INFORM
Interpretability of Deep Neural Networks for Radiomics

DMP: Data Management Plan
10 Nov. 2021

Project Id	
CHIST-ERA Call topic	CHIST-ERA Call 2019 - XAI Topic
Project acronym	INFORM
Project title	Interpretability of Deep Neural Networks for Radiomics
Project ID	CHIST-ERA-19-XAI-007
Start date	March 2021
End date	February 2024
Website	www.inform-project.eu
Coordinator of the project	
Name	Panagiotis Papadimitroulas
Organisation	BIOEMTECH
Country	Greece
Telephone	+30 2106548192
E-mail	panpap@bioemtech.com

BIOEMTECH

Laboratoire de Traitement de l'Information Médicale
Laboratory of Medical Information Processing

UNIWERSYTET WARSZAWSKI

1. Introduction

The current document aims to the definition of a Data Management Plan (DMP) that outlines how data are to be handled both during the INFORM research project, and after the project is completed. The main aspects of this document are:

- the **handling of research data during and after the end** of the project
- **what data** will be collected, processed and/or generated
- **which methodology and standards** will be applied
- whether **data will be shared/made open access** and
- how data will be **curated and preserved** (including after the end of the project).

The INFORM DMP provides a structured description of the personal and generic data. These data will be used for the achievement of the purpose of the research of the project. Additionally, this deliverable touches upon the preliminary vision of the type of processing, storage and possible re-use of the data sets. The deliverable will also introduce the internal processes that will be followed for collecting, processing and storing data, as well as the policies and procedures for strengthening the data security and privacy.

This DMP followed the main guidelines provided by the EU document “Guidelines on FAIR Data Management in Horizon 2020”.

2. Description of the data

2.1 Type of study

The INFORM consortium proposes to investigate explainable artificial intelligence (XAI) with a dual aim of i) building high performance deep neural networks (DNN)-based classifiers and ii) developing novel interpretability techniques specifically adapted for radiomics use cases. First, if the available clinical data are insufficient to properly train DNNs, we will investigate Monte-Carlo simulations (MCS) combined with generative adversarial networks (GAN) for producing highly realistic simulated/synthetic data in order to increase the number of cases to facilitate training. Second, we will tackle the interpretability of DNN-based feature engineering and latent variable modeling with innovative developments of saliency maps and related approaches for relevance scores. Both supervised and unsupervised learning will be used to generate features, which can be interpreted in

terms of input voxels, conventionally engineered, and expert-derived features. Third, we propose to build explainable AI models that incorporate both conventional radiomic and DNN-based features. By quantitatively understanding the interplay between expert-derived and DNN-based features, our models will be easier to understand and to translate into clinical use. Fourth, preliminary evaluation will be carried out with the help of clinical collaborators on predicting outcome of patients in a realistic setting, using real clinical data. These proposed DNN models, specifically developed to reveal their innerworkings, will leverage the robustness and trustworthiness of expert-derived features that medical practitioners are familiar with, while providing quantitative and visual feedback. Overall, our methodological research and clinical application will advance interpretability of feature engineering, generative models, and DNN classifiers with applications in radiomics and broad medical imaging.

2.2 Types of data

In the project there will be the following types of data:

a) Clinical data (Medical images and associated metadata, clinical information such as demographics or histopathology data)
b) Generated / synthesized data (MC simulated and GAN-based image synthesis)
c) Algorithms & DNN-based models for radiomics explainability.
d) Administrative data (management, dissemination, communication, exploitation, training purposes)

2.3 Format and scale of data

Partner 1 (BIOEMTECH) :

- 3D medical imaging PET/Ct data → input for MC simulations
  Format: .dicom, .img/hdr (includes metadata in header files)
  Total space: <10 Gb
- Python scripts → processing of medical imaging and simulated data.
  Format: .py
  Total space: <1 Gb
- Txt files (dat), incorporating macro-commands → input for MC simulations
  Format: .dat, .txt, .mac
  Total space: <1 Gb

- Root files and 3D medical imaging data → output from MC simulations

  Format: .root, .img/hdr

  Total space: <50 Gb
- Text files, documents, spreadsheets, presentations, figures, videos → management purposes, meetings' minutes, partners presentations, training material, dissemination and exploitation files.

  Format: .txt, .doc, .xls, .ppt, .avi, .jpeg, .png, .bmp

  Total space: <10 Gb

Partner 2 (LATIM) :

- 3D medical imaging PET/CT data and associated clinical data (some publicly available, some restricted) → input for DNN / radiomics modeling

  Format: .dicom, xls, csv, txt
- Total space: <1 Tb
- Algorithms → image processing workflow, DNN training, saliency maps, related outputs.

