# Preprocessing of Pixel Data Information in the DICOM files of the LIDC-IDRI Dataset

Miguel A. Lerma

December 14, 2023

**Abstract**

This document is intended as a brief reference about how the information from the LIDC-IDRI dataset is stored in DICOM format, and how to preprocess pixel data from DICOM files for use in machine learning applications.

## The LIDC-IDRI dataset

The Lung Image Database Consortium image collection (LIDC-IDRI) consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions [1]. Further information about the LIDC-IDRI dataset can be found in the Cancer Imaging Archive website:

`https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254`

## The Hounsfield Scale

CT scans provide information about radiodensity of tissues, i.e., how much the radiation is attenuated by the different materials found in the body. The Hounsfield Unit (HU) scale (or CT number) is a quantitative scale for expressing density of materials. In this scale distilled water has a density of 0 HU, and air has $-1000$ HU. The following table shows the density in HU for other substances:

| **Substance** | **HU** |
|---|---|
| Air | $-1000$ |
| Water | 0 |
| Lung Parenchyma | $-700$ to $-600$ |
| Fat | $-120$ to $-90$ |
| Blood | $+13$ to $+50$ |
| Bone | $+300$ to $+1900$ |

## HU windows

When looking at anatomical structures in a CT scan the HU values are clipped with an appropriate HU window that depends on what is being examined. In Figure 1 two different windows (left and right of the figure) are being shown—the image in the center is an schematic of the lung structures found. The HU windows used are as indicated in the following table:

| Type of Tissue | HU window | Density Interval |
|---|---|---|
| soft-tissue | center: 50 width: 350 | −125 to +225 |
| lung | center: −200 width: 2000 | −1200 to +800 |

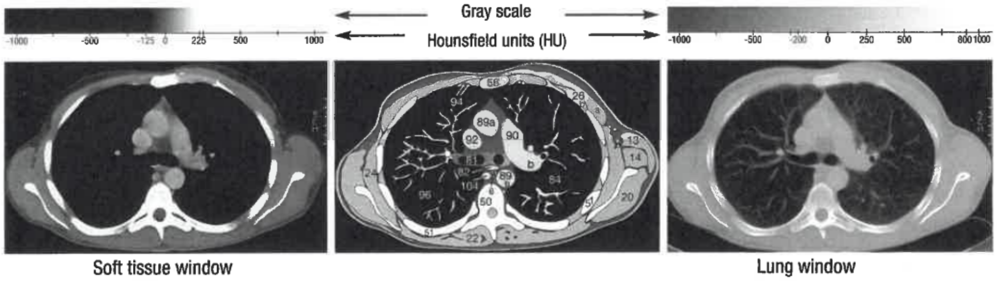

Figure 1: HU windows. The window on the left is intended to observe soft tissue, and the one on the right is the recommended window when searching for lung nodules. In the center there is a schematic of the structures found. Image credit Matthias Hofer.

When examining lung parenchyma scanning for nodules Hofer's CT Scan Manual [2] recommends the *lung window*, with center in −200 HU and width 2000 HU (−1200 to +800).

## The DICOM Format

DICOM files contain a number of fields with information about the CT scans. Here we will focus on the following fields, useful to generate data for processing by machine learning systems for the LIDC-IDRI dataset:

1. *Rows*: Number of rows of the CT scan image (for the LIDC-IDRI dataset it is always 512).

2. *Columns*: Number of columns of the CT scan image (for the LIDC-IDRI dataset it is always 512).

3. *PixelData*: This is a long string with the pixel density data stored as a sequence of 2-byte elements (the order of pixels encoded for each image plane is left to right, top

to bottom). Each element can be interpreted as a signed or unsigned 2-byte integer depending on the value of *PixelRepresentation*.

4. *PixelRepresentation*: If its value is 0 then *PixelData* is to be interpreted as a sequence of unsigned integers (`uint16`), if 1 then the pixel data must be treated as signed integers (`int16`).

5. *RescaleIntercept*: Used to convert pixel data to actual Hounsfield units. Depending on the manufacturer of the CT scanner its value could be 0, −1000, or −1024.

6. *RescaleSlope*: Used to convert pixel data to Hounsfield units. For all samples from the LIDC-IDRI it takes value 1.

7. *PixelPaddingValue*: Used for the value of data pixels that lie in the background outside the CT scan area. This value is typically a very negative value such as −2000 if pixel values are given as signed integers, and 0 if they are unsigned integers. Note that this field may be left undefined in the DICOM file.

