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[4], Martina Aineseder[4], Diego H. Milone[1], and Enzo Ferrante[1,*]

[1]Institute for Signals, Systems and Computational Intelligence, sinc(i) CONICET-UNL, Santa Fe, CP 3002, Argentina
[2]Health Informatics Department at Hospital Italiano de Buenos Aires, Buenos Aires, CP 1199, Argentina
[3]Universidad Tecnológica Nacional, Buenos Aires, CP 1179, Argentina
[4]Radiology Department, Hospital Italiano de Buenos Aires, Buenos Aires, CP 1199, Argentina
[*]corresponding author: Enzo Ferrante (eferrante@sinc.unl.edu.ar)

ABSTRACT

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. The CheXmask dataset is publicly available at: https://physionet.org/content/chexmask-cxr-segmentation-data/.

Background & Summary

Chest radiography is a pivotal imaging technique used to diagnose a variety of lung diseases, including pneumonia, tuberculosis, and lung cancer. The significant role of chest X-rays (CXR) in clinical practice is ascribed to their non-invasive nature, relatively low cost, and rapid diagnostic potential. However, the interpretation of these images poses a considerable challenge due to the intricate and overlapping structures within the thoracic cavity, and the subtle manifestations of certain pathological conditions. The high demand for chest radiography and the global shortage of radiologists accentuate the need for efficient and reliable automated analysis systems.

In recent years, methods based on deep learning (DL) have demonstrated exceptional prowess in interpreting medical images, rivaling and occasionally surpassing expert human performance[1,2]. Convolutional neural networks (CNN) have been particularly instrumental in facilitating such computer-aided diagnosis (CADx) systems[3,4]. Nonetheless, the success of these algorithms is closely tethered to the availability of accurately annotated data, with sufficient quantity and diversity, to train the models.

An essential task within this framework is segmentation – the delineation of specific anatomical structures or pathological lesions within an image. In the context of CXR, this might involve the demarcation of anatomical structures such as lungs or heart, or the location of disease abnormalities[5]. Accurate and robust segmentation can serve as a precursor to other downstream tasks, for example providing significant information about the location and size of specific organs or detected abnormalities. However, manual segmentation is a time-consuming process, demanding substantial expertise, and thus, does not scale well to the size of large databases required for DL model training[6].

HybridGNet, a deep learning model for realistic organ contouring, offers a solution for the generation of anatomically plausible CXR segmentations[7,8]. Utilizing a hybrid approach, it combines conventional convolution operations for image encoding with graph generative models for the anatomically-guided delineation of organ contours. The HybridGNet model was initially introduced with a small CXR landmark dataset to demonstrate its efficacy. In this work, we leverage this model to accomplish our main objective: introducing a large-scale segmentation dataset, named CheXmask, which provides anatomical masks with their corresponding quality index, for 5 extensive chest X-ray databases: Chest x-ray8[9], Chexpert[2], MIMIC-CXR-JPG[10], Padchest[11] and VinDr-CXR[12]. These databases collectively represent a wide variety of geographical locations, patient demographics, and disease spectra, enabling the development of a broad, diverse segmentation dataset.

As the original databases lack manually curated ground-truth segmentations, we perform quality control by implementing our own Reverse Classification Accuracy (RCA) framework[13]. RCA allows to estimate the accuracy of a segmentation method for an individual image with no ground-truth (GT) masks, which is particularly valuable for large-scale image analysis studies like ours. The fundamental concept behind RCA involves training an auxiliary model (known as the reverse classifier) solely on the individual image, using its predicted segmentation as pseudo-GT. This model is then evaluated on a reference database that contains GT data to obtain a performance metric, which is expected to correlate with the performance that would be measured for the individual image if its GT was available. We validated this method by comparing it to traditional performance evaluation on a subset of test images with masks manually segmented by an expert physician. Additionally, since large-public CXR databases built from automatic analysis of electronic health records (EHR) are subject to errors both in image selection and image annotation, we found that RCA is a useful tool to detect out-of-distribution samples (e.g. poor-quality images). Thus, the RCA metrics for HybridGNet segmentations stand out as a powerful quality metric to handle large databases for downstream tasks, by detecting not only low quality segmentation masks, but also images that should be filtered out.

Our comprehensive analysis underscores the capacity of the HybridGNet model to generate high-quality segmentations of lungs and heart structures in CXR, and presents the RCA method as a way to use these segmentations for detection of poor-quality images in large volumes of data. CheXmask dataset provides a key resource to the medical imaging community and represents a significant stride towards democratizing access to diverse, large-scale segmentation datasets, thereby propelling the advancement of automated CXR analysis research.

In this context, it is important to notice that there are limited publicly available chest X-ray datasets with expert segmentation masks, and most studies mainly rely on the Shenzhen Hospital[14], Montgomery County[15], and JSRT[16] datasets. Taken together, they correspond to less than 1000 annotations and the segmented structures are not consistent (for example, while JSRT includes lung, hearth and clavicle annotations, Shenzhen and Montgomery only include lung annotations). Concurrently to our work, larger datasets have recently emerged, like the one released by Seibold and co-workers[17]. However, even though this dataset provides segmentations for a large number of X-ray images featuring different anatomical structures, they are not real X-ray images, since they were generated as digitally reconstructed radiographs, by projecting CT images from the PAX-Ray++ Dataset[18]. As highlighted by the authors, real X-ray images tend to use a harder type of radiation than standard CTs. This fact, together with the contrast agents used in CTs, might lead to a divergence of visual quality between the projected X-rays and real ones[17]. Moreover, they only provide pixel-level masks, while we also provide landmark based segmentations with correspondences. In this context, CheXmask stands out as a unique large scale dataset for real X-ray images, with quality control validated through objective indices.

Methods

Data Preparation

Image datasets

In this study, we utilized five extensive CXR datasets: ChestX-ray8 (available in a dedicated website[19]), CheXpert (available in a dedicated website[20]), MIMIC-CXR-JPG (available at PhysioNet[21]), Padchest (available in a dedicated website[22]), and VinDr-CXR (available at PhysioNet[23]), original sources to download each dataset are provided in the web references. Figure 1 provides an overview of the complete study, which was repeated five times, one for each dataset individually. The ChestX-ray8 dataset[9] contains 112,120 frontal-view X-ray images from 30,805 unique patients, annotated with text-mined fourteen disease image labels. The CheXpert dataset[2] comprises 224,316 chest radiographs collected from Stanford Hospital between October 2002 and July 2017. The MIMIC-CXR-JPG dataset[10] includes 377,110 CXR images from 227,827 imaging studies involving 65,379 patients; these were collected at the Beth Israel Deaconess Medical Center Emergency Department in the United States from 2011 to 2016. The Padchest dataset[11] is composed of 160,861 images from over 67,000 patients, collected from Hospital San Juan in Spain from 2009 to 2017. Lastly, the VinDr-CXR dataset[12] consists of 18,000 manually annotated images gathered from two primary hospitals in Vietnam.

Inclusion-exclusion criteria

Our study incorporated only frontal images, captured in posteroanterior (PA) or anteroposterior (AP) views (no lateral views were included). Specific selection criteria varied among the datasets due to their differing metadata, as summarized in Figure 2.

ChestX-ray8, and VinDr-CXR datasets contain only frontal images, so all images were included. We incorporated images from the Padchest dataset based on the “Projection” metadata field, including only those labeled as PA or AP, reducing the initial count of 160,861 images to 96,287. For the CheXpert dataset, images labeled as “Frontal” in the “Frontal/Lateral” metadata

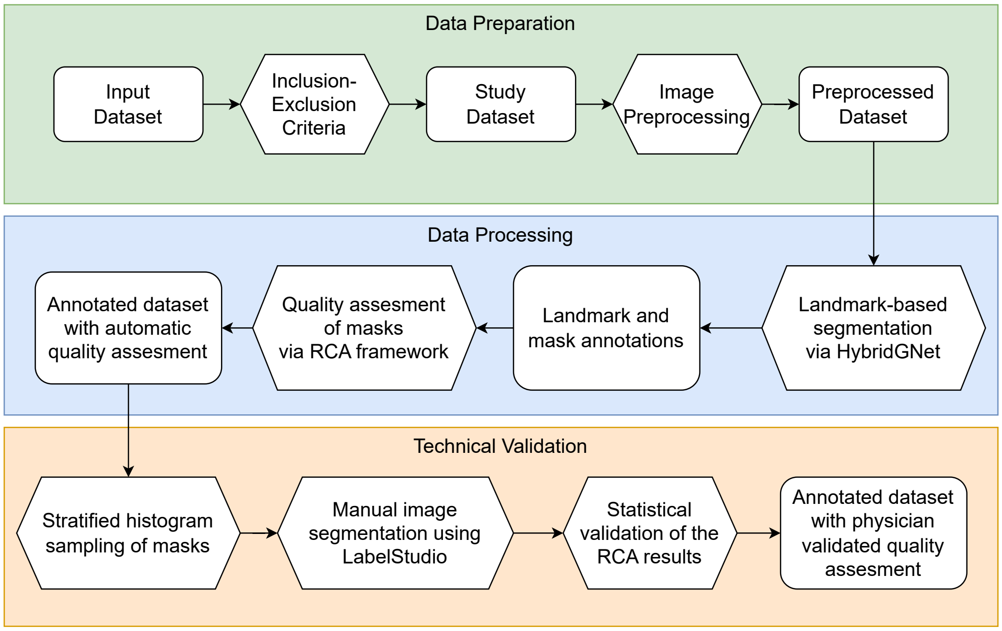
Data Preparation
Input Dataset
Inclusion-Exclusion Criteria
Study Dataset
Image Preprocessing
Preprocessed Dataset
Data Processing
Landmark-based segmentation via HybridGNet
Landmark and mask annotations
Quality assesment of masks via RCA framework
Annotated dataset with automatic quality assesment
Technical Validation
Stratified histogram sampling of masks
Manual image segmentation using LabelStudio
Statistical validation of the RCA results
Annotated dataset with physician validated quality assesment

Figure 1. Data processing flowchart depicting the main steps involved in the building of the CheXmask dataset.

