AI-BOOST

Delivering the next level of European **AI Open competitions**

GENERATIVE AI FOR ENHANCEMENT OF CLINICAL DATASETS

CHALLENGE DESCRIPTION

1. ANNEX 2: CHALLENGE DESCRIPTION

1 CHALLENGE DESCRIPTION

1.1 Title

GENERATIVE AI FOR ENHANCEMENT OF CLINICAL DATASETS

1.2 Organisation Description

Organization name: *EUCAIM – Cancer Image Europe*
Website: *https://cancerimage.eu*
Sector / industry: Medical Imaging AI companies
Country / region: EU

EUropean Federation for CAncer IMages (EUCAIM) is the cornerstone of the European Commission-initiated European Cancer Imaging Initiative, a flagship of Europe's Beating Cancer Plan (EBCP), which aims to foster innovation and deployment of digital technologies in cancer treatment and care to achieve more precise and faster clinical decision making, diagnostics, treatment and predictive medicine for cancer patients.

Funded by the European Union

1.3 Challenge Description

Medical Imaging is a key source of information for developing AI models that can support diagnosis and prognosis. Training data has been traditionally obtained through clinical trials and research projects, and more recently from healthcare routine databases (Real World Data - RWD). While RWD can be larger and more representative of a specific population, it often lacks sufficient quality in terms of completeness, consistency, and harmonization.

The availability of cohorts tailored to the characteristics of a target population is essential for both the training and validation of AI-based software tools. Existing datasets may not fully match these populations and may be incomplete with respect to critical variables or imaging modalities. Moreover, AI models may show limited generalizability when applied to patient subgroups that are underrepresented in the training data. This is particularly challenging in low-prevalence subgroups, where restricted data availability can compromise both robust model development and proper validation. Generative AI can help address these limitations by creating synthetic data to complement existing datasets and improve their quality. This is especially relevant in situations involving imbalance or underrepresentation across demographic features, sex, comorbidities, concurrent conditions, tumor locations, and other clinically meaningful variables.

Generative AI can be used to enhance, complete, or generate missing values, and even full datasets to be used in training and benchmarking. These enhanced datasets can help to reduce bias and lead to models that are better adapted to specific populations.

OBJECTIVES

1. Develop a generative AI model that can generate synthetic cohorts based on the existing data provided by the challenge.
2. Identify the key demographic, clinical, and imaging characteristics to be used as input to the model.
3. Use the model to augment existing datasets and improve balance across relevant populations subgroups.
4. Evaluate the quality, realism, and consistency of the generated data with respect to the original data distribution.
5. Demonstrate the added value of synthetic data on the evaluation of the metrics on fidelity, bias and cohort imbalance reduction, completion and harmonisation.

EXPECTED OUTCOMES AND TRL LEVEL

1. Generative AI model for generating synthetic data
2. A tool that identifies the key variables driving synthetic cohort generation.
3. Augmentation of existing datasets
4. Report on the quality, realism, and consistency of the generated data with respect to the original data distribution with concentration to underrepresented variable subgroups.

The challenge is expected to produce a prototype model that can be used to improve the existing cohorts in EUCAIM and to create new synthetic cohorts for model training and validation (TRL 7). The model will be integrated in the infrastructure with the recognition of the developers, open to different exploitation paths and non-exclusive use in the platform.

1.4 Expected Impacts and KPIs

KPI 1. Synthetic Data Fidelity
Measure how well the synthetic cohort reproduces the statistical properties of the original data. Synthetic Data Fidelity will be evaluated over the entire cohort.
Metric:

- Average Kolmogorov-Smirnov (KS) distance across all continuous clinical variables.
- Average absolute difference in prevalence (%) for categorical variables.

Target:

- KS distance ≤ 0.10 for at least 80% of variables.
- Difference in category prevalence ≤ 5% for at least 80% of variables.

Funded by the European Union

KPI 2. Bias Reduction and Cohort Balancing
Measure the improvement in representation of underrepresented patient groups.
Metric:

- Reduction in imbalance ratio between the largest and smallest predefined demographic subgroup (e.g., sex, age group, tumor location).