So, to get the pixel information from a DICOM file we must read the *Rows* and *Columns* fields to determine the size of the array, then *PixelRepresentation* to determine how to interpret the pixel data (int16 or uint16), then *RescaleIntercept* and *RescaleSlope*, and finally use the following formula to convert the data to Hounsfield units:

$$HU = PixelData * RescaleSlope + RescaleIntercept \tag{1}$$

## Pixel Data Preprocessing and Normalization

The pixel data obtained after reading DICOM files is expressed in different scales and needs to be normalized to make it suitable for display or to be fed to a machine learning application—see e.g. [3] for an exhaustive description of all the steps in preparing CT imaging datasets for deep learning in lung nodule analysis. The process, as stated in [3] section 2.2.4.3.1., starts with the following steps:

1. Transform the pixel data values in the DICOM files to the Hounsfield scale as described in eq. (1) above.

2. Apply an appropriate (clipping) lung window such as the one recommended in the CT Scan Manual.

The result can be further normalized to usual ranges for images or tensors: 0 to 255 (integers), 0.0 to 1.0 (float), −1.0 to 1.0 for PyThorch tensors, etc.

A possible alternative is to simply apply *min-max* normalization to the original pixel data, however, although simpler, it has the following drawbacks:

1. It may distort the apparent density of the tissues in the CT scan, so that areas in two different images with the same final intensities after normalization may correspond to areas with different densities in the original CT scans.
2. The maximum and minimum of a distribution are highly sensitive to outliers, so noise in the CT scan may yield inconsistent image representations (see Figure 5 for an example).
3. The pixel padding values may distort the density information even more since padding values are typically outside all possible ranges of densities of actual substances (see Figure 2 for an illustrative example).

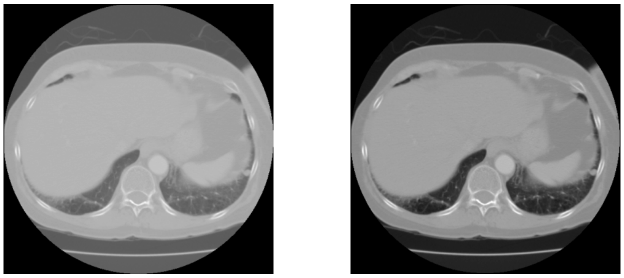

Figure 2: Left: CT scan with nodule number 930 after min-max normalization of the original pixel data. Right: Same CT scan after converting pixel data to HU and clipping with the lung window recommended by the CT Scan Manual.

Figure 2 illustrates the problem of using min-max normalization on the original pixel data (left image). The minimum in this case is the pixel padding value assigned to the background outside the actual CT scan area. As a result the range of values in the CT scan area are compressed within a relatively small interval, reducing the contrast. On the other hand (right image) converting the data to HU and clipping avoids the pixel padding value, improving the contrast.

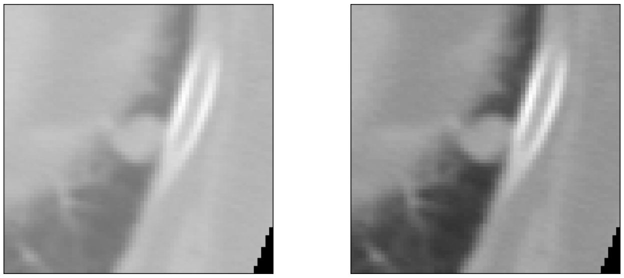

Figure 3: Same as Figure 2 but with a cropping of the image with the nodule in its center. Left: CT scan containing nodule number 930 cropped, and with min-max normalization of the original pixel data. Right: same cropped CT scan after conversion to HU and clipping before converting to image for display.

Figure 3 shows the CT scan for nodule number 930 after cropping the area containing the nodule (in the center). This cropping captures part of the background area filled with padding values (black area in the lower right corner), and that has an impact in the contrast unless clipping is performed. The histogram of the cropped CT scan for nodule number 930 is shown in Figure 4. We notice a short spike on the left area of the histogram, corresponding to pixels outside the CT scan which have been assigned a padding value of −2048. This sets the minimum pixel data value to −2048.

