Radiomic feature extraction

Valeria Amato

20/12/2021

Abstract

This report describes a framework able to extract radiomic features, thanks to the open-source python package: pyradiomics (1). Radiomic data has the potential to uncover disease characteristics that fail to be appreciated by the naked eye. The central hypothesis of radiomics is that distinctive imaging algorithms quantify the state of diseases, and thereby provide valuable information for personalized medicine. The paper defines how to extract lung tumor nodules characteristics, which are in The Cancer Imaging Archive, exactly in the collection LIDC-IDRI. The framework provides a simple interface for the extraction of the main features by only inserting paths to the DICOM directories.

1 Introduction

The framework is able to extract radiomic features, thanks to the open-source python package: pyradiomics (1). This package establishes a reference standard for Radiomic Analysis, and provide a tested and maintained open-source platform for easy and reproducible Radiomic Feature extraction(Warning: not intended for clinical use.).The platform supports both the feature extraction in 2D and 3D and can be used to calculate single values per feature for a region of interest.

The framework is implemented to extract features from the LIDC-IDRI dataset in The Cancer Imaging Archive (TCIA), a service which de-identifies and hosts a large publicly available archive of medical images of cancer. TCIA is funded by the Cancer Imaging Program (CIP), a part of the United States National Cancer Institute (NCI), and is managed by the Frederick National Laboratory for Cancer Research (FNLCR). line

2 How to use it?

2.1 Setup

2.1.1 Cancer Imaging Archive - Data Download

The imaging data in The Cancer Imaging Archive are organized as “collections” defined by a common disease (e.g. lung cancer), image modality or type (MRI, CT, digital histopathology, etc) or research focus. DICOM is the file format used by TCIA for LIDC-IDRI.

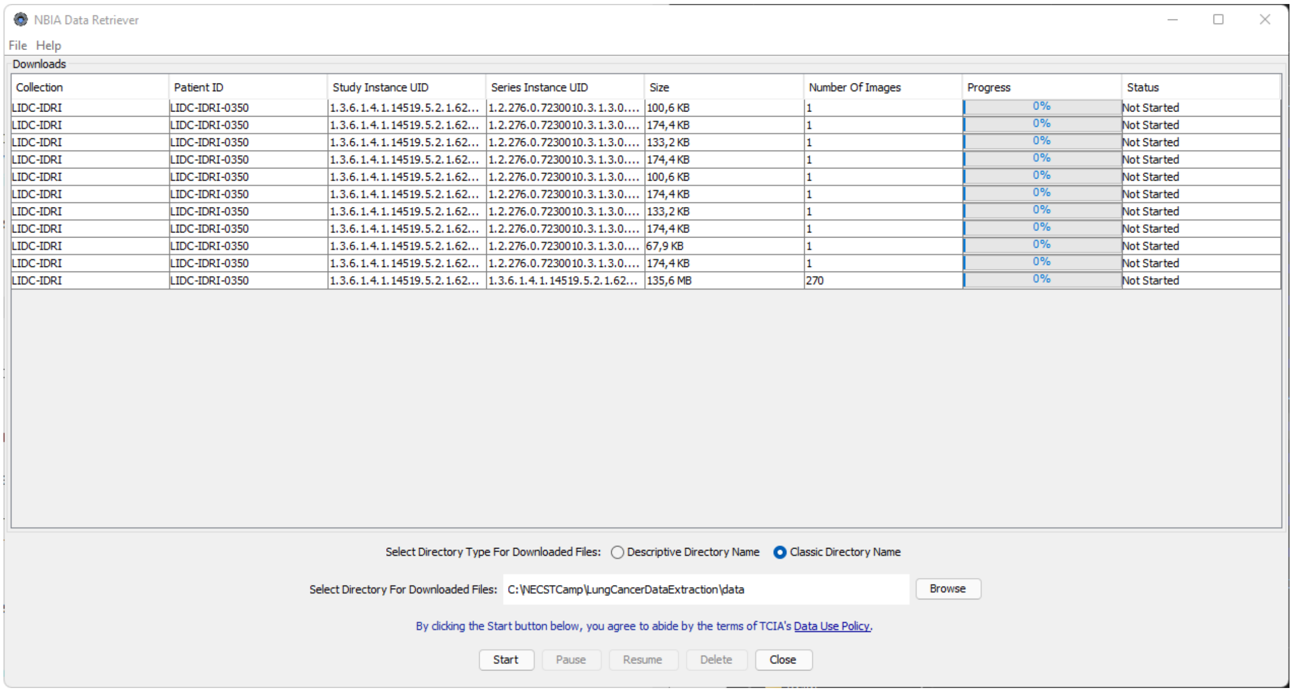
Once in the LIDC-IDRI dataset in The Cancer Imaging Archive, add in the chart a collection ID and download it.

