



Karolinska
Institutet

MICCAI/AFRICAI Summer school 2023

Model development (*part 1*)

Data-centric best practices and common pitfalls and open access infrastructures

Apostolia Tsirikoglou

Postdoctoral researcher on AI for Breast Imaging

Department of Oncology-Pathology

Karolinska Institutet

apostolia.tsirikoglou@ki.se

Martijn P. A. Starmans

PostDoc AI for Integrated Diagnostics in Oncology

Department of Radiology and Nuclear Medicine &

Department of Pathology, Erasmus MC, Rotterdam, NL

m.starmans@erasmusmc.nl

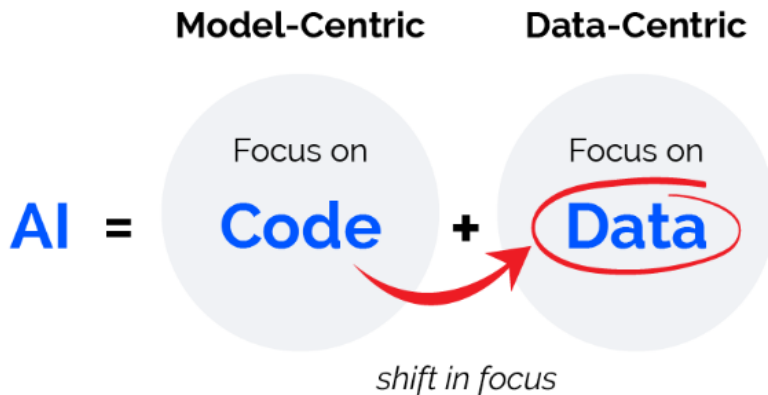
Agenda

- Data-centric AI
- Open access datasets
- Know your data
 - Data inspection and curation
- Data splitting
- Data preprocessing
- Data augmentation
- Data documentation

Data-centric AI

What Is Data-Centric AI?

Data-Centric AI is the discipline of systematically engineering the data used to build an AI system. Think of a Data-Centric AI system as programming with **focus on data instead of code**. Industries of all types continue to adopt AI solutions, and while AI models have improved over the years, a fundamental shift is needed to truly unleash AI's full potential.



SOURCE:  LANDING AI

Open access datasets

The Cancer Imaging Archive (TCIA)

