

A post processing pipeline to prepare raw data for machine learning algorithms in cardiac magnetic resonance imaging

*Original*

A post processing pipeline to prepare raw data for machine learning algorithms in cardiac magnetic resonance imaging / Agostini, F; Campese, S; Vianello, R; Pizzi, M; Cipriani, A; Zanetti, M. - In: EUROPEAN HEART JOURNAL. CARDIOVASCULAR IMAGING. - ISSN 2047-2404. - 23:Supplement\_2(2022). [10.1093/ehjci/jeac141.017]

*Availability:*

This version is available at: 11583/2971584 since: 2022-09-21T16:22:44Z

*Publisher:*

Oxford Academic

*Published*

DOI:10.1093/ehjci/jeac141.017

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

# A post processing pipeline to prepare raw data for machine learning algorithms in cardiac magnetic resonance imaging

Doctor Agostini F<sup>1</sup>, Mr Campese S<sup>1</sup>, Mr Vianello R<sup>1</sup>, Mr Pizzi M<sup>1</sup>, Doctor Cipriani A<sup>1</sup>, Professor Zanetti M<sup>1</sup>

<sup>1</sup>University of Padua, Padova, Italy

**Funding Acknowledgements:** Type of funding sources: None.

**Background:** Artificial Intelligence is an emergent tool in clinical practice for post processing of medical images. Machine Learning (ML) pipelines are created for data of interest extraction and algorithm application. A common issue in data extraction is represented by noisy datasets, like those of CMR studies, characterized by multiple images, acquired by different techniques, axis orientation and contrast timing.

**Purpose:** A ML pipeline for extraction of LGE images from raw DICOM data is presented. Additionally, steps for normalization of image number and automatically heart localization are outlined.

**Methods and Results:** 642 consecutive CMR studies were analyzed.

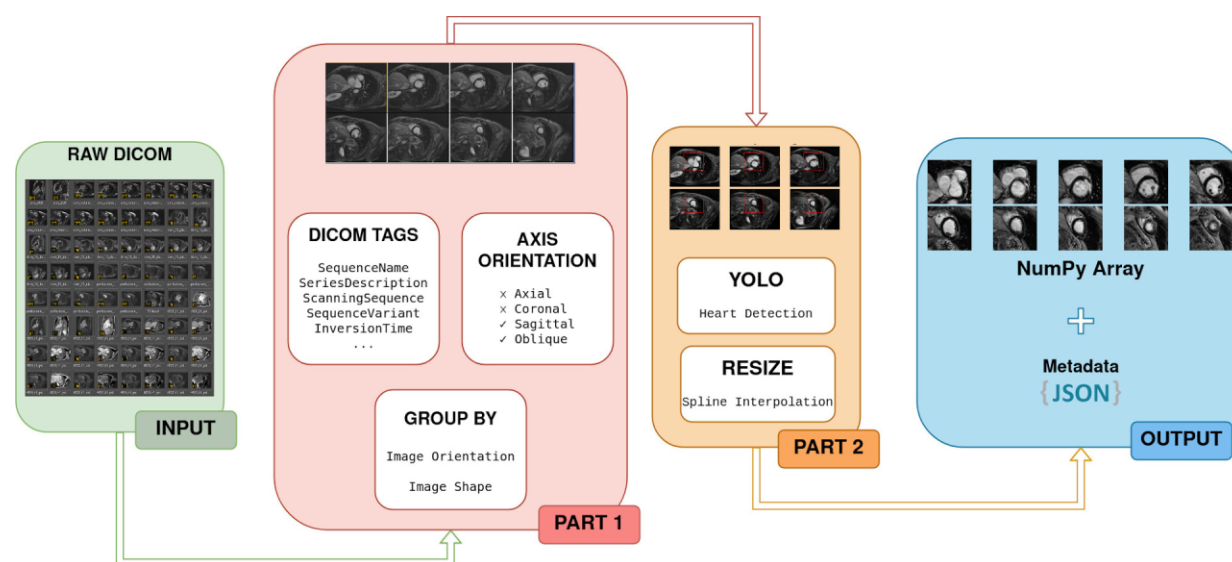
Pipeline, Part 1. By looking at the metadata in raw files, 'SequenceName' tag was used to discard cine images, 'ScanningSequence' tag to select Gradient Recall and Inversion Recovery techniques (Inversion Time > 100 ms), 'SequenceVariant' tag to discard Steady State images (See Fig. 1). Orientation of the major axis was computed and 'Axial' or 'Coronal' images removed. Scans were 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 was selected. Finally, DICOMs were grouped by image shape (demanding a min number of elements), and only the series with the highest resolution was 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 were not homogeneous between subjects.

Pipeline, Part 2. Given a desired final number of slices and resolution, the 3D-array was 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 was applied on all the slices (Fig. 2); 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.

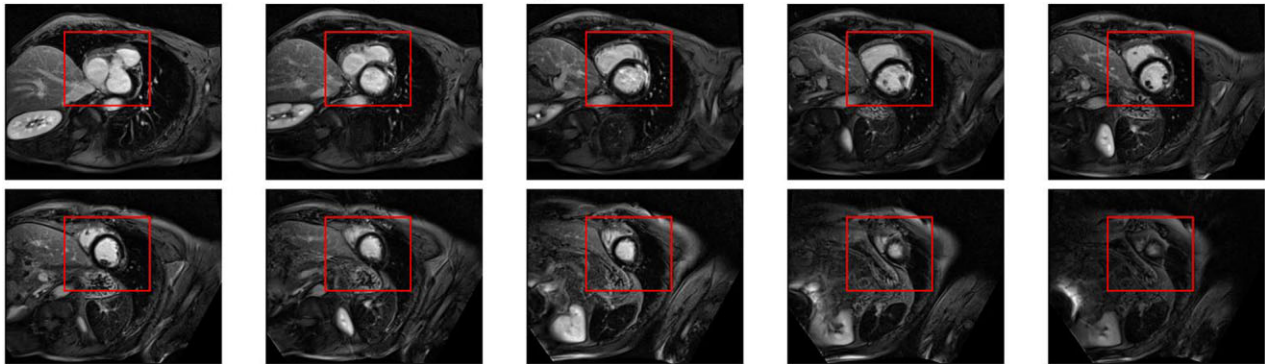
At the end of the ML pipeline, images can be reduced to a common resolution and forwarded to ML algorithms.

By using this pipeline, a great amount of information not needed for LGE analysis can be discarded, granting a significant reduction in terms of data storage. In our series, 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.

**Conclusions:** In this work, we presented a post-processing pipeline for CRM images and LGE analysis. Given in input raw CRM the pipeline is able to (i) remove useless data, (ii) extract heart ROIs storing also Dicom metadata, (iii) normalize slices and image resolution, and (iv) store the processed CRM into ready-format for ML techniques.



**Input images with YOLO box**



**Extracted images after the pipeline**

