpose one.
- **Models:**
    - Full Automatic segmentation of Glioblastoma Multiforme (HULAFE)[3]. A model trained to segment the entire tumor in a single process.
    - Subregional Automatic segmentation of Glioblastoma Multiforme (HULAFE)[1]. The same architecture as the full segmentation model, but trained separately for different tumor subregions (enhanced tumor, necrosis, and peritumoral edema). The final segmentation is obtained by merging the partial segmentations.
    - Minimal interactive tumor segmentation (Erasmus MC)[4]. A model that generates a full tumor segmentation based on six user-provided points as input. It does not require a specific training process for this task; however, it requires human intervention from expert clinicians.
- **Dataset:** The PerProGlio[3] dataset is a multicenter collection of glioblastoma images and clinical data gathered across Europe from 2019 to 2023. The subset comprising cases from the HULAFE data holder is the one used for this benchmarking. The three classification models presented in this benchmarking were tested using this dataset. Both architectures, HULAFE's and ERASMUS MC's, were trained with other cases. The HULAFE model was trained and validated using the *Brain Tumor Segmentation Challenge 2021 (BraTS2021)*[5] image repository. The ERASMUS model was trained and validated using the public, retrospective, and multicenter *WORC* database. More information is available in the related bibliography.

[3] Beser-Robles, M., Castellá-Malonda, J., Martínez-Gironés, P. M., Galiana-Bordera, A., Ferrer-Lozano, J., Ribas-Despuig, G., Teruel-Coll, R., Cerdá-Alberich, L., & Martí-Bonmatí, L. (2024). *Deep learning automatic semantic segmentation of glioblastoma multiforme regions on multimodal magnetic resonance images*. International journal of computer assisted radiology and surgery, 19(9), 1743–1751. https://doi.org/10.1007/s11548-024-03205-z

[4] Spaanderman, D. J., Starmans, M. P. A., van Erp, G. C. M., Hanff, D. F., Sluijter, J. H., Schut, A.-R. W., van Leenders, G. J. L. H., Verhoef, C., Grunhagen, D. J., Niessen, W. J., Visser, J. J., & Klein, S. (2024). *Minimally interactive segmentation of soft-tissue tumors on CT and MRI using deep learning*. arXiv. https://arxiv.org/abs/2402.07746

[5] U.Baid, et al., The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, arXiv:2107.02314, 2021.

3.1.2 Benchmarking two classification models performing the same task on different datasets.

- **Objective:** Compare the performance of two classification models performing the same task but trained on different datasets.
- **Analysis:** Evaluate model performance using standard survival metrics across diverse training datasets.
- **Models:**
    - Neuroblastoma survival model (HULAFE).
    - Neuroblastoma survival model (QUIBIM).
- **Dataset**: The PRIMAGE[6] dataset is a multicentric collection of neuroblastoma images and clinical data gathered across Europe from 2018 to 2023. The two classification models presented for this benchmarking were trained and validated using different subsets of the complete dataset. QUIBIM's model was trained on all available cases and tested on external cases, while HULAFE's model was trained exclusively on non-HULAFE cases and tested on HULAFE cases.

3.2 Results

3.2.1 Pipeline evaluation

The benchmarking process encountered challenges and required adaptive solutions to advance the project's goals. Below are the key findings and resolutions:

Lack of Ground Truth Data for Benchmarking

During the initial design phase of the PoC, it became evident that no ground truth dataset was available for the benchmarking tasks.

- **Solution.** Through collaboration with T7.3 partners, HULAFE stepped forward to provide the PerProGlio dataset. Since the dataset resides at HULAFE's premises, they performed the benchmarking manually by applying the models to their dataset, calculating the metrics, and delivering the results (see ***Figure 1***).
- **Outcome.** This adaptive approach allowed the segmentation model benchmarking to proceed despite the initial data limitations.

[6] Martí-Bonmatí L, Alberich-Bayarri Á, Ladenstein R, Blanquer I, Segrelles JD, Cerdá-Alberich L, Gkontra P, Hero B, García-Aznar JM, Keim D, Jentner W, Seymour K, Jiménez-Pastor A, González-Valverde I, Martínez de Las Heras B, Essiaf S, Walker D, Rochette M, Bubak M, Mestres J, Viceconti M, Martí-Besa G, Cañete A, Richmond P, Wertheim KY, Gubala T, Kasztelnik M, Meizner J, Nowakowski P, Gilpérez S, Suárez A, Aznar M, Restante G, Neri E. PRIMAGE project: predictive in silico multiscale analytics to support childhood cancer personalised evaluation empowered by imaging biomarkers. Eur Radiol Exp. 2020 Apr 3;4(1):22. doi: 10.1186/s41747-020-00150-9. PMID: 32246291; PMCID: PMC7125275.