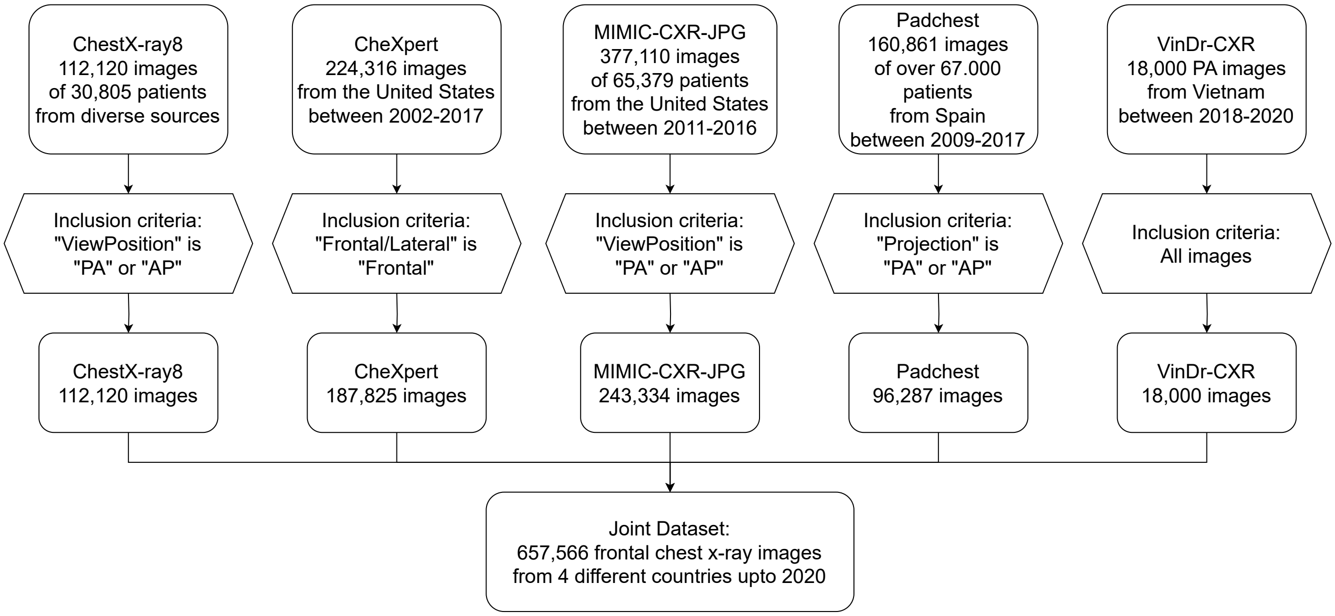
ChestX-ray8 112,120 images of 30,805 patients from diverse sources
CheXpert 224,316 images from the United States between 2002-2017
MIMIC-CXR-JPG 377,110 images of 65,379 patients from the United States between 2011-2016
Padchest 160,861 images of over 67.000 patients from Spain between 2009-2017
VinDr-CXR 18,000 PA images from Vietnam between 2018-2020
Inclusion criteria: "ViewPosition" is "PA" or "AP"
Inclusion criteria: "Frontal/Lateral" is "Frontal"
Inclusion criteria: "ViewPosition" is "PA" or "AP"
Inclusion criteria: "Projection" is "PA" or "AP"
Inclusion criteria: All images
ChestX-ray8 112,120 images
CheXpert 187,825 images
MIMIC-CXR-JPG 243,334 images
Padchest 96,287 images
VinDr-CXR 18,000 images
Joint Dataset: 657,566 frontal chest x-ray images from 4 different countries upto 2020

Figure 2. Summary diagram of inclusion-exclusion criteria.

column were included, resulting in 187,825 out of the total 224,316 images. Lastly, images from the MIMIC-CXR-JPG dataset were included if the "ViewPosition" tag was listed as "PA" or "AP", which resulted in 243.334 out of the initial 377,110 images.

Image preprocessing

For compatibility with HybridGNet input format, we preprocessed images to attain a uniform size suitable for generating pseudo-landmark contours. The standard size of the training data was 1024x1024 pixels, obtained by padding the images to square dimensions then resizing them to the required size, when needed.

Images from ChestX-ray8 dataset already satisfied the required dimension and format, thus requiring no pre-processing. CheXpert, MIMIC-CXR-JPG, Padchest, and VinDr-CXR datasets underwent a standard pre-processing pipeline, which included padding the images to a square shape and then resizing them to 1024x1024 pixels. Additionally, VinDr-CXR images needed extraction from DICOM files before the pre-processing. All these pre-processing procedures are reversible, which allows for the preservation of the original image shapes for releasing the generated annotations.

Data processing

Anatomically plausible segmentation via HybridGNet

The DL model HybridGNet segments organ contours by detecting the coordinates of anatomical landmarks. Thus, it is trained to minimize the distance between the predicted landmark positions and their GT location. The model incorporates an encoder-decoder architecture that combines standard convolutions for image encoding and graph generative models to achieve anatomically plausible representations. Pixel-level masks are obtained by filling in the contours defined by the landmarks predicted by HybridGNet. A detailed description of the HybridGNet model can be found in the work of Gaggion et al.[7,8]. A modified training procedure was presented in a later work[24], to avoid domain memorization issues which emerge when dealing with heterogeneous labels in multi-centric scenarios (i.e. images from some medical centers contain only lung annotations, while others include both lung and heart masks).

In this work, we used the modified multi-centric training procedure[24] to retrain the HybridGNet model with the complete dataset used in prior studies (i.e., both the train and test splits of previous works). This dataset, which will be referred to as Chest-Xray-Landmark dataset, was released previously and is available on GitHub: `https://github.com/ngaggion/Chest-xray-landmark-dataset`. The Chest-Xray-Landmark dataset is curated from a variety of sources, providing annotations for a total of 911 images. These images are sourced from different repositories, including the JSRT dataset[16] (available from its dedicated website[25]), a small subset of the Padchest dataset[11] (accessible on its designated website[22]), as well as the Montgomery[15] and Shenzhen[14] datasets (both obtainable from the NIH National Library of Medicine[26]). Please note that the Chest-Xray-Landmark dataset does not provide the images themselves; users are required to source the original images from their respective origins. All 911 images within this dataset contain lung annotations and only 383 images also include heart labels. The training process incorporated various data augmentation techniques, such as random rotations, scaling, and color shifting, to enhance robustness and generalizability.

A fully working demo of the model used to segment the complete dataset is available on Hugging Face Spaces at `https://huggingface.co/spaces/ngaggion/Chest-x-ray-HybridGNet-Segmentation`.

Reverse classification accuracy to measure segmentation quality

We employed RCA to estimate the Dice Similarity Coefficient (DSC)[27] in the absense of ground truth. DSC is a widely used metric for assessing the similarity between two sets of binary labels. By using RCA, it is possible to produce accurate estimations of DSC, thereby making it easier to perform quality control in the absence of GT. The DSC quantifies the overlap between the predicted segmentation and the GT segmentation, ranging from 0 (no overlap) to 1 (perfect overlap). It is defined as $DSC = \frac{2TP}{2TP+FP+FN}$, where TP, FP, and FN denote the number of true positive, false positive, and false negative pixels, respectively.

To obtain the estimated DSC for a segmentation mask, the RCA framework consists in training a new segmentation method using solely this image with the predicted mask as GT. This new trained model (i.e., the reverse classifier) is then applied to segment a reference image set, with known GT masks. The hypothesis is that the segmentation accuracy of the reverse model on the reference set correlates to the segmentation quality of the original image: if the original segmentation of the evaluated image was good, the segmentation accuracy of the reverse model on this reference set should also be good.

In the original work, the authors experiment with three different methods for implementing the RCA classifier: Atlas Forests, Deep Learning, and Atlas-based Label Propagation. In this case, we used the last one, which employs a non-rigid registration technique for single-atlas label propagation, where the single atlas corresponds to the evaluated image together with its predicted segmentation. A summary value of the distribution of DSC across the reference set images is then used as performance metric. We will use for these values the *mean* and *max* RCA-estimated DSC. The RCA-estimated DSC is expected to correlate well with the real DSC that one would obtain if GT data was available. Figure 3 showcases the distribution of RCA-estimated DSC for the complete CheXmask dataset with examples sampled from each bin.