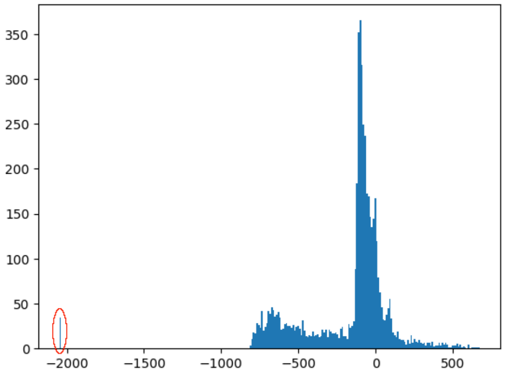

Figure 4: Histogram of original pixel data in the cropped CT scan for nodule number 930. The cropped image contains a few pixels outside the actual CT scan area, which has been assigned the padding value of −2048. Note the short spike at the left of the histogram in the vicinity of −2000. This sets the minimum value of pixel data in the cropped area to −2048 and distorts min-max normalization.

The impact of noise producing outliers should not be underestimated either, e.g. nodule 350 contains just a few (about 14) noisy pixels with extreme values, but that is enough to drastically alter the min-max normalization operation, as shown in Figure 5.

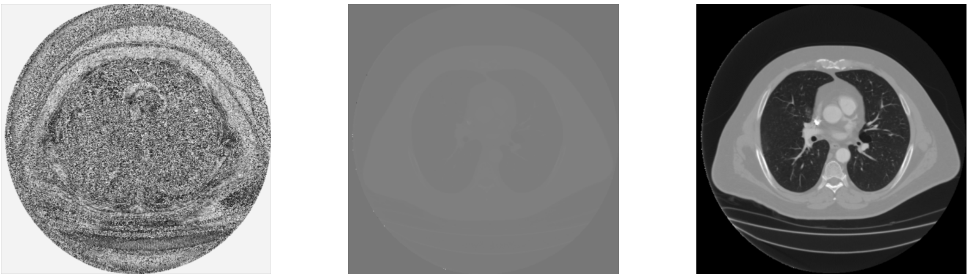

Figure 5: This is an example of the impact of noise in the pixel data values. Left: CT scan containing nodule number 350 with min-max normalization of the original pixel data. The salt and pepper effect is caused by overflow in some intermediate computations with short scalars. Center: Image recomputed using floats to avoid overflowing. Right: same CT scan after conversion to HU and clipping with the recommended lung window before converting to image for display.

### Summary and Conclusions

We have presented a quick review of the main concepts on how pixel data is stored in DICOM files for the LIDC-IDRI dataset, and what to take into account when retrieving it. Also two methods for standardization of pixel data values have been examined: min-max normalization, and conversion to Hounsfield units followed by clipping. Drawbacks of plain min-max normalization have been presented. We note that preprocessing the pixel data by converting to HU and using a clipping window is the method recommended in studies such as [3].

## References

[1] Armato SG 3rd et al. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011 Feb;38(2):915-31. doi: 10.1118/1.3528204. PMID: 21452728; PMCID: PMC3041807.

[2] Matthias Hofer (2007). CT Teaching Manual: A Systematic Approach to CT Reading, 3rd Edition. Institute for Diagnostic Radiology, Heinrich Heine University, Düseldorf, Germany.

[3] Wang J et al. (2023). Preparing CT imaging datasets for deep learning in lung nodule analysis: Insights from four well-known datasets, Heliyon, Volume 9, Issue 6, 2023, e17104, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2023.e17104. (https://www.sciencedirect.com/science/article/pii/S2405844023043128)

## Appendix: Records with Inconsistent Ratings

The LIDC-IDRI documentation includes the following warning:

> For a subset of approximately 100 cases from among the initial 399 cases released, inconsistent rating systems were used among the 5 sites with regard to the spiculation and lobulation characteristics of lesions identified as nodules > 3 mm. The XML nodule characteristics data as it exists for some cases will be impacted by this error. [...]
>
> Contrary to previous documentation (prior to March 2010), the correct ordering for the subjective nodule lobulation and nodule spiculation rating scales stored in the XML files is 1=none to 5=marked. The issue of consistency noted above still remains to be corrected.

Given that it is impossible to determine which of the initial 399 cases contain the wrong annotation all 399 records should be removed in any study involving spiculation or lobulation.

In order to obtain the list of cases with inconsistent ratings, click the "Versions" tab in the following page:

`https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254`

In the notes about Version 1 (2011/06/23) there is a link to a *table which allows mapping between the old NBIA IDs and new TCIA IDs.* That table contains 399 rows that correspond to the cases with the inconsistent ratings.