The LUNA16 Dataset Description

Guillermo Torres and Debora Gil

Contents

The LIDC-IDRI dataset [1] comprises over 1000 low-dose computed tomography (CT) scans with pulmonary nodules, along with expert radiologists' annotations for the nodules in DICOM and XML formats. To enhance access to the scans and annotations and provide a curated dataset, the LUNA16 dataset [2] was created as a subset of the LIDC-IDRI dataset. LUNA16 was originally utilized for a pulmonary nodule detection challenge [3] and has since been widely adopted for research studies related to nodule detection, segmentation, and classification.

1 Metadata

Nodule's annotations

The LUNA16 dataset includes CT scans from patients, with each patient identified anonymously by an ID, such as LIDC-IDRI-0001, LIDC-IDRI-0003, and so on. Additionally, each scan has a unique series ID number. For instance, the patient LIDC-IDRI-0001 has a CT scan with the series ID 1.3.6.1.4.1.14519.5.2.1.6279.6001.179049373636438705059720603192.

The annotation process followed a two-phase reading approach. During the first phase, each radiologist reviewed the CT scan individually, identifying and annotating all nodules without any knowledge of the other radiologists' annotations. In the second phase, the radiologists reviewed their own annotations in comparison to the annotations made by the other radiologists, and had the opportunity to make changes to their annotations based on this comparison. In their annotations, the radiologists subjectively characterized several properties of each nodule using the following set of features [4, 5, 6]:

- **Location**: The coordinate (x, y, z) in millimeters representing the central location of the nodule in the CT scan.
- **Diameter**: Nodule's diameter measured in millimeters.
- **Subtlety**: Difficulty of detection. Higher values indicate easier detection.
- **Internal structure**: Internal composition of the nodule.
- **Calcification**: Pattern of calcification, if present.
- **Sphericity**: The three-dimensional shape of the nodule in terms of its roundness.
- **Margin**: Description of how well-defined the nodule margin is.
- **Lobulation**: The degree of lobulation ranging from none to marked.
- **Spiculation**: The extent of spiculation present.
- **Texture**: Radiographic solidity: internal texture (solid, ground glass, or mixed).
- **Malignancy**: Subjective assessment of malignancy made by visual inspection of lesion and clinical factors.

For the aforementioned features, Table 1 displays the possible descriptions along with their corresponding values.

To determine whether a nodule is malignant or benign, only the «Malignancy» feature is taken into consideration. Given that a nodule have multiple annotations from different

Table 1: Values and description of the nodule' annotations.

Feature	Range	Value and description
Subtlety	1-5	1. 'Extremely Subtle' 2. 'Moderately Subtle' 3. 'Fairly Subtle' 4. 'Moderately Obvious' 5. 'Obvious'
Internal Structure	1-4	1. 'Soft Tissue' 2. 'Fluid' 3. 'Fat' 4. 'Air'
Calcification	1-6	1. 'Popcorn' 2. 'Laminated' 3. 'Solid' 4. 'Non-central' 5. 'Central' 6. 'Absent'
Sphericity	1-5	1. 'Linear' 2. 'Ovoid/Linear' 3. 'Ovoid' 4. 'Ovoid/Round' 5. 'Round'
Margin	1-5	1. 'Poorly Defined' 2. 'Near Poorly Defined' 3. 'Medium Margin' 4. 'Near Sharp' 5. 'Sharp'
Lobulation	1-5	1. 'No Lobulation' 2. 'Nearly No Lobulation' 3. 'Medium Lobulation' 4. 'Near Marked Lobulation' 5. 'Marked Lobulation'
Spiculation	1-5	1. 'No Spiculation' 2. 'Nearly No Spiculation' 3. 'Medium Spiculation' 4. 'Near Marked Spiculation' 5. 'Marked Spiculation'
Texture	1-5	1. 'Non-Solid/GGO' 2. 'Non-Solid/Mixed' 3. 'Part Solid/Mixed' 4. 'Solid/Mixed' 5. 'Solid'
Malignancy	1-5	1. 'Highly Unlikely' 2. 'Moderately Unlikely' 3. 'Indeterminate' 4. 'Moderately Suspicious' 5. 'Highly Suspicious'

radiologists, a criterion is needed to determine its diagnosis. Specifically, the **nodule's diagnosis criteria** is as follows: if at least two radiologists have annotated the nodule with a value greater than 3 (see Table 1 to match descriptions with their values) in the «Malignancy» feature, it is classified as malignant. Otherwise, it is considered benign.

To illustrate the application of the diagnostic criteria, we can use the «MetadatabyAnnotation.xlsx» file, which contains the radiologist's annotations for each nodule. For instance, consider the case of nodule 1 of patient LIDC-IDRI-0001, which has been characterized by four radiologists (annotation_id of 84, 85, 86, and 87). The radiologists have annotated the «Malignancy» feature of this nodule with the descriptions ['Highly Suspicious', 'Highly Suspicious', 'Highly Suspicious', 'Moderately Suspicious'], which correspond to the values [5, 5, 5, 4] (refer to Table 1 to match descriptions with their values). According to the diagnostic criteria, this nodule is classified as malignant because at least two radiologists assigned a value greater than 3 to the «Malignancy» feature. Similarly, for the case of nodule 1 of patient LIDC-IDRI-0004, the «Malignancy» feature was characterized by four radiologists using the descriptions ['Moderately Unlikely', 'Highly Unlikely', 'Highly Unlikely', 'Highly Unlikely'], which correspond to [2, 1, 1, 1]. This resulted in its classification as benign, according to the diagnostic criteria.