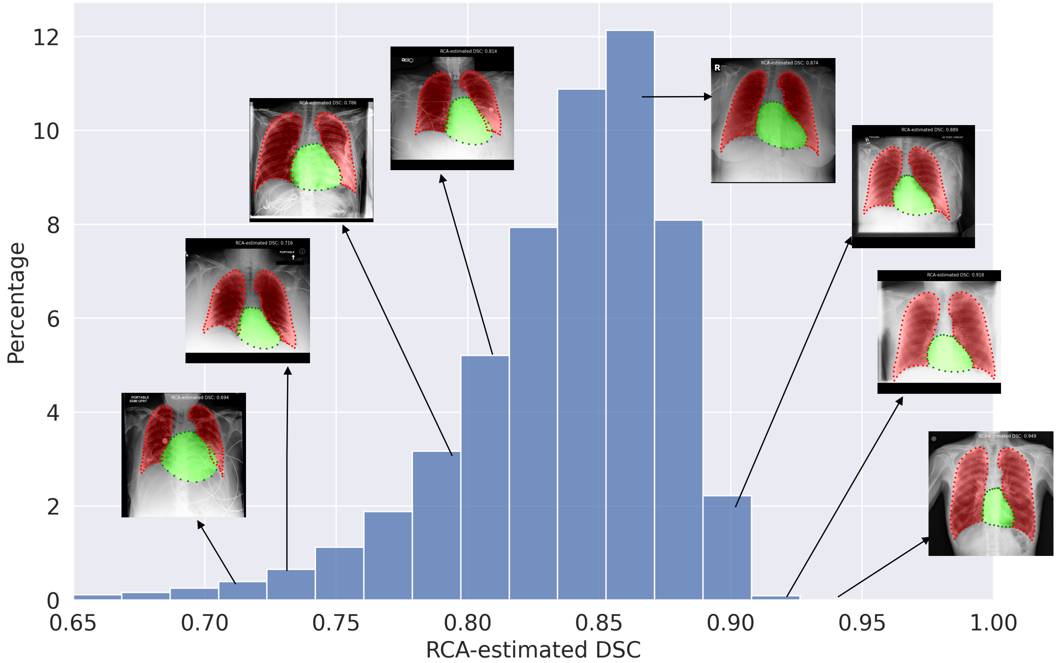

Figure 3. Histogram showcasing the distribution of the RCA-estimated DSC for the complete CheXmask database. Example images with their landmark-based segmentations were drawn with lines to their corresponding histogram bin. Reproduction of x-ray images was allowed upon request to the original sources.

The use of DL in atlas registration has shown improved performance over classic multi-atlas approaches, while also offering a significant speed-up in computation. In order to speed up the evaluation process, we implemented a DL-based atlas registration procedure[28] as label propagation method (that we refer as Deep RCA), which is several orders of magnitude faster than traditional registration methods. Specifically, we divide the registration procedure into two stages: first performing rigid registration to globally align the corresponding images, and then proceeding with deformable registration to correct local misalignments.

Data Records

The CheXmask dataset is structured as one file of comma-separated values (CSV) for each CXR dataset. The CSV content is explained in Table 1. We do not include images or metadata from the original datasets. We include an image ID in the first column of the CSV files, which has the same column name as the ID column of the respective dataset, and allows to match the rows in this CSV with the original dataset's instances. We also include the pre-processed version of the masks, allowing to use all datasets in the same image resolutions (please also note that ChestX-ray8 masks are already in the desired resolution). This database[29] is available at PhysioNet[30], in the following repository: `https://physionet.org/content/chexmask-cxr-segmentation-data/`.

Technical Validation

The technical validation of the CheXmask dataset was conducted through a three-tiered approach: (1) evaluating the quality of HybridGNet segmentations by measuring the DSC on a gold-standard subset revised by an expert physician; (2) validating the relevance of the RCA-estimated DSC as performance metric; (3) evaluating the quality of the masks for the five source datasets separately, by exploring their suitability as a training GT set for the segmentation task.

1) Validation via physician annotations as gold-standard

We built a gold-standard set of masks (for images sampled from the five original datasets) labeled by two experienced physicians, referred as P1 and P2, to evaluate the segmentation quality of CheXmask lungs and heart masks. Both physicians, counting with more than five years of experience, performed a manual revision of the right lung, left lung and heart landmark-based segmentations. We used an open-source labeling platform called LabelStudio to do so. The landmarks predicted by HybridGNet

Table 1. Description of the columns in CheXmask CSV files.

Column	Description
Image ID	Contains references to the original images as per the original metadata, thus the column name changes across dataset.
Dice RCA (Max)	Provides the maximum Dice Similarity Coefficient for the Reverse Classification Accuracy (RCA), indicating the quality of the segmentation.
Dice RCA (Mean)	Provides the mean Dice Similarity Coefficient for the Reverse Classification Accuracy (RCA), indicating the quality of the segmentation.
Landmarks	Includes a set of points representing the contour of the organs, as obtained by the HybridGNet.
Left Lung	Contains segmentation masks of the left lung in run-length encoding (RLE).
Right Lung	Contains segmentation masks of the right lung in RLE.
Heart	Contains segmentation masks of the heart in RLE.
Height	Height of the segmentation mask, necessary to decode RLE.
Width	Width of the segmentation mask, necessary to decode RLE.

for a subset of images were uploaded as predictions to LabelStudio, where the annotator corrected these predicted landmarks by moving them to their correct specific contour positions. This guarantees that the GT masks are based on the same collection of nodes than the evaluated masks, and reduces the time needed for annotation. Figure 4 shows the annotation interface. We include the source code to launch this interface in our repository, as an easy tool to perform further validations. We measured the DSC of HybridGNet's masks comparing with the manually-corrected landmark-annotations, and report statistics on the DSC distribution for the five public datasets.

The sample size of the subset used as gold-standard evaluation was chosen based on the work by El Jurdi and Colliot[31], which suggests sampling 100 images for a 4%-wide confidence interval in segmentation tasks. We sampled 50 images per dataset, obtaining a final gold-standard set of 250 images. We combined random sampling, histogram-based sampling and a random selection of images which included radiological findings (disease or abnormalities) to ensure the presence of easy and challenging cases. For each dataset, ten images were sampled based on a 10-bins histogram of the RCA-estimated DSC of all the dataset's images, taking one random image per bin. Another twenty images were then randomly selected from the whole dataset. Finally, 20 additional images were taken from each dataset, ensuring the presence of pathology or findings such as edema, consolidation, pelural effusion or support devices, which usually result in more challenging segmentation scenarios.

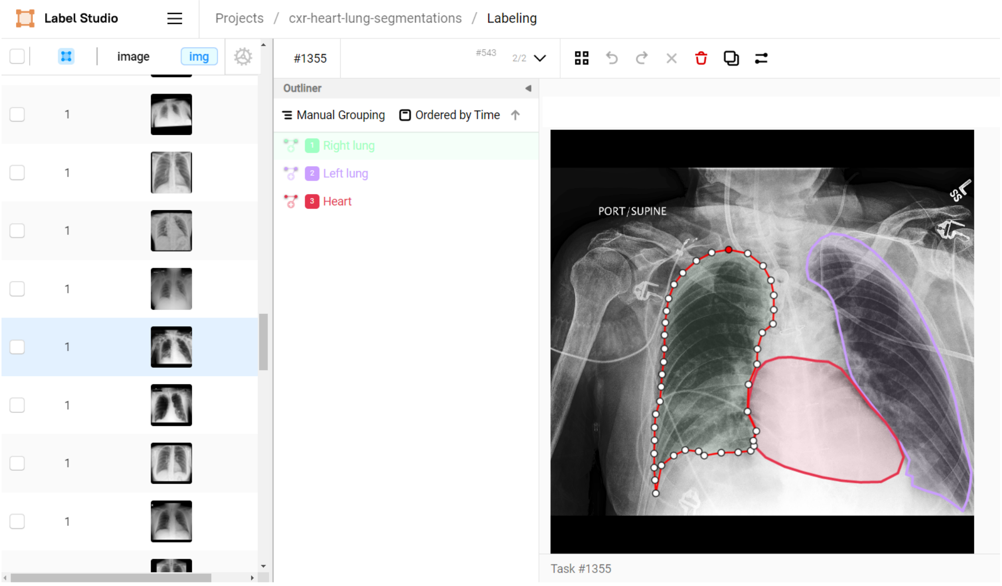

Figure 4. Label Studio setup for manual landmark-based segmentation. Reproduction of x-ray images was allowed upon request to the original sources.

Table 2 summarizes the results of physician validation for lung and heart segmentation across the five datasets on their 50-samples subsets. We report DSC, Hausdorff Distance (HD), and Hausdorff 95% Distance (HD95) as evaluation metrics. HD and HD95 are distance error metrics, providing insights into the difference between the contours of the GT masks and the predicted masks.

In terms of lung segmentation, the datasets generally exhibit high DSC values, ranging from 0.949 to 0.982. This indicates a strong overlap between the GT masks and the predicted lung masks. Among the datasets, VinDr-CXR exhibits the highest agreement for both physicians, while CheXpert has the lowest mean DSC.