If it is the first time, you may not have installed the NBIA Data Retriever on your computer. You must install the NBIA Data Retriever before you begin downloading. You only have to install it once to use it in future downloading sessions.

From this point, you must have installed NBIA Data Retriever.

When the download of the TCIA file is completed, click on it and let NBIA open it. You must agree to the data usage policy before you can proceed with downloading the data in your cart. Note that this policy is included in the license file accompanying your download.

Select directory type for downloaded files: Classic Directory Name Select directory for downloaded files. Start.



Collection	Patient ID	Study Instance UID	Series Instance UID	Size	Number Of Images	Progress	Status
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	100,6 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	174,4 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	133,2 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	174,4 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	100,6 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	174,4 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	133,2 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	174,4 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	67,9 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.2.276.0.7230010.3.1.3.0....	174,4 KB	1	0%	Not Started
LIDC-IDRI	LIDC-IDRI-0350	1.3.6.1.4.1.14519.5.2.1.62...	1.3.6.1.4.1.14519.5.2.1.62...	135,6 MB	270	0%	Not Started

Select Directory Type For Downloaded Files: ○ Descriptive Directory Name ● Classic Directory Name

Select Directory For Downloaded Files: C:\NECSTCamp\LungCancerDataExtraction\data

By clicking the Start button below, you agree to abide by the terms of TCIA's Data Use Policy.

Figure 1: A screenshot of NBIA Data Retriever with a TCIA file to load. Note: Classic Directory Name preferred

Now, you can access to the dicoms directory. The setup is finished.

3 Script clarification

3.1 Inputs

The inputs required are used to locate dicoms and metadata inside the directory. Pay attention do not cause the unicode error, for example not using two backslashes (
) for the path.

3.1.1 dirName

The variable dirName indicates the path to the DICOMs folder. In some cases, there are collection with more studies and multiple series inside each study. However, only one serie is complete of xml and many DICOMs. Take that path.

3.1.2 pid

The variable pid is a string which has the corresponding patientID.

3.1.3 imagePath

The string imagePath indicates the path in which converted DICOMs will be stored. The conversion is from a series of DICOMs into a SimpleITK object (NRRD) that can be read by the pyradiomics (1) package.

3.1.4 maskPath

The string maskPath indicates the path in which the 3D mask (file format: NRRD) of the volume is stored.

3.2 Extraction

Once all DICOMs have been loaded and converted into NRRD, the script extract the mask from the XML file referring to the patientID.

