




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

Candace Makeda Moore <sup>1</sup>, Andrew Murphy <sup>2</sup>, Oleg Sivokon<sup>3</sup>, and Patrice J Musoke <sup>4</sup>

<sup>1</sup> Netherlands eScience Center, Amsterdam, Netherlands <sup>2</sup> Department of Medical Imaging, Princess Alexandra Hospital, Brisbane, QLD, Australia <sup>3</sup> Bright Computing / NVIDIA, Netherlands <sup>4</sup> Temple University

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: 01 August 2022

## 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 developing 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 artificial intelligence (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 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](#)). Even in the case of relatively high-quality datasets, the task of ‘cleaning’ is a necessary step, to avoid the problem of poor input leading to poor performance, also known as the “garbage in, garbage out” phenomenon ([Rahm & Do, 2000](#)).

In contemporary research, many approaches to data cleaning for radiology datasets overlook the content of the images themselves. In fact, recent work reveals more open algorithms for image de-identification than for all kinds of image curation ([Diaz et al., 2021](#)) (which may or may not include analysis of image content and quality). However the quality of data, especially the image data, is often context-specific and salient to a particular AI model.

Algorithms that rely on shape detection may be accomplished with contrast, and positional invariance, but specific neural networks or radiomics algorithms should not be insensitive to contrast or position. A neural network designed to detect the technical quality of a chest X-ray should not be positionally invariant, as a the rotated patient is likely to indicate poor radiographic technique, and in the case of a flipped image, would correlate with the clinical picture of situs inversus. 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, data contamination problems should be cleaned out of imaging datasets before building algorithms.

In the case of radiological datasets, the data cleaning task involves checking the accuracy of labelling and/or the quality of the images. Potential problems inside the images in large datasets include “out-of-domain data” and “label leakage”. Certain 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 multiple 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 cost tremendous amounts of time to remove from a dataset with hundreds of thousands of images.

While data cleaning can not be fully automated at present, it is unrealistic for many data science practitioners and researchers to afford the hours of an imaging specialist for every data cleaning task. This package speeds up data cleaning, and gives researchers basic insights into an examined datasets of images. Instead of examining all data by hand or writing bespoke functions for cleaning every specific dataset, the software allows users to decrease the amount of data that needs to be reviewed by hand, explore, and clean data automatically. It also has functions for augmenting X-ray images so that the resultant images are within domain data.

Automated data cleaning and augmentation can improve datasets in terms of quantity, quality and diversity of images and labels. This work includes open code initially built to help with automatic chest X-ray dataset exploratory data analysis and data cleaning. It was expanded to include functions for DICOM processing, image data normalization, and augmentations. The majority of the functions can be used to clean up a dataset of any two-dimensional images; the software has generalizability. Several algorithms for identifying out-of-domain data in a large dataset of chest X-rays are facilitated by the functions in this library.

## Acknowledgements

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

## References

- Diaz, O., Kushibar, K., Osuala, R., Linardos, A., Garrucho, L., Igual, L., Radeva, P., Prior, F., Gkontra, P., & Lekadir, K. (2021). Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Physica Medica*, *83*, 25–37. <https://doi.org/10.1016/j.ejmp.2021.02.007>
- Harvey H., G. B. (2019). A standardised approach for preparing imaging data for machine learning tasks in radiology. In A. P. (eds). Ranschaert E. Morozov S. (Ed.), *Artificial intelligence in medical imaging*. Springer International Publishing. [https://doi.org/10.1007/978-3-319-94878-2\\_6](https://doi.org/10.1007/978-3-319-94878-2_6)
- Munson, M. A. (2012). A study on the importance of and time spent on different modeling steps. *SIGKDD Explor. Newsl.*, *13*(2), 65–71. <https://doi.org/10.1145/2207243.2207253>
- Rahm, E., & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*, 3–13.
- Tizhoosh, F., H. R. (2021). COVID-19, AI enthusiasts, and toy datasets: Radiology without radiologists. *European Radiology*. <https://doi.org/10.1007/s00330-020-07453-w>
- Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, *59*. <https://doi.org/10.18637/jss.v059.i10>