For heart segmentation, the DSC ranges differ between the different annotators, in one case from 0.957 to 0.979 and, in the other, from 0.912 to 0.948. In both cases there is a strong overlap between the GT and predicted heart masks across all datasets (heart segmentation is a more difficult task). The HD and HD95 values for heart segmentation are generally lower compared to lung segmentation, indicating a smaller distance between the contours. When comparing between the two expert physicians, the results vary in the same way as with the predictions. More difficult datasets such as CheXpert have the largest inter-reviewer differences, meanwhile VinDr-CXR has both better performance in each reviewer and the lowest inter-reviewer difference.

Overall, the results indicate that the segmentation models perform well in capturing the lung and heart structures. VinDr-CXR consistently demonstrates the highest agreement for both lungs and heart, while CheXpert exhibits relatively lower agreement. These findings highlight the variations in segmentation performance across different datasets and provide valuable insights for further improving the accuracy of the segmentation algorithms.

Per-landmark error analysis

In addition to the segmentation metrics, we also computed the mean squared error (MSE) per landmark across the entire set. This measure provided deeper insights into the positional accuracy of the landmarks predicted by HybridGNet. Interestingly, the results suggested that certain landmarks were more prone to prediction errors than others. Specifically, we observed higher MSE values for landmarks corresponding to the lower part of the lungs and the heart, indicating that these regions are more difficult for the network.

Figure 5 illustrates the MSE per landmark across the whole dataset per expert reviewer. Each landmark is color-coded based on its respective MSE value, using a logarithmic scale for better visualization. The size of each landmark also corresponds to the magnitude of the MSE, with larger landmarks indicating higher errors. As can be seen in this figure, the bottom part of the lungs and the heart landmarks tend to be larger, denoting higher prediction errors. This aligns with the observations made by both expert physicians responsible for the manual annotations, who noted that the boundaries of the heart and the lower section of the lungs tend to exhibit more blurriness compared to the upper part.

Table 2. Results of clinical validation. Comparison of predictions vs P1 and P2 ground truth annotations ($n = 50$ per dataset)

	DSC Lungs	**HD Lungs**	**HD 95 Lungs**	**DSC Heart**	**HD Heart**	**HD 95 Heart**
Pred vs P1						
CheXpert	0.949 (0.067)	49.9 (49.7)	30.0 (38.8)	0.912 (0.083)	40.8 (34.0)	36.4 (32.6)
ChestX-Ray8	0.966 (0.078)	35.4 (53.7)	18.6 (37.6)	0.921 (0.111)	34.8 (45.3)	30.9 (43.1)
MIMIC-CXR-JPG	0.962 (0.055)	37.5 (40.4)	21.1 (30.5)	0.919 (0.093)	36.1 (38.5)	32.5 (36.6)
Padchest	0.965 (0.067)	36.7 (35.9)	17.9 (23.0)	0.948 (0.048)	26.5 (23.1)	22.7 (20.6)
VinDr-CXR	0.982 (0.035)	27.2 (33.0)	10.2 (15.4)	0.912 (0.098)	39.8 (42.8)	35.8 (41.3)
Pred vs P2						
CheXpert	0.958 (0.068)	42.1 (49.4)	24.3 (37.5)	0.966 (0.070)	21.5 (34.7)	16.9 (28.5)
ChestX-Ray8	0.969 (0.077)	33.3 (51.6)	16.6 (35.1)	0.965 (0.105)	18.2 (46.8)	16.2 (44.0)
MIMIC-CXR-JPG	0.962 (0.062)	36.9 (42.6)	18.3 (26.2)	0.945 (0.083)	30.1 (43.6)	26.2 (39.6)
Padchest	0.974 (0.070)	27.2 (38.1)	11.9 (22.6)	0.979 (0.037)	14.4 (26.4)	11.7 (21.9)
VinDr-CXR	0.982 (0.039)	22.5 (35.3)	8.0 (16.0)	0.957 (0.102)	21.1 (47.3)	19.3 (44.5)
P1 vs P2						
CheXpert	0.962 (0.035)	43.9 (55.9)	23.0 (37.8)	0.918 (0.064)	40.4 (29.6)	33.5 (25.0)
ChestX-Ray8	0.976 (0.031)	27.9 (25.5)	12.2 (14.1)	0.942 (0.064)	24.5 (20.4)	21.2 (19.0)
MIMIC-CXR-JPG	0.966 (0.036)	29.0 (21.8)	15.8 (12.9)	0.932 (0.059)	31.5 (22.4)	26.6 (19.9)
Padchest	0.974 (0.027)	30.8 (23.1)	13.3 (12.5)	0.951 (0.038)	25.3 (19.8)	20.8 (16.3)
VinDr-CXR	0.983 (0.024)	22.7 (18.5)	8.1 (9.0)	0.943 (0.037)	25.3 (15.4)	21.6 (14.5)

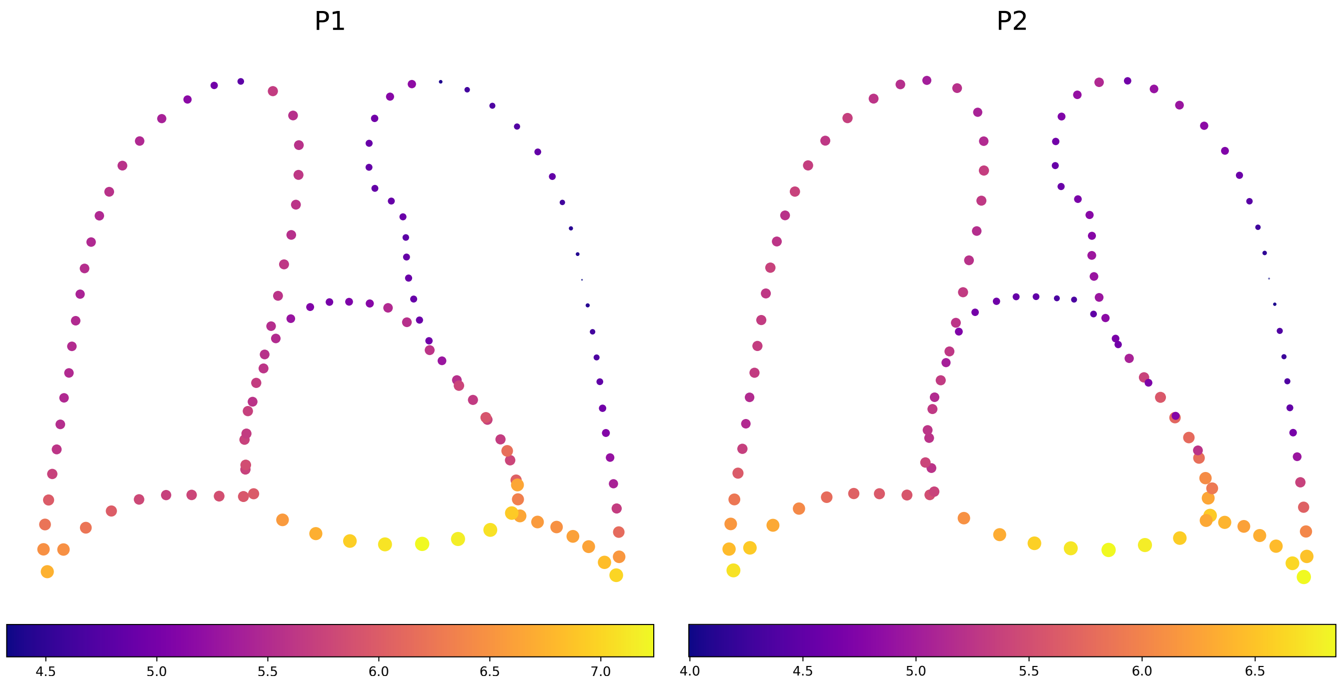

Figure 5. Illustration of the Mean Squared Error (MSE) per landmark across the entire gold-standard set per expert reviewer, depicted in logarithmic scale for improved visual clarity. The color intensity of each landmark represents the magnitude of the MSE. Moreover, the size of each landmark is proportional to its respective MSE, thus larger landmarks indicate a greater prediction error.

2) Validation via RCA-estimated DSC

As previously mentioned, computing the RCA-estimated DSC requires training a rigid and deformable registration model based on deep learning. To this end, we employed 80% of the ChestXray-Landmarks database for training. This model was used to propagate the predicted labels from the image atlas to the reference images (composed of the 5 most similar images to the atlas from the training set), and finally compute the RCA-estimated DSC (mean and max across the 5 values).

To validate that RCA-estimated DSC accurately approximates the real DSC, we proceeded to use simulated segmentation masks of various qualities for which we can compute the real coefficient. To this end, we generated a set of candidate segmentations of various quality (i.e. with real DSC ranging from low to high values), by training 12 UNet models on the same dataset but halting at various epochs (models trained for a few epochs produced lower quality segmentations while those trained until convergence produce better segmentations). Then, we assessed the correlation of RCA-estimated DSC on the remaining 20%test-split of the ChestXray-Landmarks database. The Pearson Correlation scores for the Max and Mean RCA-estimated DSC across the reference set were 0.93 and 0.94 respectively, demonstrating a strong correlation (Figure 6). We found that the RCA-estimated value tends to slightly sub-estimate the true DSC, suggesting the performance measured with RCA follows a conservative approach (i.e., more pessimistic than the actual true performance). We selected the *mean* RCA-estimated DSC as RCA estimation metric for the following experiments since it achieved better correlation.