<https://www.cancerimagingarchive.net/>

- 10.000 + public (free) radiology and histopathology imaging datasets

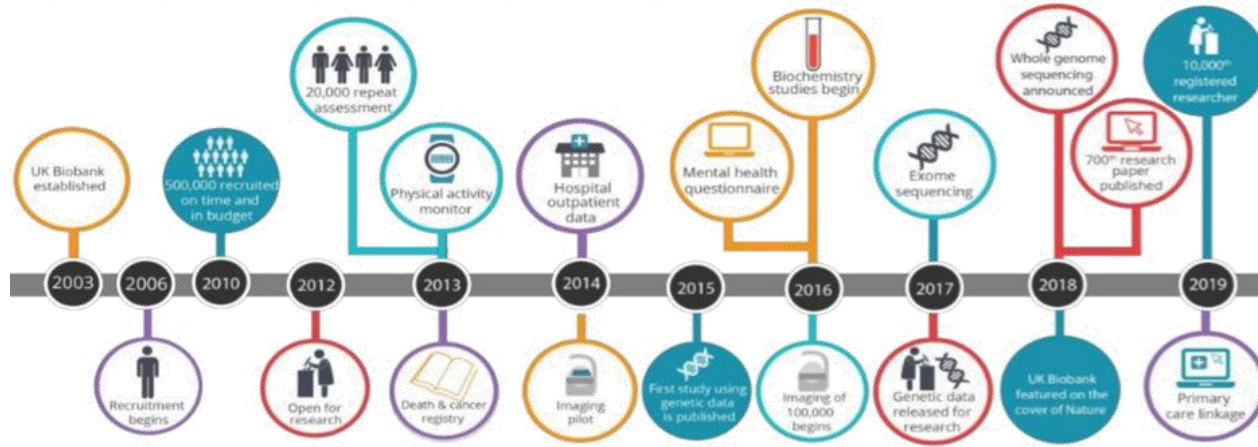
Collection	Cancer Type	Location	Species	Subjects	Data Types	Supporting Data	Access	Status	Updated
APOLLO-5	Breast Cancer, Bladder Urothelial Carcinoma, Cholangiocarcinoma, Colon adenocarcinoma, Cutaneous Melanoma, Endocrine Miscellaneous, Esophageal Carcinoma, Gastrointestinal Stromal Tumor, Head and Neck Squamous Cell Carcinoma, Kidney Chromophobe, Kidney Clear Cell Renal Cell Carcinoma, Kidney Renal Papillary Cell Carcinoma, Liver Hepatocellular Carcinoma, Lung Adenocarcinoma, Lung Other, Lung Squamous Cell Carcinoma, Major Salivary Gland, Miscellaneous, Neuroendocrine Tumors (all sites), Ovarian Cancer, Pancreatic adenocarcinoma, Pathologically Benign, Prostate Adenocarcinoma, Soft Tissue, Thymoma, Thyroid Carcinoma, Uterine Corpus Endometrial Carcinoma	Bile duct, Bladder, Breast, Colon, Endocrine, Esophagus, GI, Head-Neck, Kidney, Liver, Lung, Ovary, Pancreas, Prostate, Skin, Soft Tissue, Thymus, Thyroid, Uterine corpus	Human	273	CT, MR, NM, PT, US, XA		Limited	Ongoing	2023-08-31
HNSCC-miF-miHC-comparison	Head and Neck Cancer	Head-Neck	Human	8	Pathology	Image Analyses	Public	Complete	2023-08-31
CT-Phantom4Radiomics	Phantom	Phantom	Human	1	CT, SEG	Image Analyses	Public	Complete	2023-08-23
Vestibular-Schwannoma-MC-RC	Vestibular Schwannoma (non-cancer)	Ear	Human	124	MR, SEG		Public	Complete	2023-08-23
CPTAC-UCEC	Corpus Endometrial Carcinoma	Uterus	Human	250	CT, MR, PT, US, Pathology	Clinical, Genomics, Proteomics	Public	Ongoing	2023-08-18
CPTAC-CCRCC	Clear Cell Carcinoma	Kidney	Human	222	CT, MR, Pathology	Clinical, Genomics, Proteomics	Public	Ongoing	2023-08-18
CPTAC-PDA	Ductal Adenocarcinoma	Pancreas	Human	168	CT, MR, PT, US,	Clinical	Public	Ongoing	2023-08-18

- Note: The Cancer Genome Atlas Program (TCGA) has genomics and some histology data.

Population studies / biobanks

<https://www.ukbiobank.ac.uk/>

- Collect numerous datatypes of initially healthy population over a long time period
- Example: UK Biobank



- “Public”: submit access request, potential fees

Challenges

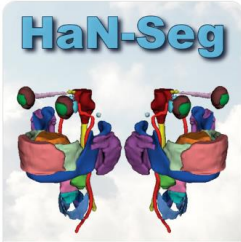
- Idea: compare and benchmark methods on a common (hidden) dataset.
- Substantial role within MICCAI (<http://www.miccai.org/special-interest-groups/challenges/miccai-registered-challenges/>)

Leading platform:

<https://grand-challenge.org/>

Grand Challenge Challenges Algorithms ... Help Sign In Register

179 challenges found




HaN-Seg

The Head and Neck Org...

Algorithm submission challenge

Accepting submissions for Preliminary Test Phase until Oct 31 2023 at 23:59

294 5 2022




Shifts Challenge 2022

Algorithm submission challenge

Accepting submissions for MS Lesion Segmentation: Phase II until Apr 09 2024 at 00:59

