Systematic Detection and Correction of DICOM Label Discrepancies in Large-Scale Chest X-ray Datasets

Frank Li, PhD, Postdoctoral Research Fellow, Radiology, Emory University
Theo Dapamede, MD, PhD; Mohammadreza Chavoshi, MD; Bardia Khosravi, MD, MPH, MHPE; Janice Newsome, MD; Aawez Mansuri, MS; Rohan Satya Isaac, MS; Hari Trivedi, MD; Judy Gichoya, MS, MD, FSIIM

Background/Problem Being Solved

Artificial intelligence models require large, high-quality annotated datasets for development. Manual labeling of radiology datasets is expensive and time-consuming; hence most labels are extracted from existing radiology reports, with the most common tool being CheXpert labeler. Beyond labeling, radiology images also require harmonization and de-identification of DICOM metadata and pixel data. Verifying the quality of the resultant curation process remains extremely challenging due to dataset scale and heterogeneity, and manual review is time-prohibitive. We describe our curation process for our CXR dataset (containing > 2 million images) and highlight common pitfalls and solutions.

Intervention(s)

Many errors were encountered within DICOM files during data curation that would introduce noise and affect downstream analyses (Table 1), including 1) mislabeled DICOM tags for ViewPosition (PA, AP, and lateral); 2) images with Contrast Limited Adaptive Histogram Equalization (CLAHE) without differentiation in DICOM tags; and 3) duplicate images with unique SOP Instance UID within the same study.
To correct the view position, we developed an in-house view position deep learning classifier trained on CheXpert that achieved AUC=1.00 on test data. To identify duplicates and CLAHE images, cosine similarity was computed between image pairs within each study. A cosine similarity of 1 indicated identical images. For images with high similarity (but < 1), image noise was quantified using a high pass filter (Figure 1a), with the noise signal's histogram fitted to a Poisson distribution to obtain the distribution parameter (λ) (Figure 1b). For image pairs exceeding the chosen similarity threshold, images with greater noise (or λ) were classified as CLAHE, based on the assumption that CLAHE processing increases image noise (Figure 1c).

Barriers/Challenges

The large sample sizes of medical imaging datasets make pairwise comparison of images computationally intensive and time-consuming. Parallel computation techniques can significantly accelerate this process by distributing the workload across multiple processors.

Outcome

Our interventions identified 46,512 (1.87%) lateral images incorrectly labeled as AP view in the DICOM tags, 397,384 (15.98%) CLAHE images, and 1,782 (0.07%) duplicate images among the total 2,486,502 images.

Conclusion/Statement of Impact/Lessons Learned

DICOM tags are inaccurate as labels for dataset curation. Computational approaches leveraging deep learning and image processing techniques like noise quantification can improve curation at scale with minimal human input. Such tools can be incorporated in orchestration engines for routing medical images to AI platforms to improve match rate.

Figure(s)

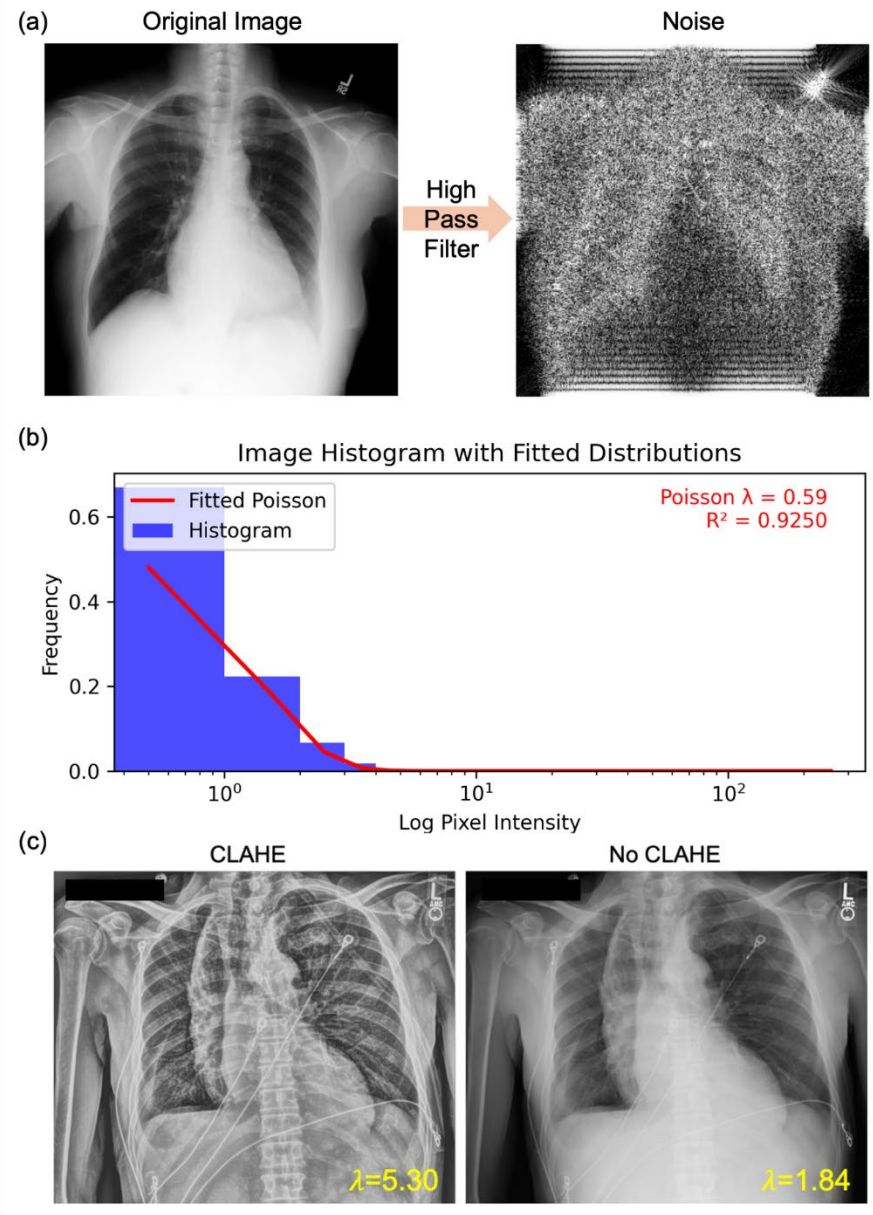

Figure 1. Figures showing (a) noise extracted from a CXR using a high pass filter, (b) The characteristic of the noise (λ) quantified by Poisson curve fitting, and (c) Image pair which has high similarity with different λ. The image with greater λ is considered as a CLAHE image.

DICOM Tags	Description	Values
(0018, 5101) View Position	Radiographic view associated with Patient Position	AP, PA, LL …
(0008, 0018) SOP Instance UID	Unique Identifier for DICOM instance	Dot-separated numbers
(0008, 0008) Image Type	Image identification characteristics	['ORIGINAL', 'PRIMARY'] ['DERIVED', 'SECONDARY'] …

Table 1. Frequently mislabeled DICOM tags

Keywords

Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity