Dual center validation of deep learning for automated multi-label segmentation of thoracic anatomy in bedside chest radiographs

Computer Methods and Programs in Biomedicine 234 (2023) 107505

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Dual center validation of deep learning for automated multi-label segmentation of thoracic anatomy in bedside chest radiographs

Felix Busch[a,b,*], Lina Xu[a], Dmitry Sushko[a], Matthias Weidlich[a], Daniel Truhn[c], Gustav Müller-Franzes[c], Maurice M. Heimer[d], Stefan M. Niehues[a], Marcus R. Makowski[e], Markus Hinsche[a], Janis L. Vahldiek[a], Hugo JWL. Aerts[f,g,h,i], Lisa C. Adams[a,f,#], Keno K. Bressem[a,f,g,#]

[a] Department of Radiology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany
[b] Department of Anesthesiology, Division of Operative Intensive Care Medicine, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany
[c] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany
[d] Department of Radiology, Ludwig-Maximilians-University of Munich, Munich, Germany
[e] Department of Radiology, Technical University of Munich, Munich, Germany
[f] Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
[g] Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA
[h] Departments of Radiation Oncology and Radiology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, MA, USA
[i] Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands

ARTICLE INFO

Article history:
Received 29 May 2022
Revised 17 February 2023
Accepted 21 March 2023

Keywords:
Anatomy
Active learning
Chest radiograph
Artificial intelligence
Deep learning
Convolutional neural network

ABSTRACT

Background and Objectives: Bedside chest radiographs (CXRs) are challenging to interpret but important for monitoring cardiothoracic disease and invasive therapy devices in critical care and emergency medicine. Taking surrounding anatomy into account is likely to improve the diagnostic accuracy of artificial intelligence and bring its performance closer to that of a radiologist. Therefore, we aimed to develop a deep convolutional neural network for efficient automatic anatomy segmentation of bedside CXRs.

Methods: To improve the efficiency of the segmentation process, we introduced a "human-in-the-loop" segmentation workflow with an active learning approach, looking at five major anatomical structures in the chest (heart, lungs, mediastinum, trachea, and clavicles). This allowed us to decrease the time needed for segmentation by 32% and select the most complex cases to utilize human expert annotators efficiently. After annotation of 2,000 CXRs from different Level 1 medical centers at Charité – University Hospital Berlin, there was no relevant improvement in model performance, and the annotation process was stopped. A 5-layer U-ResNet was trained for 150 epochs using a combined soft Dice similarity coefficient (DSC) and cross-entropy as a loss function. DSC, Jaccard index (JI), Hausdorff distance (HD) in mm, and average symmetric surface distance (ASSD) in mm were used to assess model performance. External validation was performed using an independent external test dataset from Aachen University Hospital ($n = 20$).

Results: The final training, validation, and testing dataset consisted of 1900/50/50 segmentation masks for each anatomical structure. Our model achieved a mean DSC/JI/HD/ASSD of 0.93/0.88/32.1/5.8 for the lung, 0.92/0.86/21.65/4.85 for the mediastinum, 0.91/0.84/11.83/1.35 for the clavicles, 0.9/0.85/9.6/2.19 for the trachea, and 0.88/0.8/31.74/8.73 for the heart. Validation using the external dataset showed an overall robust performance of our algorithm.

Abbreviations: ASSD, average symmetric surface distance; CNN, convolutional neural network; CXR, chest radiograph; DSC, Dice similarity coefficient; HD, Hausdorff distance; ICU, intensive care unit; JI, Jaccard index; PA, posterior-anterior view; SD, standard deviation.

* Corresponding author.
E-mail address: felix.busch@charite.de (F. Busch).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2023.107505

Conclusions: Using an efficient computer-aided segmentation method with active learning, our anatomy-based model achieves comparable performance to state-of-the-art approaches. Instead of only segmenting the non-overlapping portions of the organs, as previous studies did, a closer approximation to actual anatomy is achieved by segmenting along the natural anatomical borders. This novel anatomy approach could be useful for developing pathology models for accurate and quantifiable diagnosis.

Short abstract

This study presents a novel real anatomy-based multi-label segmentation approach for automated anatomical prediction of the lungs, heart, clavicles, trachea, and mediastinum in bedside chest radiographs (CXRs). Segmentation efficiency was optimized using a human-in-the-loop approach with active learning, including 2,000 anterior-posterior view bedside CXRs from different Level 1 medical centers. A 5-layer U-ResNet was trained for 150 epochs with combined soft Dice similarity coefficient (DSC) and cross-entropy as a loss function. Model performance was evaluated based on DSC, Jaccard index (JI), Hausdorff distance (HD) in mm, and average symmetric surface distance (ASSD) in mm. The final training/validation/test dataset consisted of 1,900/50/50 segmentation masks of each anatomical structure. Our model achieved a mean DSC/JI/HD/ASSD of 0.93/0.88/32.1/5.8 for the lungs, 0.92/0.86/21.65/4.85 for the mediastinum, 0.91/0.84/11.83/1.35 for the clavicles, 0.9/0.85/9.6/2.19 for the trachea, and 0.88/0.8/31.74/8.73 for the heart.