Evaluation of segmentation quality through RCA

As part of the technical validation on the quality of lungs and heart masks, we report a statistical description of the distribution of mean RCA-estimated DSC for all images in each dataset (Table 3). The full histograms are presented in Figure 7. For all datasets, the mean RCA-estimated DSC is higher than 0.83. As explained above, the true DSC distribution probably has a larger mean, since the RCA-estimated DSC was found to under-estimate the performance.

Demographic bias analysis through RCA

To explore potential biases[32,33]in the masks quality, we conducted a detailed analysis incorporating the metadata associated with the MIMIC-CXR-JPG, Padchest, ChestX-ray8, and CheXpert datasets. We took into account parameters such as disease findings, the sex and age of the patients, and the X-ray capture view (PA or AP).

Figure 8 presents histograms that clearly demonstrate a superior RCA-estimated DSC for PA images compared to AP images across all investigated datasets. This is consistent with the fact that AP images tend to come from hospitalized patients, who are more difficult to position in standard views and usually include artifacts or cables that may act as confounders or

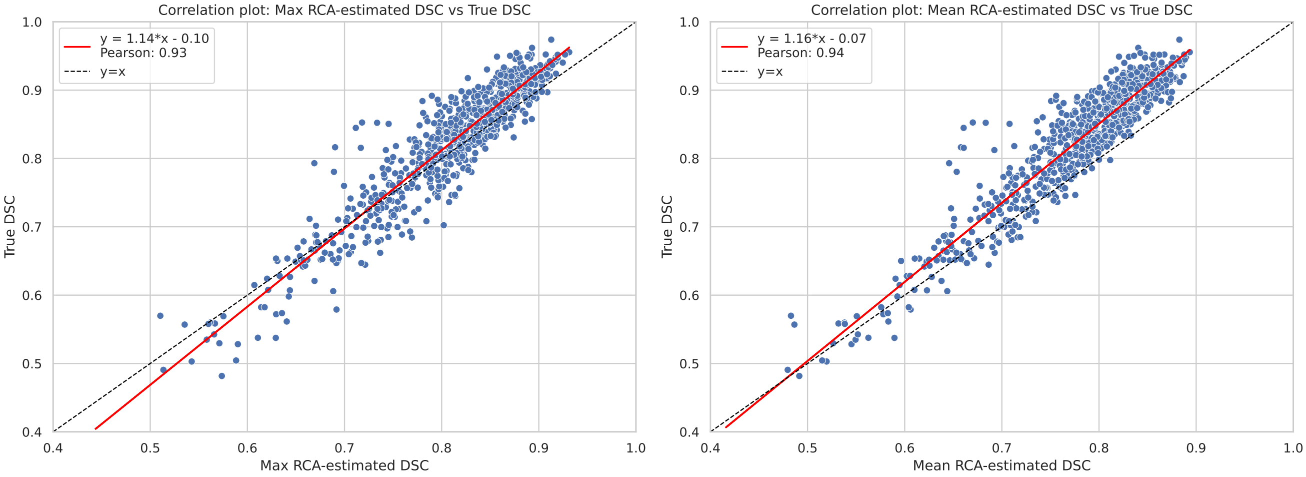

Figure 6. Correlation between the RCA values and the true DSC on the 20% test-subset of the ChestXray-Landmarks database

Table 3. Mean image level RCA-estimated DSC: statistical analysis

Dataset name	Sample size (n)	Mean	Std	Min	1%	5%	25%	50%	75%	Max
Chestx-ray8	112120	0.841	0.044	0.148	0.691	0.764	0.823	0.850	0.869	0.925
CheXpert	187825	0.830	0.038	0.400	0.718	0.761	0.808	0.835	0.857	0.918
MIMIC-CXR-JPG	243334	0.831	0.051	0.149	0.650	0.738	0.811	0.842	0.864	0.921
Padchest	96184	0.851	0.042	0.103	0.672	0.790	0.838	0.858	0.874	0.927
VinDr-CXR	18000	0.850	0.033	0.469	0.738	0.794	0.835	0.855	0.872	0.919

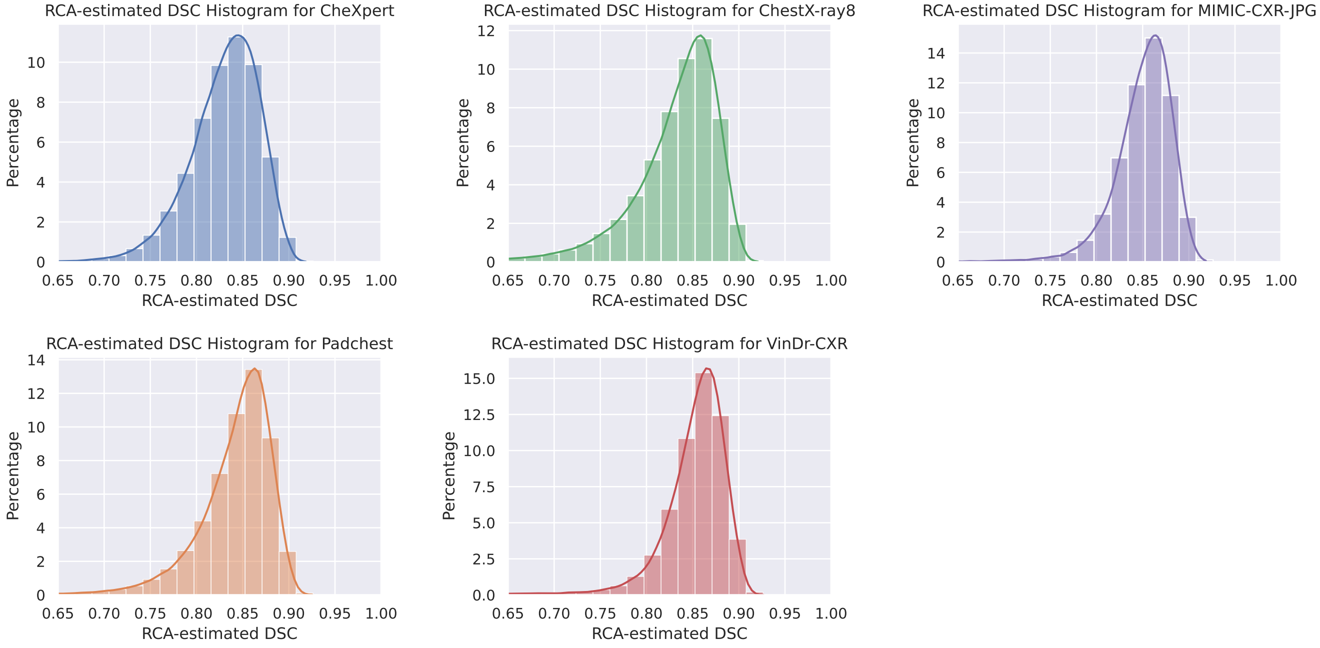

Figure 7. Histograms illustrating the distribution of segmentation quality across the five datasets. Each histogram represents the distribution of RCA (Reverse Classification Accuracy) estimations of DSC for a specific dataset, providing a visual representation of the segmentation performance. The histograms are truncated at 0.65 for visualization purposes. The full range of values is considered in the statistical analysis (Table 3).

Table 4. Dataset-level estimation of the quality of the segmentation via retraining the HybridGNet model from scratch.

	Lungs ($n = 911$)			**Heart** ($n = 383$)		
Training Dataset	**DSC**	**HD**	**HD 95**	**DSC**	**HD**	**HD 95**
Chestx-ray8	0.960 ± 0.019	36.2 ± 22.7	17.7 ± 10.3	0.927 ± 0.031	39.9 ± 16.5	35.2 ± 16.3
CheXpert	0.946 ± 0.028	42.0 ± 23.5	22.9 ± 12.1	0.888 ± 0.053	58.2 ± 27.0	54.0 ± 26.4
Padchest	0.956 ± 0.028	37.7 ± 23.7	19.0 ± 12.3	0.924 ± 0.037	40.9 ± 19.6	36.6 ± 19.2
MIMIC-CXR	0.956 ± 0.020	37.5 ± 22.8	19.2 ± 10.5	0.921 ± 0.031	44.2 ± 19.9	39.6 ± 19.6
VinDr-CXR	0.959 ± 0.023	35.4 ± 23.3	17.7 ± 11.7	0.925 ± 0.035	38.4 ± 16.8	34.0 ± 16.2
20% test split		**Lungs** ($n = 181$)			**Heart** ($n = 76$)	
HybridGNet model trained on 80% train split[24]	0.967 ± 0.017	32.5 ± 20.6	15.6 ± 10.9	0.937 ± 0.029	34.6 ± 14.8	29.4 ± 13.4

occlude body parts, making anatomical segmentation more challenging. In contrast, discrepancies based on sex were less pronounced, with a slightly diminished segmentation quality observed in females. This is also consistent with the fact x-ray imaging of the upper thorax women's breasts tend to occlude the imaged organs, resulting in poorer image contrast for the relevant anatomy[34] and making the segmentation task more challenging. Furthermore, the influence of disease findings on segmentation quality varied amongst datasets, with significant disparities evident in the ChestX-ray8 and CheXpert datasets.

