A Post Processing Pipeline to Prepare Raw Data for Machine Learning Algorithms in Cardiac Magnetic Resonance Imaging

F. Agostini[1,2] S. Campese[1] R. Vianello[3] M. Pizzi[3] A. Cipriani[3] M. Zanetti[1]

[1]Department of Physics and Astronomy, University of Padua
[2]Department of Control and Computer Engineering, Polytechnic University of Turin
[3]Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padua

federico.agostini@polito.it *stefano.campese@phd.unipd.it* *alberto.cipriani@unipd.it*

Background

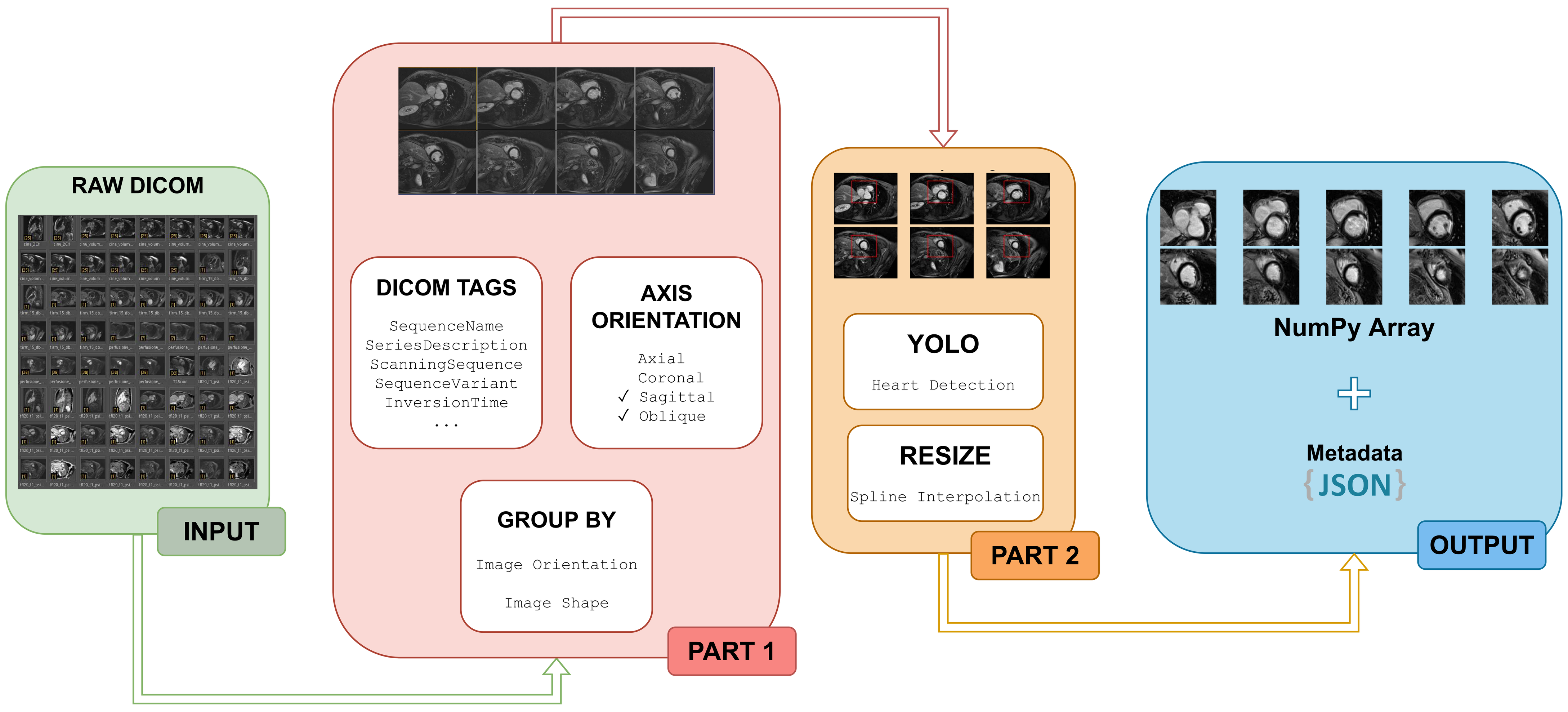
Artificial Intelligence is an emergent tool in clinical practice for post processing of medical images, which requires ad hoc pipelines to extract relevant data. A common issue is represented by **noisy datasets**, like those of **Cardiac Magnetic Resonance** (CMR) **studies**, characterized by multiple images, different acquisition techniques, slices, axis orientation and contrast timing.

