Research Engineer Entrance Test

At Hera-MI we work with medical imaging. Our ambition is to create insightful tools for clinicians to help them in breast cancer screening and diagnosis. Many of the algorithms rely on machine learning techniques, and deep learning, in particular. Well-curated data are crucial to perform the relevant experiments, including training and evaluation.

In this assignment, you will need to build a complete experimental pipeline, starting from curation, through the design of an algorithm and to evaluation of its performances. That is, we need to collect data from an external source, prepare them for a machine learning experience, run such an experience, and reconcile the results of the experience in a way that could be shared with other researchers.

There are 5 tasks. Some of them have optional subtasks that would allow you to score more. You can also ask for hints by email.

1. You need to get the data from the CMMD dataset. The dataset is available on the Cancer Imaging Archive website (https://www.cancerimagingarchive.net). We need a tool that collects the data in an automated manner. We might not need or be able to download the whole dataset so your code should be allow to process a portion of it.

   Bonus: wrapping the tools in Docker container will be useful.

2. Once the images and the provided clinical data are downloaded, you need to prepare the dataset for the deep learning experiments. First, you should structure the data in a way to allow them to be used and queried effectively. You are free to choose any manner that you find appropriate.

3. Next, we need a code that would split data in a consistent and reproducible manner. You need to choose an appropriate splitting strategy. In this assignment, our main target would be to train a classifier to separate benign and malignant samples.

4. Now we need to run a deep-learning experiment. While we are not targeting the top-performing algorithm, we are looking for a meaningful evaluation compared to the state of the art. Your code should be able to read input data that are split according to the strategy defined in the previous step. The result of the experiments shall be reusable and sharable metrics. You are free to use any metrics you consider appropriate for the task to make them comparable to the state of the art.

   The DICOM data might not be the most optimal for running experiments, so you can implement a pipeline step that would preprocess and store the images in a different format.

   Bonus 1: While the main target of our experiment is separating benign and malignant samples, you are free to introduce other classification tasks of your choice (e.g., type of abnormality, breast density).

Bonus 2: You are welcome to address uncertainty and explainability of the algorithm in any way you find relevant.

5. Finally, we want the generated metrics to be shared with others. There are plenty of tools for the job and you are free to use any that you find the most appropriate.

Please keep in mind a few things:

- There's a myriad of frameworks and tools, but if you opt for a neural network we would prefer if you used Python as programming language, Docker for containerization, and Tensorflow for deep learning.
- While building Docker container is explicitly stated as Bonus in the first task, it is applicable to all the other tasks as well.
- You are free to dive as deep as you want in the deep learning experiments.
- You are expected to use a code repository (private gitlab or github repository is fine) and commit the code carefully. Send an invite to **mickaeltardy**.

Your code should be comprehensive. Please comment when needed.

Your deliverables shall include:

- source codes;
- build and execution instructions;
- file of files with metrics and illustrations;
- references of any used state-of-the-art methods.
- any additional instructions you judge necessary.

Note, that while the metrics are part of the deliverables, your work is not evaluated by the level of the obtained performances.

For any questions related to this assignment, or requests for hints, please reach out to Mickael Tardy, mickael.tardy@hera-mi.com, (Phone/WhatsApp: +33-637-936-235)