The Impact of Standard Image Preprocessing on Deep Learning Models for Chest Radiographs: An Overlooked Source of Performance Variability

Daniel C. Kargilis,[1,2] Adway Kanhere, [1,2] Sean Garin,[2] Zachary Murphy, MD/MSE,[3] Christopher Hafey,[4] Vishwa S Parekh, PhD,[2] Paul H Yi, MD[2]

[1]Johns Hopkins University School of Medicine and/or Whiting School of Engineering, Baltimore, MD
[2]University of Maryland Medical Intelligent Imaging (UM2ii) Center, Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD
[3]Department of Anesthesiology, University of Michigan, Ann Arbor, MI
[4]Amazon Web Services

BACKGROUND

- Deep learning (DL) models have shown promise in diagnosis on chest radiographs (CXR)
- Image preprocessing is an essential first step in DL pipelines
- It is unclear how commonly-used image preprocessing methods affect performance

STUDY OBJECTIVE

- **To evaluate the impact of three standard types of image preprocessing on DL model performance for CXR diagnosis.**

METHODS

- We assessed for changes in model predictions across 3 datasets, 4 models, and 11 preprocessing techniques (Fig 1).

Techniques:	Algorithm
Resize	Bilinear, interarea, cubic, Lanczos
Histogram Equalization	Contrast Limited Adaptive Histogram Equalization (CLAHE)
Compression	JPEG2000 Lossless, Lossy
Datasets:	**N, CXR**
NIH-CXR14	3206
CheXpert	2684
MIMIC	2540
Models:	**Training Details**
XRV-NIH	TorchXRayVision (XRV) model trained on NIH data
XRV-CHEX	XRV model trained on CheXpert data
XRV-MIMIC	XRV model trained on MIMIC data
CHEX-Official	Official Stanford CheXpert model

METHODS CONT.