  Format: .py, ck, pt, npy

  Total space: <100 Gb

Partner 3 (UW) :

- 3D medical imaging PET/CT data (publicly available datasets) and associated meta data

  Format: .dicom, .csv, .xls, .txt

  Total space: <10 Gb
- Python scripts for image analysis

  Format: .py

  Total space: <1 Gb
- Deep learning models

  Format: .ck, .pt

  Total space: <1 Gb
- Importance estimators, saliency maps, and related outputs

  Format: .npy

  Total space: >100 Gb

3. Data procedures: collection / management / reuse / standards

3.1 Reused Data / Data collection

For the needs of the INFORM project, already acquired medical data will be used without requiring any generation of new patients' data who will undergo any medical procedure for the purposes of the project. Data are going to be used through open databases and through access in databases from the HECKTOR MICCAI challenge. INFORM partners will ensure the anonymization of the data before any use and processing for the project, in case the data has not been anonymized already (for instance, the publicly available HECKTOR dataset is already anonymized).

More specifically, datasets from MICCAI HECKTOR 2021 and 2022 challenges will be used prioritarily as a use case in the project, including data from 325 and >880 patients respectively. The available data are PET/CT, F18-FDG, head & neck cancer type, in .dicom format along with clinical information (csv file) and are already anonymized. Two other publicly available datasets will be exploited, namely the Lung Image Database Consortium image collection (LIDC-IDRI) and A Lung Nodule Database (LNDb). In total, 1312 patients anonymized CT scans of the lungs are available, with more than 320k slices.

3.2 Technical details for data sharing

PLACIS[1] is a high capacity storage and intensive computing platform located in the LaTIM, in Brest, France. It offers 800 CPU cores and 72 GPUs dedicated to calculations, and a storage capacity of about 200TB. An "INFORM" account was created on PLACIS and a dedicated storage space (1 TB) has been allocated. This account can be accessed from outside partners (Bioemtech and Univ Warsaw) through an "Univ Brest" account that is available for use by both partners.

3.3 Data storage and curation

All image data will be provided from modalities that abide by the DICOM standard. Furthermore, all imaging modalities, should follow routine Quality Assurance (QA) and Quality Control (QC) that are

[1] http://placis.univ-brest.fr/english

detailed in national authorities' standards and protocols[2] [3] (HAMP (Hellenic Association of Medical Physicists) and SFPM (French Society of Medical Physics)). That means that when an inconsistency is observed then appropriate correcting mechanisms will be applied to return the modality to the proper fine-tuned status.

All generated data of the partners including, partners' presentations, videos, deliverables, training material, posters, publications, will be stored in BIOEMTECH's cloud server as well as, will stored in a dedicated location in the official INFORM's website, accessible only to authorized members.

3.4 Metadata standards and data documentation

Each collected or generated data will be followed by a metadata file. The metadata will define information about the type, format, creation date, purpose of use, creator and level of access of the data. The basic metadata fields that must be filled from each partner that collects or creates metadata are:

- **Data information**
  - Data description (max 150 characters)
  - Data origin (collected, created, or reused)
  - Data type (clinical, synthetic/simulated)
  - Data format
- **Data storage**
  - Data repository during the project (Fileserver)
  - Data repository after the project is complete (Fileserver or other)
  - Data size (in KBs)
- **Data Access Control**
  - Data Ownership (max 250 characters)
  - Data sharing (Shared only to the consortium, Shared outside the consortium)
- **Data life cycle**
  - Status of data at the end of the project (destroyed or not)
  - The duration of the data preservation after the end of the project (in years)

[2] https://www.efie.gr/index.php/gr/protocols-title

[3] http://www.sfpm.asso.fr/download/

The above metadata structure is subject under change during project if there is need for the utilization data for the research's purposes.

3.5 Data preservation strategy and standards

Medical Data

The collected medical data is planned to be stored for the duration of the project and for five years after the end of the project. After this period, all the medical stored data (collected for INFORM, not through Open databases), will be stored, anonymized, in open public repositories. The only data that will remain in the cloud storage will be the outcomes of statistical or data analyzes, and the system's logs.

3.6 Data security and open access standards

Also, the EU General Data Protection Regulation (GDPR) and HIPAA directives will be followed to enforce project's sensitive data protection. FAIR (Findability, Accessibility, Interoperability, and Reuse) principles will be followed to the data collected and generated (where applicable) within the project. INFORM will adhere to the EC's Open Access Guidelines rules, ensuring the open access policy to the produced results. All the academic beneficiaries of the suggested project will apply and get an Ethical Approval from their Institutions.

4. Conclusion

The purpose of this deliverable is the creation of research Data Management Plan for the project INFORM. During the project, four types of data will be used/generated (see section 2.2).

The collection, processing and sharing of the data will follow specific procedures that will be compliant with GDPR guideline from EU. For each type of data, the project's researchers will use specific data format and they must keep specific metadata information. The data storage and sharing will be done in PLACIS (LaTIM) and in BIOEMTECH's server. At the start of the project, there is no anticipation for sharing research data with non-consortium entities. In any case, the current Data Management Plan will continuously be updated when there is a need for new type of data or decisions for data sharing that are mandatory for the implementation of the research during project.