Upon evaluating the relationship between patient age and RCA-estimated DSC in the CheXpert dataset, no substantial patterns emerged. It is important to note, however, that all subjects are above the age of 18. On the contrary, when analyzing results for the ChestX-ray8 dataset which includes patients under 18, we observed lower RCA-estimated DSC for these cases. This is consistent with results reported in a previous analysis of age patterns for the HybridGNet model[8], where subjects below the age of 18 demonstrated subpar performance due to their absence in the training set. We hypothesize a similar trend may be affecting the sparse set of ChestX-ray8 X-rays taken for individuals under the age of 18.

These findings underscore the importance of desaggregated analysis with respect to image acquisition protocol, disease presence and demographic attributes when understanding segmentation quality.

Out-of-distribution detection through RCA

In this section we focus on analyzing those segmentation masks which received lower RCA-estimated DSC values. Through empirical observation, we discovered that RCA-estimated DSC values lower than 0.7 tend to be indicative of out-of-distribution cases. To exemplify this, we randomly selected a subset of six images from each dataset with some of the lowest RCA values, which are showcased in Figure 9. By performing a manual and qualitative analysis of the lowest RCA-estimated DSC masks for each dataset, we found a wide variety of challenging scenarios, ranging from images with high noise levels or mislabeled lateral CXR, to an X-ray image of a mobile phone from the Padchest dataset. That is why we recommend users of the CheXmask dataset to only use segmentations whose RCA-estimated DSC is higher than 0.7.

Table 3 can be useful to estimate the prevalence of such out-of-distribution images in each dataset. Notably, the VinDr-CXR dataset does not include any non-posteroanterior (PA) images nor many out-of-distribution images, as all images were manually annotated by physicians. However, the inclusion of zoomed-out images capturing both arms, which were absent in the HybridGNet training set, leads to lower-quality segmentations in those cases. In contrast, the larger datasets contain a higher number of cases mislabeled as PA or AP in their metadata. These mislabeled images could be easily identified and discarded by thresholding the RCA-estimated DSC, for example using the 1% or 5% quartile of a dataset as threshold. This approach can help to improve the overall quality and reliability of downstream applications developed using this datasets.

3) Validation via retraining segmentation models with CheXmask

We performed a final quality control at the dataset-level to evaluate the quality of HybridGNet segmentations per dataset (not individually as we did with RCA-estimated DSC). The hypothesis is that if the segmentations are good, they can be used as training set for a segmentation model that will obtain high performance. Thus, we trained five new HybridGNet models, by using the set of predicted masks of each of the five public datasets as training GT. We then evaluated model performance on the complete ChestXray-Landmarks database. All models were trained using the same configurations and hyperparameters as the original HybridGNet model[24]. As a comparison reference, we include the performance metrics reported on prior studies for the 20% test split of the ChestXray-Landmarks database which can be seen as an upper bound in performance, as this model was trained with the training split of the ChestXray-Landmarks database.

Table 4 shows the results for lungs and heart segmentation. A slight decrease is observed in terms of DSC, decreasing from 0.967 to 0.955 when using the segmented datasets as training set, in comparison to the reference values. This indicates that the overall quality of the segmentations is good enough to train models which perform well when compared with existing expert annotations.

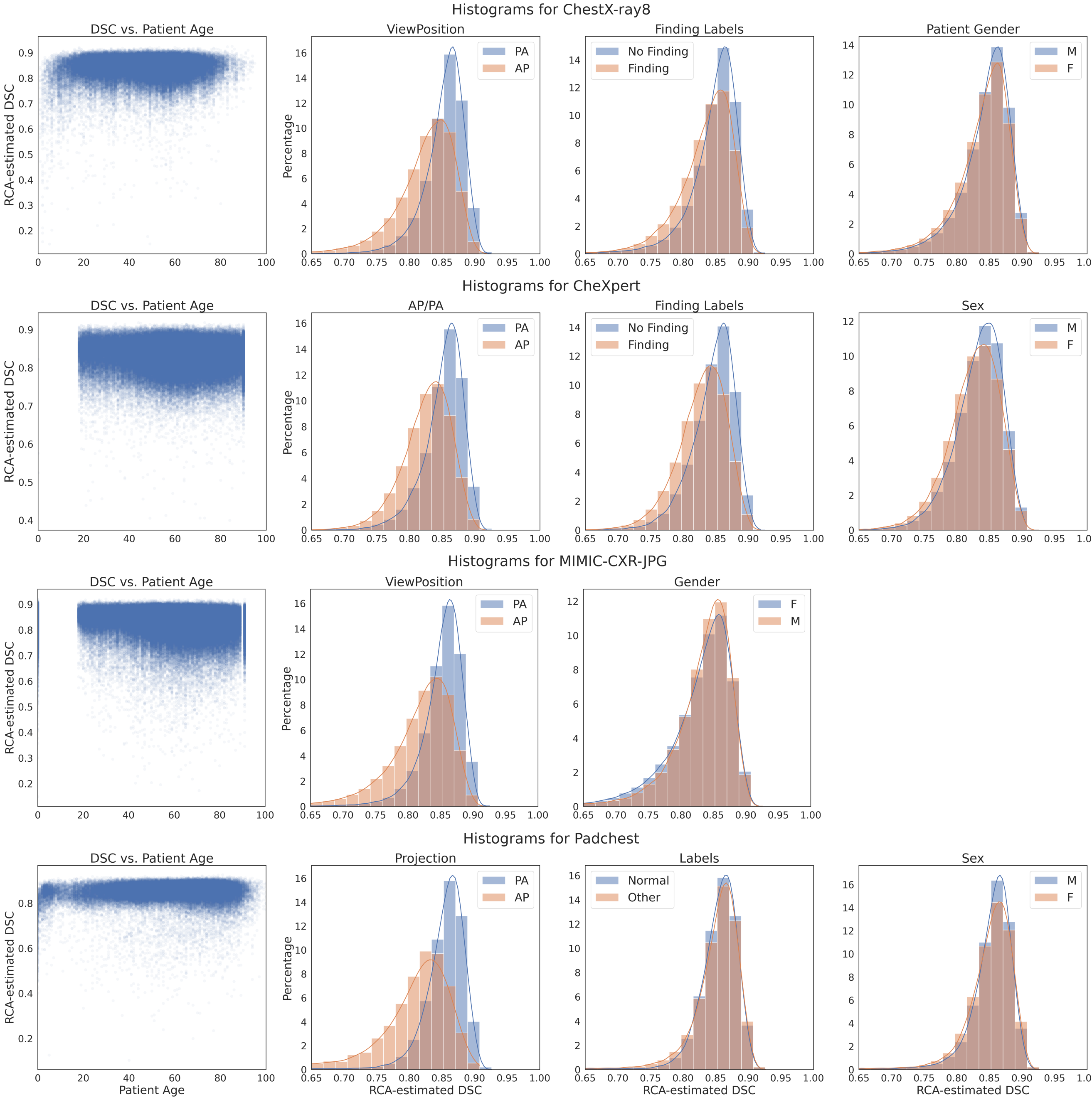

Figure 8. Histograms depicting RCA-estimated DSC distribution across diverse metadata categories, namely imaging protocol, sex, and disease findings, supplemented by scatter plots correlating RCA-estimated DSC and age. These visual aids elucidate the relationship between segmentation quality and metadata categories, particularly highlighting differences associated with X-ray capture view (PA versus AP), age and the presence of disease findings.