Target:

- At least 50% reduction in imbalance compared with the original dataset. Example: If the original ratio is 10:1, the synthetic cohort should achieve a ratio of 5:1 or better.

KPI 3. Missing Data Completion and Imaging Harmonization
Measure the ability of the solution to complete missing clinical variables, generate missing imaging acquisitions, and harmonize imaging studies across acquisition protocols, considering the following variables: local_recurrence_progression; clinical_staging_date; clinical_stage_group; metastasis_lung; metastasis_pleura; metastasis_lymph_nodes; metastasis_adrenal_gland; metastasis_liver; metastasis_brain; metastasis_bone; metastasis_other; treatment_intent; smoking_status; date_smoking_status; ecog_performance_status; metastasis_clinical_category; progression_recurrence; distant_metastasis_pr; tumor_clinical_category; regional_nodes_clinical_category. And dosage information from KVP DICOM Tag (0018,0060).

Clinical Data Completion
Metrics:

- Accuracy for categorical variables.
- Mean Absolute Error (MAE) for continuous variables.

Target:

- Accuracy ≥ 90% for categorical variables.
- MAE improvement ≥ 20% compared with a baseline median imputation strategy.

Imaging Data Completion
The solution should be able to generate missing CT acquisitions to create complete imaging studies when one or more of the following series are unavailable:

- Non-contrast CT.
- Contrast-enhanced CT.
- Low-dose CT (non-contrast).

Metric:

- Imaging Completion Rate (ICR): percentage of missing imaging acquisitions successfully generated.

Target:

- ICR ≥ 95% of missing imaging acquisitions.

Imaging Harmonization
The solution should harmonize CT studies acquired using different tube voltage (kVp) settings. The relevant parameter is DICOM Tag (0018,0060) – KVP (Peak kilo voltage output).
Metric:

- Reduction in differences between CT image intensity distributions from scans acquired at different kVp values, measured using the Wasserstein distance before and after harmonization.

Target:

- ≥ 30% reduction in variability between scans acquired at different kVp settings after harmonization, while preserving imaging biomarker stability (median CCC ≥ 0.85 for the GLCM Joint Entropy radiomic feature).

Funded by the European Union

KPI 4. Robustness
Measure whether synthetic and harmonized data preserve quantitative imaging biomarkers under controlled perturbations.
Evaluation protocol: A reference lung segmentation tool will be applied to all CT scans. Radiomic feature GLCM Joint Entropy, measuring global texture randomness and heterogeneity, will be extracted from original and synthetic images after different Gaussian noise perturbations.
Metric:

- Concordance Correlation Coefficient (CCC) of the GLCM Joint Entropy radiomic feature before and after perturbation.

Target:

- Median CCC ≥ 0.85 at each of the three perturbation levels individually (σ = 5, 10, and 20 HU).

1.5 Data Framework

DATASETS PROVIDED

A dataset of Computer Tomography images from 1088 subjects with ages between 29 years and 91 years (both Male and Female), collected in the CHAIMELEON project for training and validation of AI models (https://zenodo.org/records/14046442)..

Data types provided:
The Challenge will provide 1088 DICOM Thorax CT imaging studies from baseline time-point (before treatment started), and a JSON file with the following variables: subject_id; date_baseline_ct; age_at_baseline; gender; tumor_histotype; pd_l1; pd_l1_unknown; local_recurrence_progression; death_related_to_cancer; clinical_staging_date; clinical_stage_group; metastasis_lung; metastasis_pleura; metastasis_lymph_nodes; metastasis_adrenal_gland; metastasis_liver; metastasis_brain; metastasis_bone; metastasis_other; treatment_intent; smoking_status; date_smoking_status; ecog_performance_status; metastasis_clinical_category; progression_recurrence; distant_metastasis_pr; tumor_clinical_category; regional_nodes_clinical_category; death_date.