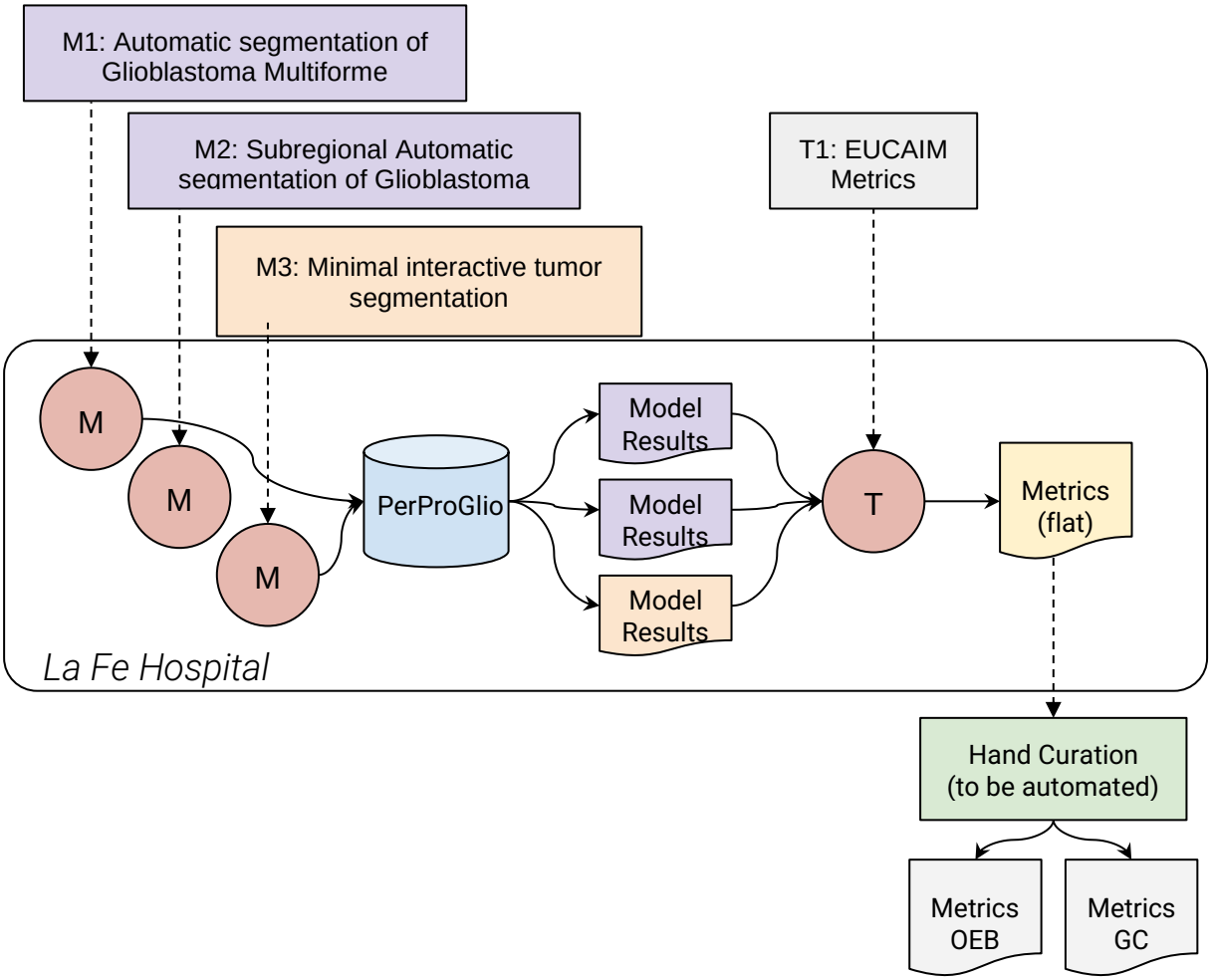

Figure 1. General description of the benchmarking pipeline used in the PoC. *The two evaluated models (M1 and M2) were deployed with the HULAFE premises and applied on the PerProGlio datasets. The results of the models were given to the "EUCAIM Metrics", a python package deployed at the HULAFE premises, to extract the desired metrics. The metrics were then distributed across the contributors to hand curate and tailor them for Grand Challenge and OpenEBench.*

Inability to Execute the Second Benchmarking Event

The second benchmarking event—comparing two classification models trained on different datasets—could not be conducted. The models were trained using data from the "combined dataset for new validation," creating a circular dependency that would render metrics unreliable and the report inaccurate.