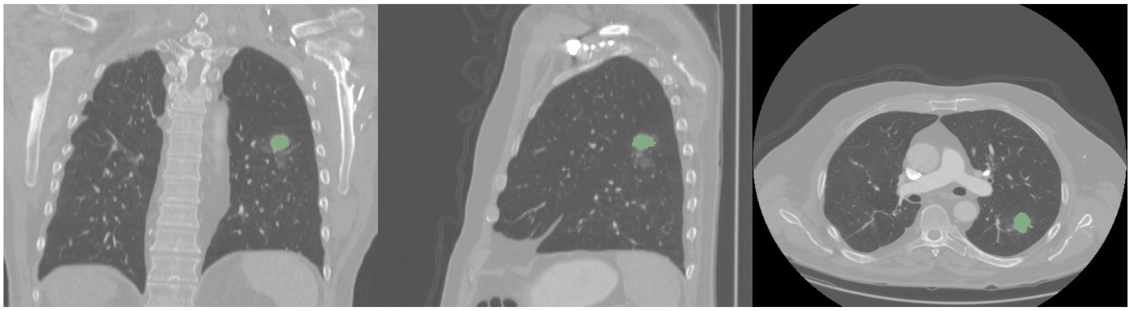


Figure 2: Three sides of lung image with the segmented mask (green).

Then, the extraction can begin.
The extraction has the following parameters:

- minimumROIDimensions: 2
- minimumROISize: None
- normalize: False
- normalizeScale: 1
- removeOutliers: None
- resampledPixelSpacing: None
- interpolator: 'sitkBSpline'
- preCrop: False
- padDistance: 5
- distances: [1]
- force2D: False
- force2Ddimension: 0
- resegmentRange: None
- label: 1
- additionalInfo: True

and any filter is applied.

3.3 Outputs and Results

3.3.1 No DICOMs in the directory

The script prints on stdout the absence of DICOMs and ends.

3.3.2 Features extracted

All features enabled (extracted):

- firstorder
- glcm
- gldm
- glrlm
- glszm
- ngtdm
- shape

4 Conclusion

The framework is able to extract various statistics, which are singly explained below.

4.1 First Order Feature

First-order statistics describe the distribution of voxel intensities within the image region defined by the mask through commonly used and basic metrics.

4.2 Gray Level Co-occurrence Matrix (GLCM) Features

A Gray Level Co-occurrence Matrix (GLCM) of size N×N describes the second-order joint probability function of an image region constrained by the mask.

4.3 Gray Level Dependence Matrix (GLDM) Features

A Gray Level Dependence Matrix (GLDM) quantifies gray level dependencies in an image. A gray level dependency is defined as a the number of connected voxels within distance that are dependent on the center voxel.

4.4 Gray Level Run Length Matrix (GLRLM) Features

A Gray Level Run Length Matrix (GLRLM) quantifies gray level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same gray level value.

4.5 Gray Level Size Zone Matrix (GLSZM) Features

A Gray Level Size Zone (GLSZM) quantifies gray level zones in an image. A gray level zone is defined as a the number of connected voxels that share the same gray level intensity. A voxel is considered connected if the distance is 1 according to the infinity norm (26-connected region in a 3D, 8-connected region in 2D). In a gray level size zone matrix P(i,j) the (i,j)th element equals the number of zones with gray level i and size j appear in image. Contrary to GLCM and GLRLM, the GLSZM is rotation independent, with only one matrix calculated for all directions in the ROI.

4.6 Neighbouring Gray Tone Difference Matrix (NGTDM) Features

A Neighbouring Gray Tone Difference Matrix quantifies the difference between a gray value and the average gray value of its neighbours within distance. The sum of absolute differences for gray level i is stored in the matrix.

4.7 Shape

In this group of features we included descriptors of the two and three-dimensional size and shape of the ROI. These features are independent from the gray level intensity distribution in the ROI and are therefore only calculated on the non-derived image and mask.

Unless otherwise specified, features are derived from the approximated shape defined by the triangle mesh. To build this mesh, vertices (points) are first defined as points halfway on an edge between a voxel included in the ROI and one outside the ROI. By connecting these vertices a mesh of connected triangles is obtained, with each triangle defined by 3 adjacent vertices, which shares each side with exactly one other triangle.

References

[1] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107. ‘https://doi.org/10.1158/0008-5472.CAN-17-0339 ¡https://doi.org/10.1158/0008-5472.CAN-17-0339¿‘