DATA RIGHTS (LEGAL & ETHICAL)

Ownership & Access Rights: The applicants should bind to the Terms and Conditions of EUCAIM (https://dashboard.eucaim.cancerimage.eu/eucaim_usage_policy.pdf)

Personal & Sensitive Data: The datasets have been anonymised, but the data is sensitive, as it has been obtained from clinical practice. The data will only be accessible through a Secure Processing Environment deployed on top of the AI-BOOST resources.

Data Collection & Legal Basis: The data has been transferred to EUCAIM on the basis of a Data Transfer Agreement and they will be temporarily copied in a secure processing environment deployed on top of the AI-BOOST resources. The instructions of usage of the environment are provided in https://github.com/EUCAIM/upv-node-workstation-images/blob/main/usage-guide.md

Known Biases & Limitations: Imaging data have not been annotated (e.g. segmentation of the nodes). Data cannot be downloaded or copied to the personal area of the user in the HPC resources, but processed in the Secure Environment offered in the challenge. This limitation is a requirement to guarantee the traceability of the data access and to fulfill the conditions of the Data Transfer Agreement.

Funded by the European Union

Legal & Ethical Restrictions: Applicants should adhere to the General Terms and conditions of EUCAIM data as well as the specific terms and conditions for this dataset (https://chaimeleon-eu.i3m.upv.es/dataset-service/web/terms-and-conditions.pdf).

1.6 Evaluation Metrics and Protocol

The baseline is the original dataset provided in the challenge, which serves as the reference for all evaluation metrics. For each KPI, participant methods are evaluated as follows:

- KPI 1 (Synthetic Data Fidelity): synthetic data are compared against the statistical distribution of the original dataset.
- KPI 2 (Bias Reduction and Cohort Balancing): improvement is measured relative to the original cohort distribution, which reflects existing demographic and clinical imbalances. The reduction in imbalance between predefined subgroups in the synthetic cohort is computed with respect to the original dataset.
- KPI 3 (Missing Data Completion and Imaging Harmonization): the original dataset serves as ground truth. Missing values and imaging acquisitions are artificially masked, and participants must reconstruct them. For harmonization, non-harmonized CT scans grouped by kVp are compared against harmonized outputs.
- KPI 4 (Robustness): original CT images are used as reference to assess the stability of radiomic features under controlled perturbations.

In all cases, performance is assessed using the original dataset as reference, according to the specific metric defined in each KPI.

1.7 Infrastructure

Infrastructure & Testing Environment:
CINECA can provide a secure development environment compatible with the EUCAIM's data and the objectives of the challenge.

Technical & Deployment Constraints: The Secure Processing Environment provides sessions with PyTorch or Tensorflow, including the following software:

- Ubuntu 24.04.4 LTS
- Python 3.12.3
- dicom2nifti 2.6.2
- Jupyterlab 4.5.8
- Keras 3.14.1
- Numpy 2.4.6
- Torch 2.12.0

1.8 Responsible AI

Dataset Diversity & Known Gaps: The main objective of the challenge is to reduce the bias in the datasets, focusing specially on gender. Therefore, this is a main concern and it will be the focus of the work.

Explainability, Traceability & Human Oversight are not mandatory.

No known Risks & Bias Mitigation different from the objective of the challenge.

The data is located in a secure environment and there are no additional considerations rather the ones stated in the Terms and conditions.

1.9 Additional Support Offered by the Challenge Owner

The challenge owner will provide the following support:

1. Clinical guidance through the experts in the infrastructure, helping on defining the rightmost variables and the success criteria.
2. Software to access securely high-sensitive data under the conditions they were collected.
3. Support on the integration of the tools in the secure environment.
4. Support on the development of sustainability models through the Secure environment.
5. The challenge owner will compare the results obtained in the original and the improved dataset with an existing AI model that assesses the CT imaging and clinical data to predict overall survival. This information will be used in the promotion of the results of the challenge and could lead to a scientific publication in which the winning participants will be invited.

Funded by the European Union