Purpose

In this work, a **ML pipeline** for **extraction of Late Gadolinium Enhancement** (LGE) **images from raw DICOM** data is presented. In addition, steps for **normalization of image number** and automatically **heart localization** are outlined.

Methods

642 consecutive **CMR studies** are analyzed. Data are composed by **raw DICOM** file plus **annotations** made by expert doctors in the form of an Excel file, where the presence of Myocardial Fibrosis is indicated along side its location in the bullseye diagram of the heart.



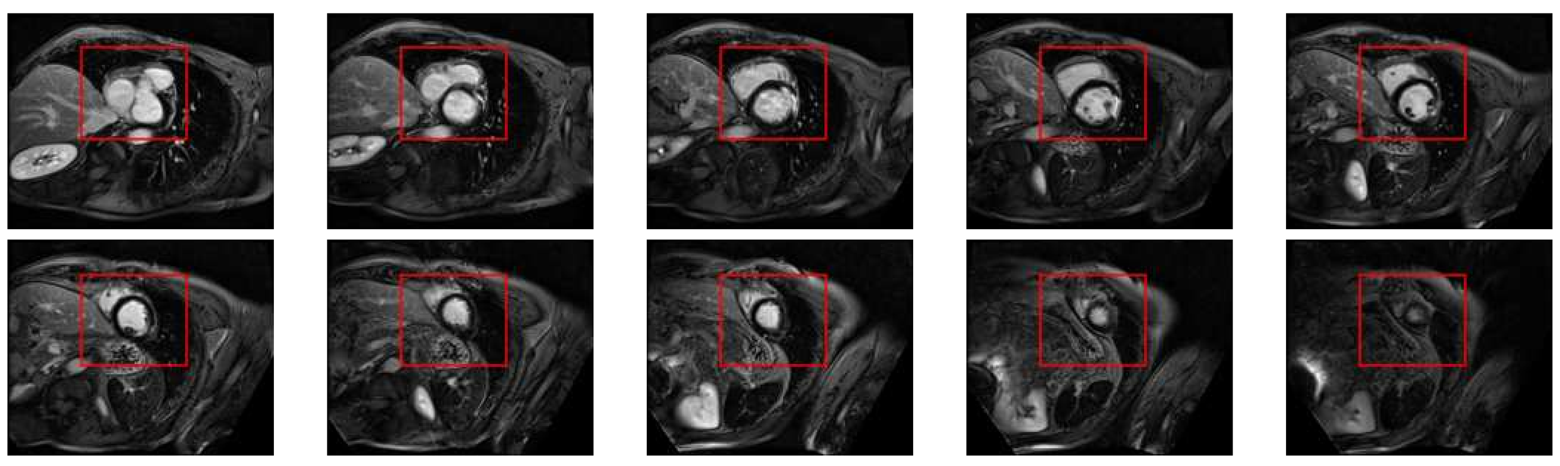
Pipeline, Part 1

By looking at the metadata in raw files, `SequenceName` tag is used to **discard cine images**, `ScanningSequence` tag to select **Gradient Recall** and **Inversion Recovery** techniques (Inversion Time > 100 ms), `SequenceVariant` tag to **discard Steady State images**. Orientation of the **major axis** is computed and **`Axial` or `Coronal` images removed**. Scans are **grouped together by image orientation** (requesting a min and max number of elements per group) and only the group with the largest number of files is selected. Finally, DICOMs are **grouped by image shape** (demanding a min number of elements), and only the series with the highest resolution is retained. Then, for each subject, the extracted series consists of a **3D-array** `(N,H,W)`, with `N` number of slices, and `(H,W)` image resolution. The **attributes** are **not homogeneous** between subjects.