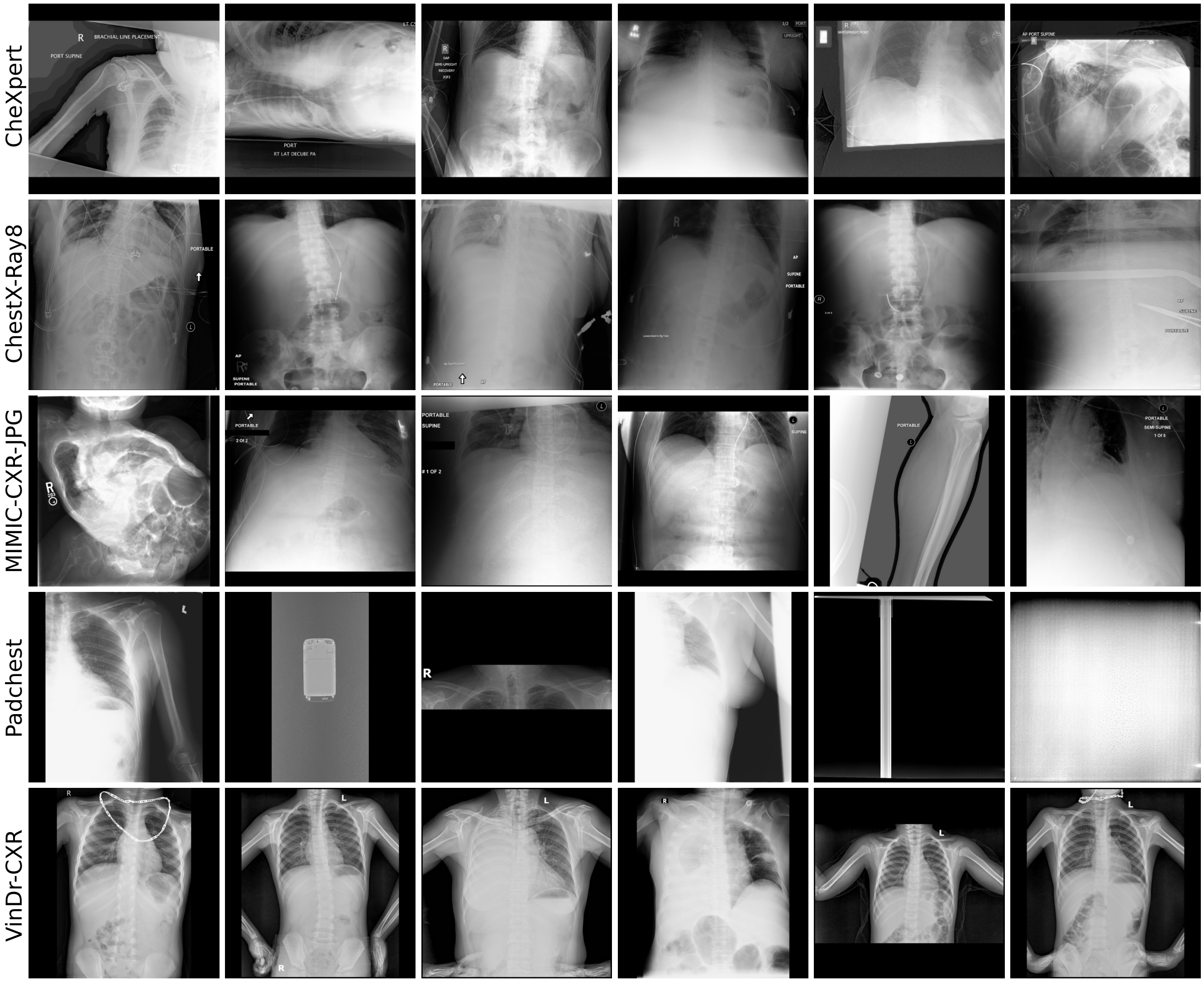

Figure 9. Illustration of a random subset of images with some of the lowest RCA values from each dataset. These images were identified as having low-quality segmentations based on their corresponding RCA scores. The selection showcases various types of images, including those with noise, mislabeled lateral CXR, and even non-medical objects like a mobile phone in the Padchest dataset. Reproduction of these images was allowed upon request to the original sources.

Usage Notes

The CheXmask dataset is intended to serve as a resource for researchers working in the field of medical imaging, particularly those interested in computational anatomy, shape understanding, CXR analysis and image segmentation. The dataset opens avenues for various practical applications, such as deep learning model development and evaluation, training of generative models, clinical decision support systems, disease detection and diagnosis, and anomaly detection. Researchers can explore novel methodologies, such as the integration of masked autoencoders for self-supervised learning, wherein the anatomical segmentation masks serve as a rich source for unsupervised feature extraction, or in the implementation of attention and deep supervision techniques guided by masks, facilitating enhanced model focus on specific anatomical regions for improved localization and feature extraction. As an example of clinical application, the task of cardiothoracic ratio estimation, can be particularly enhanced by leveraging the precise anatomical information encoded in the segmentation masks, thereby contributing to more robust and interpretable cardiomegaly detection methodologies.

For the sake of completeness, we release segmentation masks for all images included in our study. However, based on the qualitative and quantitative analyses presented in this paper (see section Technical Validation), our recommendation for downstream tasks is to only use segmentations whose RCA-estimated DSC is higher than 0.7, to avoid including out-of-distribution images as well as low quality masks. In addition, more attention should be put on specific subgroups (like pediatric patients in ChestX-ray8 or images acquired in AP position) for which we observed lower RCA-estimated DSC values overall, as discussed in Section "Validation via RCA-estimated DSC". Please also note that, when using segmentation masks for downstream tasks (e.g. computing cardiothoracic ratio from heart and lung segmentation masks), additional validation considering ground-truth annotations for the particular task should be performed to ensure the quality of the segmentation masks is good enough for the task at hand (e.g. manual measurements of cardiothoracic ratio should be employed to validate the quality of the segmentation masks).

A critical point to note is that this dataset does not include the original CXR images due to proprietary and privacy considerations. Hence, researchers intending to utilize this dataset are required to source the original images directly from the respective datasets cited in the manuscript. It is imperative that all conditions of usage and restrictions established by the providers of these datasets are fully respected and complied with.

The segmentation masks within this dataset are encoded using the Run-Length Encoding (RLE) technique, a standardized method for representing binary masks in a compact manner. In order to decode these masks back to their original binary format, researchers can use Python along with widely-used scientific computing libraries such as NumPy and Pandas. We provide the necessary scripts for encoding and decoding these masks to facilitate their use.

For researchers aiming to employ machine learning or deep learning techniques for analyzing this dataset, we recommend the use of well-established libraries. In particular, the HybridGNet model, used in this work, was developed in PyTorch. The source code for this model is made available alongside the dataset, allowing researchers to reproduce our results, conduct comparative studies, or further refine the methodology.

The segmentation masks are available in two formats. The first retains the original resolution as provided by the respective imaging source. The second consists of preprocessed masks with an uniform resolution across different imaging sources. We also provide the corresponding image preprocessing scripts which researchers can apply to the original images, enabling the use of all datasets at the same resolution. These scripts have been designed to replicate the preprocessing steps applied to the original datasets, thereby facilitating the integration of all these major CXR datasets.

Limitations of Study

There are several limitations to consider in this study. Firstly, the validation of the subset of images was performed by a single physician, which may introduce potential biases or subjective interpretations. While efforts were made to ensure accuracy and consistency, further validation by multiple radiologists or experts could enhance the reliability and generalizability of the findings.

The selection process focused on including only CXR images in the posteroanterior (PA) or anteroposterior (AP) view, aiming to ensure consistency and homogeneity. However, due to potential inaccuracies or inconsistencies in the metadata (particulary in large-scale databases built from automatic analysis of electronic health records), there is a possibility that other types of images may have been included in the PA and AP views. Despite the utilization of the Reverse Classification Accuracy (RCA) framework, which provides a reliable metric to detect low-quality segmentations, it is important to acknowledge that some images that do not meet the specific PA or AP criteria may have inadvertently leaked into the dataset. While RCA helps identify such instances, there remains a potential for the inclusion of non-PA or non-AP images that could impact the segmentation quality assessment. Therefore, when working with these large datasets, it is crucial to exercise caution and acknowledge the possibility of mislabeled or misclassified images. Further investigations and refinements in the dataset selection process, including more robust metadata validation techniques, would be beneficial to enhance the reliability and accuracy of the segmentation results.

Finally, it is important to highlight that, since segmentation masks released in CheXmask are generated with HybridGNet, they tend to follow patterns of anatomical plausibility, even in presence of complex conditions like tuberculosis, which generate different type of organ oclussions. This is of particular importance for lung segmentation, where there are usually two ways in which masks can be annotated, either following the 'air' or 'anatomy' strategy[35,36]. While the 'air' strategy focus on segmenting the air cavity of the lung, excluding areas of increased opacity due to infection, the 'anatomy' strategy provide a comprehensive view which follows the expected anatomy, including these opaque areas. Segmentation masks generated by HybridGNet (i.e. those in CheXmask) are more similar to those generated via the 'anatomy' strategy. For a complete discussion about this, see Section V.5) and Figure 6 in Gaggion et. al[8].

Code availability

The code associated with this study is available in our Github repository: `https://github.com/ngaggion/CheXmask-Database`. The repository encompasses Python 3 code for various components, including data preparation, data post-processing, technical validation, and the deep learning models. The data preparation section includes scripts for preprocessing the images. The data post-processing section provides scripts for converting the segmentation masks from run-length encoding to a binary mask format, examples of how to read the model and the necessary code to revert the pre-processing steps for each dataset. The technical validation section includes the code for the individual RCA analysis and the processing of the physician results. Additionally, the repository includes the code for the deep learning models used for image segmentation, including the HybridGNet model architecture, weights, training and inference scripts. The software prerequisites for running the code are outlined in the repository's README file.

References

1. Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

2. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 590–597 (2019).

3. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. review biomedical engineering* **19**, 221–248 (2017).

4. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017).

5. Roulet, N., Slezak, D. F. & Ferrante, E. Joint learning of brain lesion and anatomy segmentation from heterogeneous datasets. In *International Conference on Medical Imaging with Deep Learning*, 401–413 (PMLR, 2019).

6. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).