191 99 Article 2022




SynthRAD 2023

SynthRAD2023

Algorithm submission challenge

Opening submissions for Post-challenge Task 1 - MRI on Sep 20 2023 at 00:00

661 198 Oct. 8, 2023



PI-CAI

ARTIFICIAL INTELLIGENCE & RADIOLOGISTS AT PROSTATE CANCER DETECTION IN MRI

The PI-CAI Challenge

Algorithm submission challenge

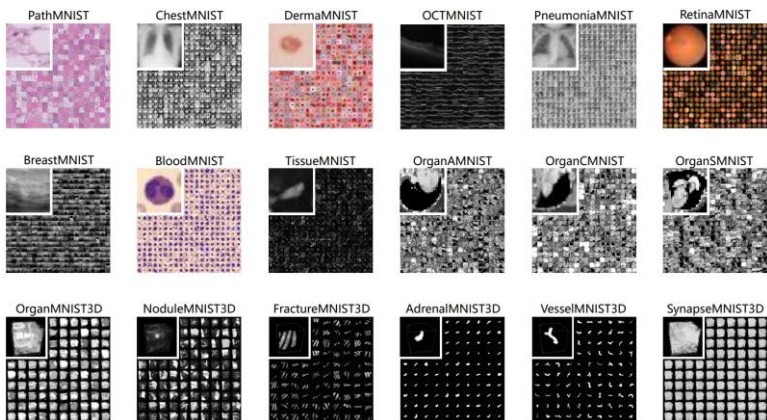
Accepting submissions for Open Development Phase - Validation and Tuning

1,444 363 Article 2022

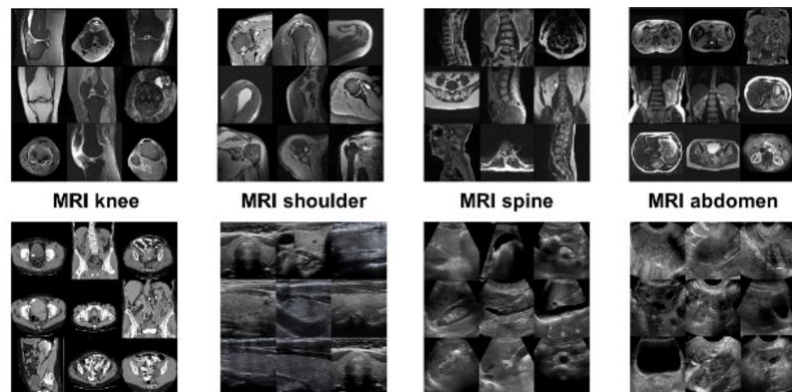
Other individual datasets

Some popular, general individual datasets or repositories, e.g.:

MedMNist (<https://medmnist.com/>)



RadImageNet (<https://www.radimagenet.com/>)



Take home message: always look for a public dataset or challenge for your specific application, which you can use for model development or benchmarking

Know your data

Data inspection and curation

- Types and modalities
 - Digital pathology (virtual microscopy)
 - Medical imaging in radiology
 - Ultrasound
 - X-Ray Imaging
 - Computer Tomography (CT)
 - Magnetic Resonance Imaging (MRI)
 - Positron Emission Tomography (PET)
 - Endoscopy

Know your data

Data inspection and curation

- inspect them according to their format and analyze all the information
 - from pixel data and ground truth labels statistics, outliers, irregularities and metadata analysis.
 - DICOM, NIfTI
- *Important note: **sensitive personal information***

Data splitting

- **Never use the same data for model training and evaluation**
- training/validation/testing
 - External clinical validation
- Techniques
 - Random
 - Stratified
 - Non-random
 - Cross-validation

Data splitting

- **Tips and tricks**

1. If there are many hyperparameters to tune, a larger validation set is preferred to optimize the model performance.
2. Validate the model after each epoch to make the model learn varied scenarios.
3. Experiment with the split ratio.

Data preprocessing

- Prepare and unify data
 - Some processes can be under data curation too, e.g., image registration
- Examples
 - image orientation correction/transformation,
 - resizing/resampling
 - cropping or padding

Data augmentation

- Standard practice
 - artificially expand the size of a training set by creating modified or synthetic data
- Examples
 - *Geometric* – flip, crop, rotate, stretch, zoom or translate
 - *Color space* – random change RGB color channels, contrast, intensify and brightness
 - *Kernel filters* – sharpen or blur an image
 - *Random erasing* – delete a part of the initial image

Data augmentation

- **Tips and tricks**
 - Choose proper augmentations for your task.
 - Be careful when applying multiple transformations on the same images.
 - Display augmented data before starting training on them.

Data documentation

- Helps to promote reflection and transparency about how these datasets might affect machine learning models
- Reveal underlying assumptions and potential risks,
- Increase reproducibility and
- Contribute to informed decision-making about whether specific datasets meet developing/evaluating needs.
- *Include a dataset datasheet in the appendix/supplementary of your paper!*

Time for hands-on!