Additionally, the «MetadatabyAnnotation.xlsx» file includes the coordinates of the central location of the nodule (coordX, coordY, and coordZ in mm) and the diameter of the nodule. These values are used to calculate the bounding box of the nodule, which represents the spatial extent of the nodule in three dimensions. The bounding box is defined by the range of values over the X, Y, and Z axes, and it is represented by the parameters bboxLowX, bboxHighX, bboxLowY, bboxHighY, bboxLowZ, and bboxHighZ (in mm). This information can be particularly useful for tasks such as nodule localization and segmentation.

CT Scan's acquisition parameters

During a CT scan, various acquisition parameters can be adjusted according to the clinical requirements and specific patient needs. These acquisition parameters include:

- **Slice thickness**: The thickness of the cross-sectional slices of the body that are imaged. Thinner slices produce higher resolution images, but also increase the amount of radiation exposure to the patient.
- **Window center**: The midpoint of the range of CT numbers displayed in the image, typically measured in Hounsfield units (HU).
- **Window width**: The range of CT numbers displayed in the image, typically measured in HU.
- **Rescale slope**: A scaling factor applied to the raw CT numbers to convert them to Hounsfield units.
- **Rescale intercept**: An offset applied to the raw CT numbers to convert them to Hounsfield units.
- **kVp**: This is the peak voltage of the X-ray beam used in the scan. Increasing the kVp can increase the contrast resolution of the images.

Before we proceed, let us define what a voxel is. A voxel is a fundamental unit of a 3D volume that represents a value at a specific point in 3D space. It can be thought of as the 3D equivalent of a pixel in 2D images. A voxel is typically represented as a cube or a rectangular prism, with height, width, and depth, as shown in Figure 1.

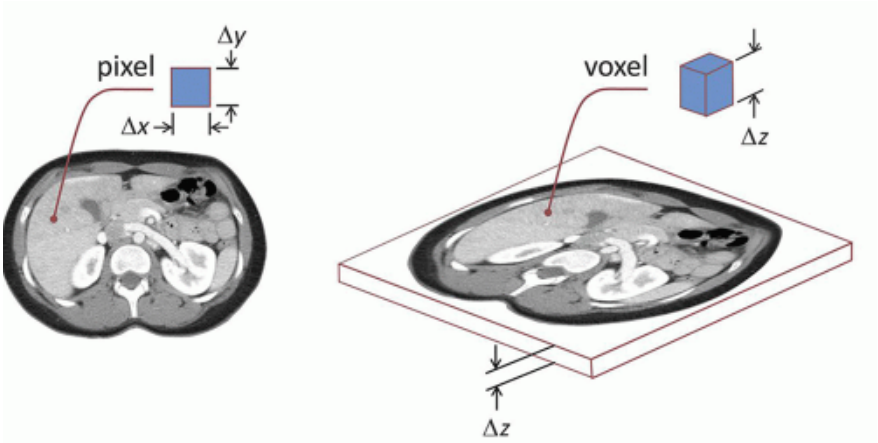

Figure 1: This image shows an axial cut of a slice from a CT scan. On the left side depicts a magnified view of a single pixel. On the right side, the same pixel is extended in a third dimension, creating a voxel, which is the smallest unit of a 3D image. (Original source: https://radiologykey.com/computed-tomography-15. Accessed date: 1 April 2023).

In the context of medical imaging, such as CT scans, a voxel represents an intensity value that is proportional to the signal intensity of the corresponding volume of tissue. The intensity value is usually measured in Hounsfield Units (HU) and is derived from the x-ray attenuation coefficient of the tissue. This value helps in distinguishing between different types of tissues, such as air, water, bone, and soft tissue. The smaller the size of the voxel, the better the quality of the image and the more accurate the representation of the underlying anatomy.

The «AcqParams.xlsx» file provides the acquisition parameters used for the CT scans. One of the important parameters is the voxel dimension, which is determined by the PixelSpacing[0], PixelSpacing[1], and SliceThickness values. The PixelSpacing[0] and PixelSpacing[1] represent the in-plane distance between adjacent voxels (i.e., the distance between voxels in the x and y directions), while the SliceThickness represents the distance between slices (i.e., the distance between voxels in the z direction). Accurate knowledge of the voxel dimension is essential for proper interpretation of medical images and for the accurate measurement of features such as nodule size and volume.

2 CT scans and VOIs

The pulmonary nodules' location information provided by the radiologists was used to extract and save them in a separate file. The CT scan and the corresponding volume of interest (VOI) containing the nodule are saved in NIFTI format, along with their corresponding binary masks. This facilitates the visualization and manipulation of nodule data, as well as integration with various machine learning algorithms.

For instance, each patient's CT scan and its corresponding mask were saved in compressed NIFTI format with the extension «.nii.gz» in the directories CT/image and CT/nodule_mask, respectively. For example, patient LIDC-IDRI-0003 has three nodules that were saved as LIDC-IDRI-0003_R_2.nii.gz, LIDC-IDRI-0003_R_3.nii.gz, and LIDC-IDRI-0003_R_4.nii.gz in the same directory pattern as before.

References

[1] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[2] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[3] grand challenge.org. Luna16 challenge. https://luna16.grand-challenge.org/Description. Accessed: 01-04-2023.

[4] Matt Hancock. Pylidc. https://pylidc.github.io/annotation.html. Accessed: 01-04-2023.

[5] Andrey Fedorov, Matthew Hancock, David Clunie, Mathias Brochhausen, Jonathan Bona, Justin Kirby, John Freymann, Steve Pieper, Hugo Aerts, Ron Kikinis, et al. Standardized representation of the lidc annotations using dicom. Technical report, PeerJ Preprints, 2019.

[6] Matthew C. Hancock and Jerry F. Magnan. Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods. *Journal of Medical Imaging*, 3(4):044504, 2016.