7. Gaggion, N., Mansilla, L., Milone, D. H. & Ferrante, E. Hybrid graph convolutional neural networks for landmark-based anatomical segmentation. In *MICCAI* (Springer International Publishing, 2021).

8. Gaggion, N., Mansilla, L., Mosquera, C., Milone, D. H. & Ferrante, E. Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. *IEEE Transactions on Med. Imaging* 10.1109/TMI.2022.3224660 (2022).

9. Wang, X. *et al.* Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106 (2017).

10. Johnson, A. E. *et al.* Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).

11. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med. image analysis* **66**, 101797 (2020).

12. Nguyen, H. Q. *et al.* Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Sci. Data* **9**, 429 (2022).

13. Valindria, V. V. *et al.* Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging* **36**, 1597–1606 (2017).

14. Jaeger, S. e. a. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Med. Imaging* **33**, 233–245, 10.1109/TMI.2013.2284099 (2014).

15. Candemir, S. & others. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Med. Imaging* **33**, 577–590, 10.1109/TMI.2013.2290491 (2014).

16. Shiraishi, J. *et al.* Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am. J. Roentgenol.* **174** (2000).

17. Seibold, C. *et al.* Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling (2023). 2306.03934.

18. Seibold, C. M. *et al.* Detailed annotations of chest x-rays via ct projection for report understanding. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022* (BMVA Press, 2022).

19. Chest-xray8 dataset. https://nihcc.app.box.com/v/ChestXray-NIHCC. Accessed: August 27, 2023.

20. Chexpert dataset. https://stanfordaimi.azurewebsites.net/datasets/8cbd9ed4-2eb9-4565-affc-111cf4f7ebe2. Accessed: August 27, 2023.

21. Johnson, A. *et al.* Mimic-cxr-jpg - chest radiographs with structured labels (version 2.0.0). https://physionet.org/content/mimic-cxr-jpg/2.0.0/, 10.13026/8360-t248 (2019).

22. Padchest dataset. https://bimcv.cipf.es/bimcv-projects/padchest/. Accessed: August 27, 2023.

23. Nguyen, H. Q., Pham, H. H., tuan linh, l., Dao, M. & khanh, l. Vindr-cxr: An open dataset of chest x-rays with radiologist annotations (version 1.0.0). https://physionet.org/content/vindr-cxr/1.0.0/, 10.13026/3akn-b287 (2021).

24. Gaggion, N., Vakalopoulou, M., Milone, D. H. & Ferrante, E. Multi-center anatomical segmentation with heterogeneous labels via landmark-based models. In *20th IEEE International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2023).

25. Japanese society of radiological technology (jsrt) dataset. http://db.jsrt.or.jp/eng.php. Accessed: August 27, 2023.

26. Nih tuberculosis chest x-ray datasets. https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/index.html.

27. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).

28. Mansilla, L., Milone, D. H. & Ferrante, E. Learning deformable registration of medical images with anatomical constraints. *Neural Networks* **124**, 269–279 (2020).

29. Gaggion, N. *et al.* CheXmask Database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 0.4). PhysioNet (2023). https://doi.org/10.13026/6eky-y831.

30. Goldberger, A. *et al.* Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000).

31. Jurdi, R. E. & Colliot, O. How precise are performance estimates for typical medical image segmentation tasks? In *IEEE International Symposium on Biomedical Imaging (ISBI 2023)* (IEEE, Cartagena de Indias, Colombia, 2023).

32. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci.* **117**, 12592–12594 (2020).

33. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *nature communications* **13**, 4581 (2022).

34. Ganz, M., Holm, S. H. & Feragen, A. Assessing bias in medical ai. In *Workshop on Interpretable ML in Healthcare at International Connference on Machine Learning (ICML)* (2021).

35. Karargyris, A. *et al.* Combination of texture and shape features to detect pulmonary abnormalities in digital chest x-rays. *Int. J. Comput. Assist. Radiol. Surg.* **11**, 99–106, 10.1007/s11548-015-1242-x (2016).

36. Larrazabal, A. J., Martínez, C., Glocker, B. & Ferrante, E. Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. *IEEE Transactions on Med. Imaging* **39**, 3813–3820 (2020).

Acknowledgements

This work was supported by Argentina's National Scientific and Technical Research Council (CONICET), which covered the salaries of E.F., D.M., L.M. and N.G. The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPU computing used for this research and Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (Agencia I+D+i PICT).

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
N. Gaggion, C. Mosquera, L. Mansilla, J. Saidman, M. Aineseder, D. H. Milone & E. Ferrante; "CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images"
Scientific Data, No. 1, pp. 511, 2024.

Author contributions statement

E.F., N.G. and C.M. conceptualized the idea of this paper. N.G. performed the experiments. C.M. designed the LabelStudio interface and analyzed the results. L.M. implemented the deformable registration module for the RCA estimation. M.A. and J.M.S manually segmented the gold-standard set. D.H.M. and E.F. analyzed the results. All authors reviewed the manuscript.

Competing interests

The authors declare not conflict of interest.

Appendix: Statistical analysis for Max RCA-estimated DSC scores

In the main manuscript we included statistical analysis for Mean RCS-estimated DSC scores since this indicator showed better correlation with real DSC. However, for the sake of completeness, here we include the same analysis for the alternative Max RCA-estimated DSC score.

Table 5. Dice RCA (Max) statistical analysis

Dataset name	Sample size (n)	Mean	Std	Min	1%	5%	25%	50%	75%	Max
Chestx-ray8	112120	0.882	0.044	0.164	0.732	0.806	0.864	0.890	0.909	0.967
CheXpert	187825	0.870	0.038	0.415	0.757	0.801	0.849	0.875	0.896	0.961
MIMIC-CXR-JPG	243334	0.873	0.050	0.179	0.690	0.784	0.854	0.884	0.904	0.968
Padchest	96184	0.893	0.042	0.128	0.714	0.831	0.882	0.901	0.917	0.970
VinDr-CXR	18000	0.892	0.035	0.501	0.778	0.831	0.876	0.898	0.915	0.964

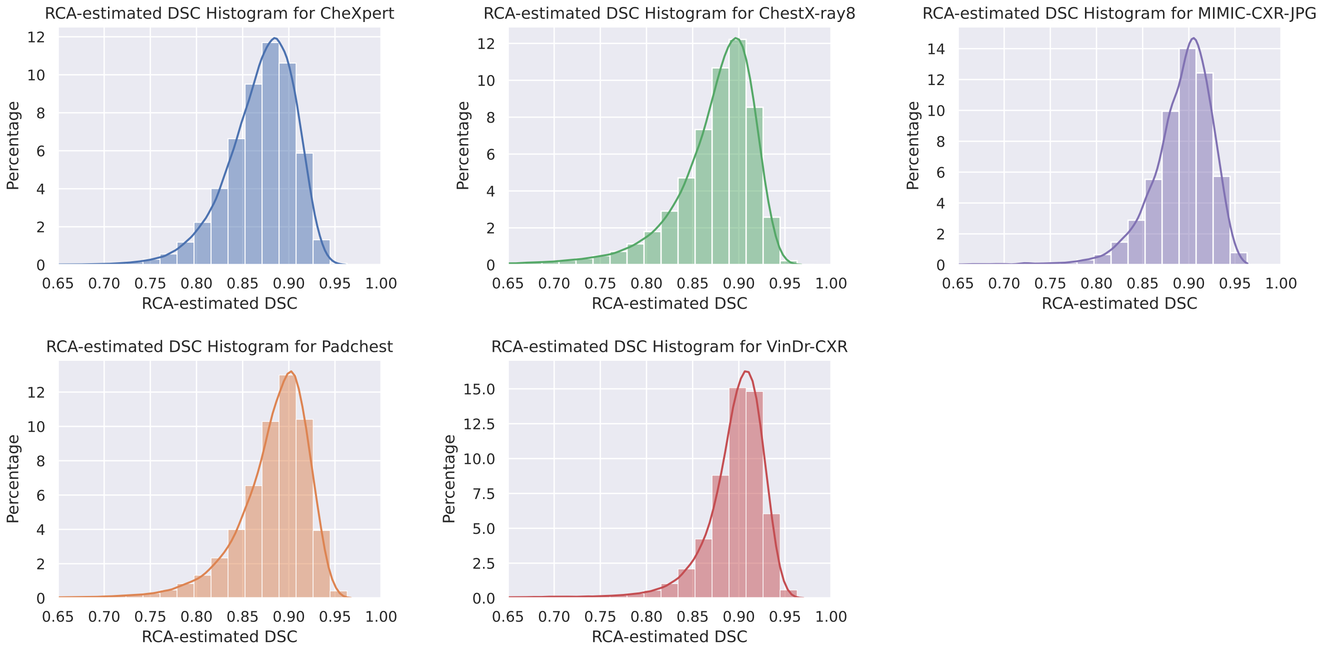

Figure 10. Histograms illustrating the distribution of segmentation quality across the five datasets. Each histogram represents the distribution of RCA (Reverse Classification Accuracy) estimations of DSC for a specific dataset using the Max value instead of the mean, providing a visual representation of the segmentation performance. The histograms are truncated at 0.65 for visualization purposes. The full range of values is considered in the statistical analysis (Table 5).