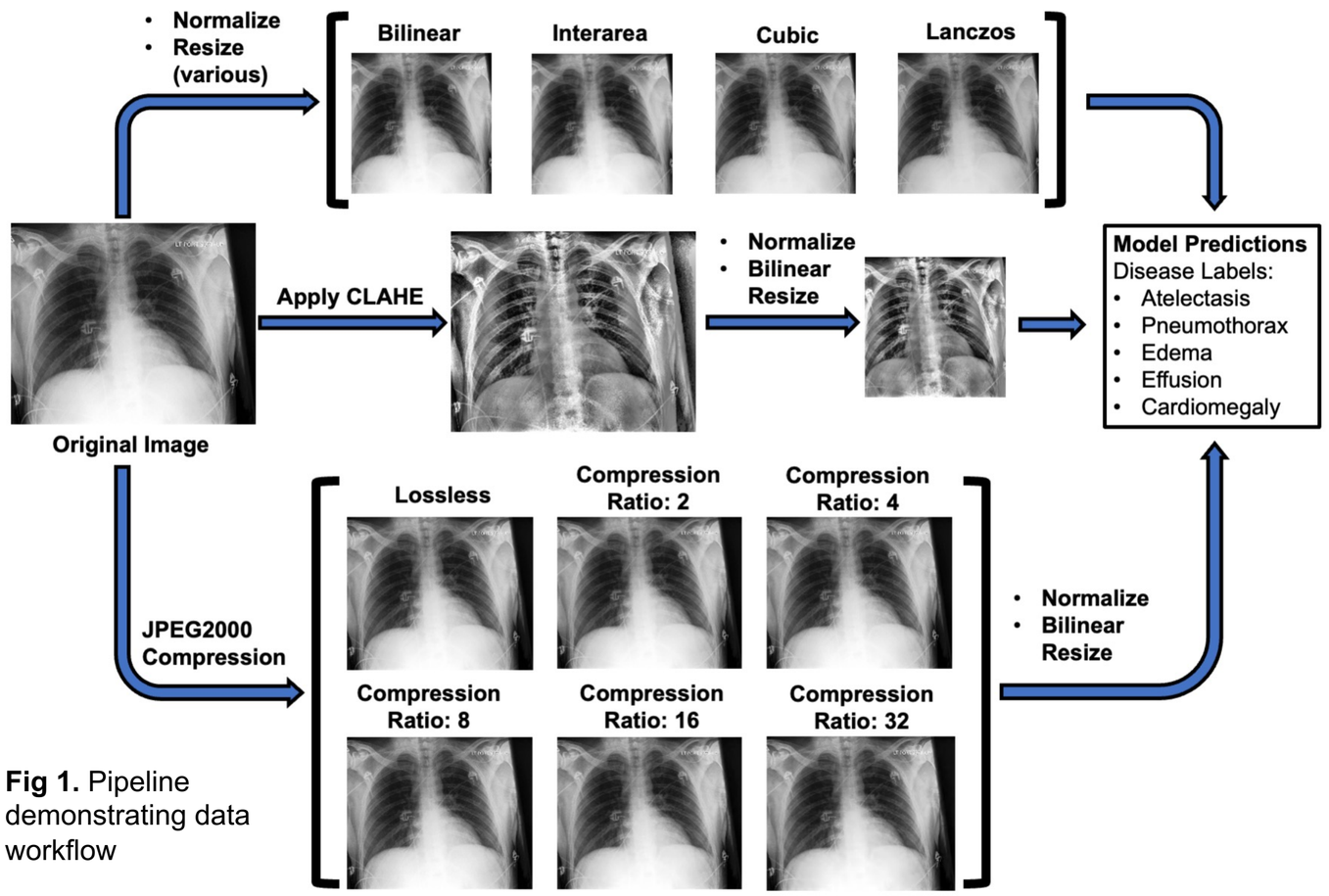

Fig 1. Pipeline demonstrating data workflow

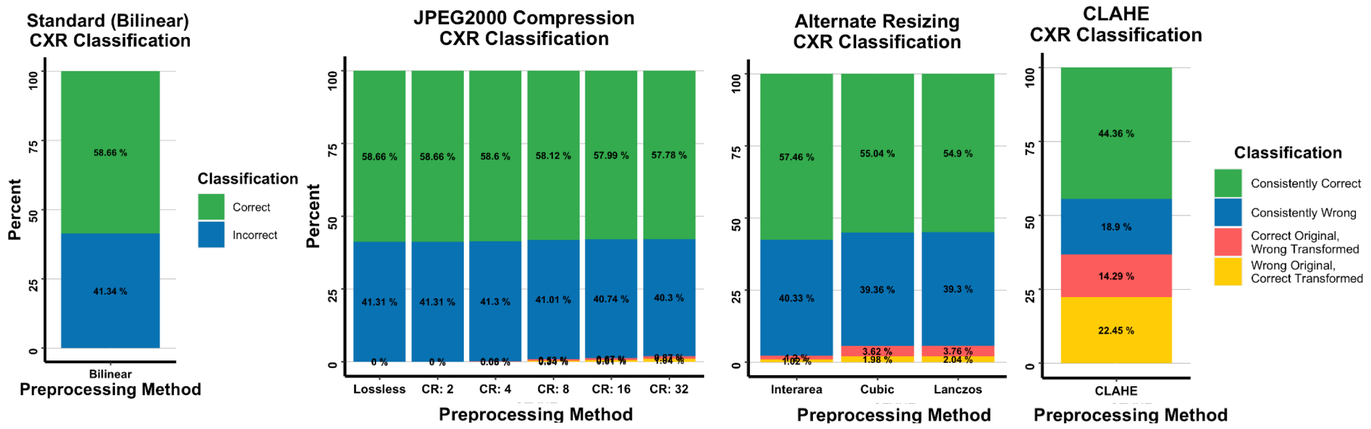

Fig 3. Average Performance in Simulated Deployment. Results shown represent averages across all tests.

RESULTS

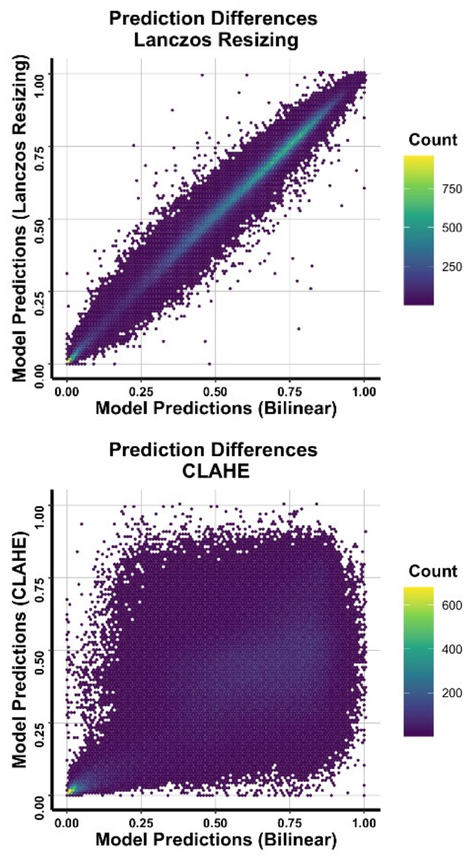

Fig 2. Heatmaps of Preprocessed Model Predictions (Y-axis) versus Standard Predictions (X-axis). Data represents all tests for each designated preprocessing method.

LIMITATIONS

- Limited diversity in model architectures
- Lower performing models may exert a bias
- Upscaling was not assessed in resizing

CONCLUSION & IMPLICATION

- DL models make inconsistent predictions on preprocessed vs. raw CXR images
- Care should be taken when preprocessing images as part of DL pipelines and/or attempts to reduce storage costs.