- **Solution.** This limitation highlighted the need for more stringent separation between training and validation datasets in future benchmarking designs. Additionally, it emphasized the importance of dataset accessibility and preparation early in the project lifecycle.
- **Outcome.** Although the classification benchmarking task was not completed, lessons learned from this challenge inform improved methodologies for subsequent benchmarking efforts.

Data structure and metrics extraction

The benchmarking design for the task of **"Benchmarking two segmentation models performing the same task on the same dataset"** involves a "blind" approach to metrics extraction (see ***Figure 1***). This ensures that the evaluation process is unbiased and standardized. The following outlines the agreed-upon structure and methodology for result organization and metric computation:

Input Format

The output of the segmentation models consists of segmentation masks in NIFTI format:

- Primary format: `.nii` (optionally `.nii.gz`)
- Alternative format: DICOM SEG (conversion possible if needed).

Output Format (EUCAIM-Interoperable)

A folder per case, with subfolders for ground truth and predictions:

- Case Folder:
  - GT Subfolder: Contains ground truth files.
  - Output Subfolder: Contains prediction files.

3.2.2 Metrics Extraction

To evaluate the segmentation results, the EUCAIM Metrics[7]' Docker container was developed. This container includes a suite of scripts specifically designed for the benchmarking challenge. The key metrics calculated include:

- **Dice Score.** It is a coefficient that measures the similarity between two sets. The coefficient ranges from 0 to 1, where 1 indicates that the two sets are identical, and 0 indicates that the two sets have no overlap.

```
Dice coefficient = 2 * |A ∩ B| / (|A| + |B|)

Where |A| represents the number of elements in set A, and |B| represents the number of elements in set B. |A ∩ B| represents the number of elements that are present in both sets.
```

- **Intersection Over Union (IoU).** Provides a ratio of the intersection to the union of the ground truth and prediction masks. The coefficient ranges from 0 to 1, where 1 indicates that the two areas fully overlap, and 0 indicates no overlap between the two areas.

```
Intersection Over Union = Area of Intersection / Area of Union
```

Considered metrics but that were discarded because of issues in calculation efficiency (but that will be include in future benchmarking events):

[7] EUCAIM Metrics' GitHub: https://github.com/josegcpa/eucaim-metrics

- **Hausdorff Distance.** It is a distance that measures how far two subsets of a metric space are from each other by identifying the greatest distance one must travel from a point in one set to reach the closest point in the other set.

$$d_H(A,B) = \max\{ \sup_{a\in A} \inf_{b\in B} d(a,b), \sup_{b\in B} \inf_{a\in A} d(b,a) \}$$

Here, d(a,b) denotes the distance between points a and b. The first term finds the point in A that is farthest from any point in B, and vice versa for the second term. The Hausdorff Distance is the larger of these two values, capturing the worst-case mismatch between the sets.

- **Normalized Surface Distance (NSD).** It is a distance that quantifies the average error between the predicted and ground truth surfaces, normalized by a tolerance threshold.

$$NSD = 1 / |S_{gt}| \ SUM_{x\in S_{gt}}(F(d(x, S_{pred}) < \tau)$$

Here the S_{gt} is the surface defined by the ground truth while S_{pred} is the surface of the prediction. Therefore the $d(x, S_{pred})$ is the shortest distance from a point x to the predicted surface. F is the indicator function (1 for true and 0 otherwise). Finally, τ is the predefined distance tolerance

Metrics publication and benchmarking visualization

Results of the benchmarking exercises were published both in OpenEBench[8] and Grand Challenge[9]. ***Table 1*** summarizes the results of the benchmarking event developed during this Proof of Concept in both OpenEBench and Grand Challenge fashion.

Table 1. Benchmarking results. *Summary of the classification of each model according to the benchmarking platform. OpenEBench summarises the results assigning each participant to a quartile (combining all metrics) while Gran Challenge uses the primary metric of the challenge (Dice Score) as main score to rank the participants.*

	OpenEBench		**Grand Challenge**	
	Dice Score	**IoU**	**Dice Score**	**IoU**
ASFL-HULAFE	Q1	Q1	0.69 ± 0.23	0.56 ± 0.24
ASFL-HULAFE-Submodule	Q3	Q3	0.65 ± 0.21	0.52 ± 0.21
MITS-ErasmusMC	Q4	Q4	0.58 ± 0.17	0.42 ± 0.16

[8]OpenEBench: https://dev-openebench.bsc.es/scientific/OEBC014/?event=OEBE0140000000
[9]Grand Challenge: https://eucaim-spoc.grand-challenge.org

Moreover, ***Figure 2***, extracted from OpenEBench, allows a visual exploration of the comparing metrics of **Dice Score** (y-axis) and **Intersection Over Union** (x-axis) of the three participants in the benchmarking event.

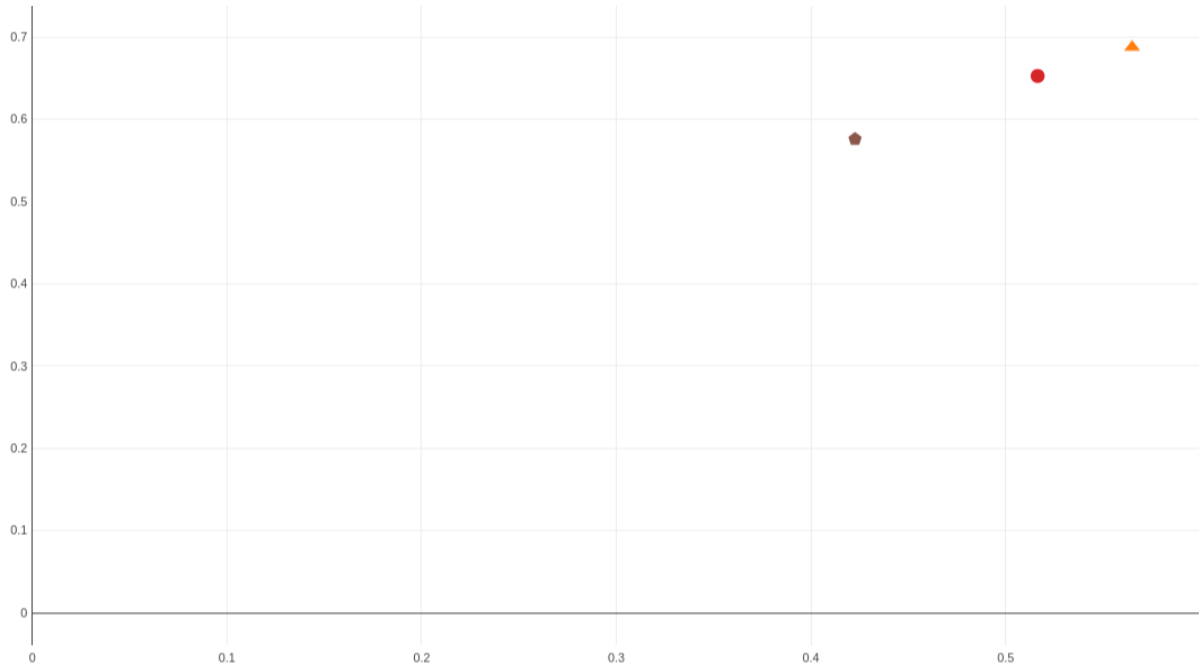

Figure 2. Scatter plot of Dice Score (DS) and Intersection Over Union (IoU) in OpenEBench. *OpenEBench allows for multiple visualizations of the metrics scored in a challenge. This figure eases the visual inspection of the relation between DS (y-avis) and IoU (x-axis) of the three participants in the challenge.*

4. Limitations

The benchmarking efforts undertaken in the EUCAIM project have been critical for advancing the integration and evaluation of AI tools. However, several limitations in the proposed and implemented approach have been identified. These challenges highlight areas requiring further attention to ensure the benchmarking process is both rigorous and scalable.

4.1 Lack of Ground Truth Datasets

The absence of predefined ground truth datasets for benchmarking tasks has been a major obstacle. This limitation compromises the ability to ensure consistent and unbiased evaluations of AI models..

