




CleanX: A Python library for data cleaning of large sets of radiology images

Candace Makeda Moore^{1,3}, Andrew Murphy, MMIS BMedImagingSc RT(R)², and Oleg Sivokon³

¹ MaxCor lab, Sapir College, Israel ² Department of Medical Imaging, Princess Alexandra Hospital, Brisbane, QLD, Australia ³ CMHM/cmhm.info, Israel

DOI: [10.21105/joss.03632](https://doi.org/10.21105/joss.03632)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Christopher R. Madan](#) 

Reviewers:

- [@henrykironde](#)
- [@anki-xyz](#)

Submitted: 31 July 2021

Published: 18 August 2021

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Radiological images of various anatomy are part of the diagnostic work-up for millions of patients for diverse indications. A considerable amount of time and resources have gone into the effort to develop automated diagnostic interpretation of these images. The purpose of this library is to help scientists, medical professionals, and programmers create better datasets upon which algorithms related to X-rays, MRIs or CTs can be based.

CleanX is a Python package for data cleaning that was developed for radiology AI.

Statement of need

CleanX is a Python package for data exploration, cleaning, and augmentation that was originally developed for radiology AI. Python is a widely used language on a global level. Data preparation for building quality machine learning algorithms is known to be a time-consuming task ([Munson, 2012](#)). Of the tasks involved, ‘data cleaning’ alone usually takes the majority of time spent on analysis for clinical research projects ([Wickham, 2014](#)). The task of data cleaning is a necessary step even in the case of relatively high-quality data to avoid the known problem of “garbage in, garbage out” ([Rahm & Do, 2000](#)).

In contemporary research, many approaches to data cleaning for radiology datasets overlook the content of the images themselves. The quality of data, especially the image data, is often context-specific to a specific AI model.

Algorithms that rely on shape detection may be accomplished with contrast and positional invariance, but many neural networks or radiomics algorithms can and should not be insensitive contrast or position. Thus scales like MIDaR ([Harvey H., 2019](#)) are necessary but not sufficient to describe data. Despite the specific nature of quality issues for each model, some potential data contamination problems should be cleaned out of imaging datasets for most algorithms.

In the case of radiological datasets, the task of data cleaning involves checking the accuracy of labelling and/or the quality of the images themselves. Potential problems inside the images themselves in large datasets include the inclusion of “out of domain data” and “label leakage.” Some types of “out of domain data” may not be apparent to non-radiologists and have been a particular problem in datasets web-scraped together by non-radiologists ([Tizhoosh, 2021](#)).

“Label leakage” depends on the desired labels for a dataset but can happen in many ways. More subtle forms of label leakage may occur when certain machines are more likely to be used on certain patients. Depending upon the goals of a model, there may be other types of “out of domain data” that are easy to see, such as inverted or flipped images. Even this can

39 cost tremendous amounts of time to remove from a dataset with hundreds of thousands of
40 images.

41 While data cleaning can not be fully automated at present, it is unrealistic for many data
42 science practitioners and researchers to afford the hours of an imaging specialist for every
43 data cleaning task. This package speeds up data cleaning, and gives researchers some basic
44 insights into datasets of images. It also has functions for augmenting X-ray images so that
45 the resultant images are within domain data.

46 Automated data cleaning can improve dataset quality on some parameters. This work includes
47 open code originally built to help with automatic chest X-ray dataset exploratory data analysis
48 and data cleaning. It was expanded to include functions for DICOM processing, and image
49 data normalization and augmentations. Some of the functions can be used to clean up a
50 dataset of any two dimensional images. Several algorithms for identifying out of domain data
51 in a large dataset of chest-X rays facilitated by the functions in this code library.

52 Acknowledgements

53 We acknowledge many contributions from Eliane Birba (delwende) and Oleg Sivokon (wvxvw)
54 during the testing and documentation of the code related to this project. We did not receive
55 any financial support for this project.

56 References

- 57 Harvey H., G. B. (2019). A standardised approach for preparing imaging data for machine
58 learning tasks in radiology. In A. P. (eds). Ranschaert E. Morozov S. (Ed.), *Artificial*
59 *intelligence in medical imaging*. Springer International Publishing. [https://doi.org/10.](https://doi.org/10.1007/978-3-319-94878-2_6)
60 [1007/978-3-319-94878-2_6](https://doi.org/10.1007/978-3-319-94878-2_6)
- 61 Munson, M. A. (2012). A study on the importance of and time spent on different modeling
62 steps. *SIGKDD Explor. Newsl.*, 13(2), 65–71. <https://doi.org/10.1145/2207243.2207253>
- 63 Rahm, E., & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data*
64 *Eng. Bull.*, 23, 3–13.
- 65 Tizhoosh, F., H. R. (2021). COVID-19, AI enthusiasts, and toy datasets: Radiology without
66 radiologists. *European Radiology*. <https://doi.org/10.1007/s00330-020-07453-w>
- 67 Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59. [https://doi.org/](https://doi.org/10.18637/jss.v059.i10)
68 [10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)