CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images

Nicolás Gaggion [1] Candelaria Mosquera [2,3] Lucas Mansilla [1]

Julia Mariel Saidman[2] Martina Aineseder[2] Diego H. Milone[1] Enzo Ferrante [1]

[1] Universidad Nacional del Litoral, CONICET, Argentina
[2] Hospital Italiano de Buenos Aires, Argentina
[3] Universidad Tecnológica Nacional, Argentina
eferrante@sinc.unl.edu.ar

Keywords: artificial intelligence, chest X-ray analysis, dataset.

The development of successful artificial intelligence models for chest X-ray analysis relies on large, diverse datasets with high-quality annotations. While several databases of chest X-ray images have been released, most include disease diagnosis labels but lack detailed pixel-level anatomical segmentation labels. To address this gap, we introduce an extensive chest X-ray multi-center segmentation dataset with uniform and fine-grain anatomical annotations for images coming from five well-known publicly available databases: ChestX-ray8, CheXpert, MIMIC-CXR-JPG, Padchest, and VinDr-CXR, resulting in 657,566 segmentation masks. Our methodology utilizes the HybridGNet model to ensure consistent and high-quality segmentations across all datasets. Rigorous validation, including expert physician evaluation and automatic quality control, was conducted to validate the resulting masks. Additionally, we provide individualized quality indices per mask and an overall quality estimation per dataset. This dataset serves as a valuable resource for the broader scientific community, streamlining the development and assessment of innovative methodologies in chest X-ray analysis.

Cite this article:
Gaggion, N., Mosquera, C., Mansilla, L. et al. CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. Sci Data 11, 511 (2024). https://doi.org/10.1038/s41597-024-03358-1