- **Impact.** The reliance on localized datasets, such as the PerProGlio dataset provided by HULAFE, is a temporary solution for a specific benchmarking challenge and does not address the broader need for universally accepted benchmarking datasets.

4.2 Limited Access to Datasets

Access to datasets required for benchmarking has been a persistent challenge. A systematic task for dataset accessibility is currently missing from the project scope.

- **Impact.** Delayed dataset access affects the timeline for benchmarking events and limits the project's ability to validate tools comprehensively.
- **Proposed Solution.** Once the federated processing infrastructure is operational, WP6 is prepared to lead a comprehensive benchmarking event using datasets provided by EUCAIM's partners and metrics from T7.3.

4.3 Low Involvement of Software Owners

The limited engagement of software owners in the benchmarking process has been a significant bottleneck. Preparing software to handle the specific data used for benchmarking is a mandatory step that requires active collaboration.

5. Future work

The Proof of Concept (PoC) has demonstrated that it is feasible to benchmark software within the EUCAIM infrastructure, setting a foundation for future developments. However, to fully realize the potential of benchmarking within the project, future efforts should focus on automation and continuous benchmarking initiatives.

5.1 Automation of Benchmarking Processes

To streamline the benchmarking workflow, automation should become a central focus. Both Grand Challenge and openEBench offer valuable services that can inform the development of automated benchmarking pipelines. By leveraging these platforms, EUCAIM can integrate tools and datasets into a cohesive, scalable benchmarking framework, reducing manual efforts and improving efficiency.

5.2 Continuous Benchmarking Events

To ensure the long-term reliability and quality of software and datasets within the EUCAIM ecosystem, we propose the establishment of continuous benchmarking events:

5.2.1 Continuous benchmarking event on software

- Any software entering the EUCAIM infrastructure should undergo proper benchmarking as part of the "Software Onboarding Strategy."
- Benchmarking should evaluate the software based on the tasks it aims to support, ensuring alignment with project goals.
- OpenEBench provides a mechanism for "continuous benchmarking events," which can serve as a model for this initiative. Preparing software to participate in such events should be added as a mandatory final step in the onboarding process.

5.2.2 Continuous benchmarking event on datasets

- All datasets within EUCAIM should be assessed for quality and bias using a comprehensive set of metrics.
- These metrics should include tools to detect obvious and hidden biases, such as sex/gender, ethnicity, race, and data completeness.
- A preliminary framework for dataset benchmarking was outlined during the T7.3 Brainstorming Workshop, which proposed metrics for sex/gender bias identification and emphasized the need for additional metrics on ethnicity, race, and completeness.

By automating benchmarking processes and instituting continuous benchmarking events, EUCAIM can ensure its infrastructure remains robust, scalable, and capable of supporting high-quality AI-driven solutions. These initiatives will help maintain the integrity of the platform, foster trust among users, and enhance the adoption of its tools and datasets across clinical and research domains.

6. Conclusions

The Proof of Concept (PoC) conducted within EUCAIM has demonstrated the feasibility of benchmarking AI models even without the appropriate infrastructure in the project. It has provided valuable insights into the platform's capabilities, identifying key technical challenges and opportunities for improvement. By integrating benchmarking results into OpenEBench and Grand Challenge, the initiative has taken a critical step toward standardizing benchmarking practices and ensuring the visibility of results within established frameworks.

However, **Task 7.3** now raises a significant concern: without a set of ground truth datasets for benchmarking, the entire benchmarking infrastructure, and the very purpose of **Task 7.3**, will be rendered ineffective. Benchmarking efforts require universally accepted, well-curated datasets to ensure consistent, reliable, and reproducible evaluations. The current reliance on locally available datasets, while a temporary solution, does not address this fundamental need. Without structured efforts to establish and maintain gold-standard datasets, the benchmarking framework will lack the necessary foundation for meaningful assessments, potentially undermining the entire EUCAIM initiative's efforts in benchmarking AI-driven tools.

Additionally, while OpenEBench and Grand Challenge have proven to be valuable platforms for publishing and disseminating benchmarking results, their integration within EUCAIM should be further optimized. Future efforts should focus on enhancing automation and ensuring that benchmarking pipelines align seamlessly with these platforms. Continuous benchmarking events leveraging OEB's methodologies could offer a pathway toward sustainability, enabling the infrastructure to support long-term AI evaluation and dataset validation. Strengthening these integrations and addressing dataset accessibility will be critical for the future success of benchmarking efforts within EUCAIM.