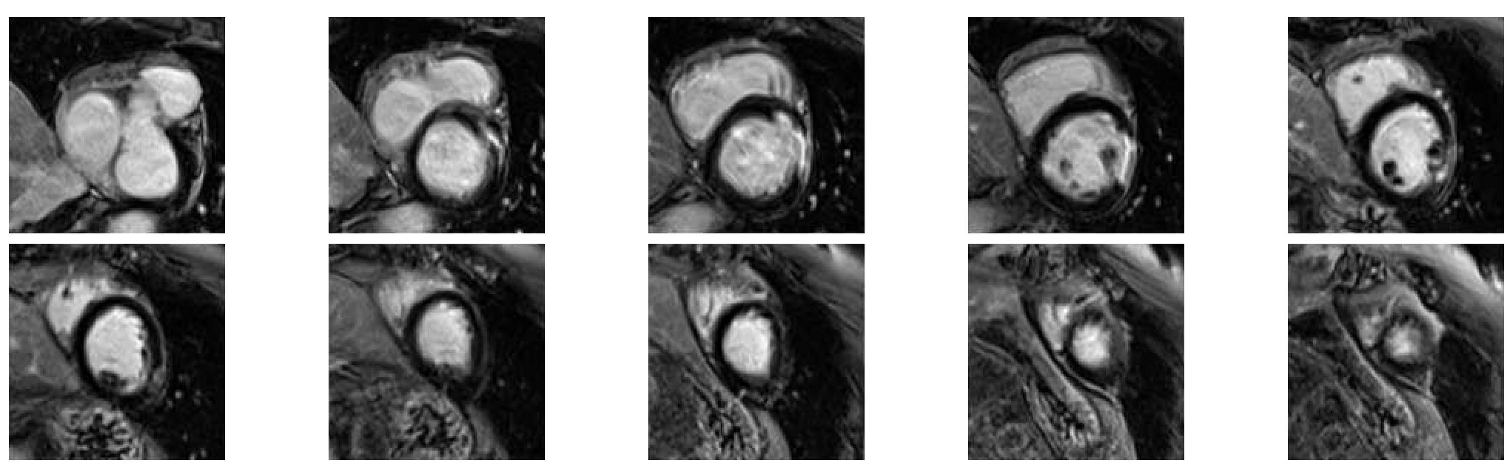
Pipeline, Part 2

Given a desired final number of slices and resolution, the 3D-array is reshaped through a **spline interpolation**. In order to have a focus on the heart, a **Region of Interest** (ROI) **extractor** was implemented, based on a **YOLO network** for **object detection**. The network is applied to all the slices; then the images were cropped by keeping the largest bounding box. This step allowed us to **remove the background** by only selecting the relevant ROIs. To manage the data more easily, images were saved as a **NumPy Array**, while other useful Dicom metadata (e.g. weight, age, ...) were stored using the **JSON** standard.

Input Images with YOLO box



Extracted Images after the Pipeline



YOLO Training

YOLO network for heart detection is **fine-tuned** starting from **darknet weights** (trained on COCO dataset). **1500 CMR** heart images are **manually labelled by experts**; in addition, syntetic examples are generated through **data augmentation** tecniques (rotations, random noise, brightness and contrast), obtaining a total of **10500 images** (7 times the original ones). In the first training stage pretrained layers are frozen to get a stable loss, then the whole model is fine-tuned. Overfitting is avoided monitoring the **validation loss** and using **early stopping**. Training is performed using a **Nvidia V100 GPU** with 32Gb of memory.

Conclusions

At the end of the pipeline, images can be reduced to a **common resolution** and forwarded to ML algorithms. In our study, the original dataset extended for about **200 GB**; by requesting **10 slices** per subject with a resolution of **128 by 128 pixels** (also extracting heart ROI) the final dimension was reduced to **108 MB**, granting also a huge reduction in terms of storage.

Declaration of Interest The authors declare not to have any past or present conflict of interest that concerns the work hereby presented.