1. Introduction

Bedside chest radiographs (CXRs) are commonly used in the emergency department or intensive care unit (ICU) to aid diagnosis and disease management but are often difficult to interpret due to limited image quality as they are usually taken under more challenging conditions [1–3]. Given that bedside CXRs are often taken in critically ill patients, accurate diagnosis is of particular relevance, e.g., for monitoring cardiopulmonary diseases or determining the location of invasive therapy devices [4–6].

Apart from image quality, the interpretation of bedside CXRs relies heavily on a thorough understanding of human anatomy. Radiologists must have a detailed knowledge of the structures within the chest, including the bones, vessels, and organs, to ensure accurate interpretation.

In clinical practice, deep learning approaches are emerging as additional diagnostic tools for CXRs. Some models reach the expert level in detecting thoracic abnormalities, e.g., pneumothorax, mediastinal widening, pneumoperitoneum, pleural effusion, atelectasis, fibrosis, cardiomegaly, or specific diseases such as SARS-CoV-2 pneumonia or tuberculosis [7–10]. Other applications include identification of therapy devices or segmentation of anatomy [11,12]. Convolutional neural network (CNN) architectures proved particularly useful for these purposes [13–17]. In a recent study, the diagnostic performance of 20 experienced radiologists was improved in 102 of 127 clinical findings when the evaluation of CXRs was supported by a comprehensive CNN model [18]. Most existing models are based on CXRs obtained in the standing or sitting position and do not consider bedside thoracic radiography, where heterogeneous imaging features and common thoracic pathologies complicate anatomy delineation.

While previous deep learning approaches succeeded in classification tasks, they were not developed taking into account anatomical context.

Using a computer-aided human-in-the-loop segmentation workflow with an active learning approach to improve segmentation efficiency, we here present the first multi-label CNN-based architecture for automatic prediction and segmentation of five major anatomical structures (lung, heart, trachea, mediastinum, and clavicles) in bedside CXRs.

The article is organized as follows: In the Materials and Methods section, we present our dataset, approach, and statistical methods and describe the external independent validation dataset. In the Results, we first present the segmentation performance in our dataset and show excellent and poor labeling results. We then report the results for the external independent dataset. In the Discussion, we provide an overview of the current state of the art in multi-label segmentation approaches for CXRs compared to our results and discuss the potential benefits and limitations of our model.

2. Materials and methods

Ethics approval was granted by the Ethics Committee of Charité – University Hospital Berlin (EA4/042/20) in line with the Declaration of Helsinki, including a waiver of informed consent due to the retrospective design of the study.

2.1. Dataset and pre-processing

Two thousand AP bedside CXRs from different level 1 medical centers at Charité – University Hospital Berlin obtained between 2009 and 2019 were randomly selected from our local PACS (Picture Archiving and Communication System) and exported to an in-hospital server using DICOM (Digital Imaging and Communications in Medicine) format. Subsequently, all CXRs were anonymized using Python (v. 3.8.13) with SimpleITK (v. 2.1.1).

Bedside radiography devices from multiple manufacturers were used for image acquisition (see Table 1), resulting in heterogeneous pixel sizes (1778 × 2092 to 3520 × 4280) and spacing (0.1 to 0.168 mm), as well as different gray levels (16 to 64 bit) for each radiograph.

2.2. Manual anatomical segmentation

Manual segmentation was performed using 3D Slicer software. For each CXR, a 2D ground-truth mask was created for the heart, lungs, mediastinum, trachea, and clavicles and stored in NRRD (Nearly Raw Raster Data) format. Segmentations were performed by two sixth-year medical students (LX, DS) and a fourth-year radiology resident (MW). Segmentation masks were then reviewed by two board-certified radiologists with six and seven years of experience in chest radiology (LCA, KKB). For quality assurance, each

Table 1
Overview of manufacturers of the devices used for chest radiography in our study with the respective number of examinations.

Manufacturer	Number of exams	%
Canon Inc.	265	12%
Carestream Health	9	0.4%
FUJI Photo Film Co., ltd.	579	26.3%
GE Healthcare	1	0.05%
Kodak	1165	53%
Philips Medical Systems	1	0.05%
Siemens	180	8.2%

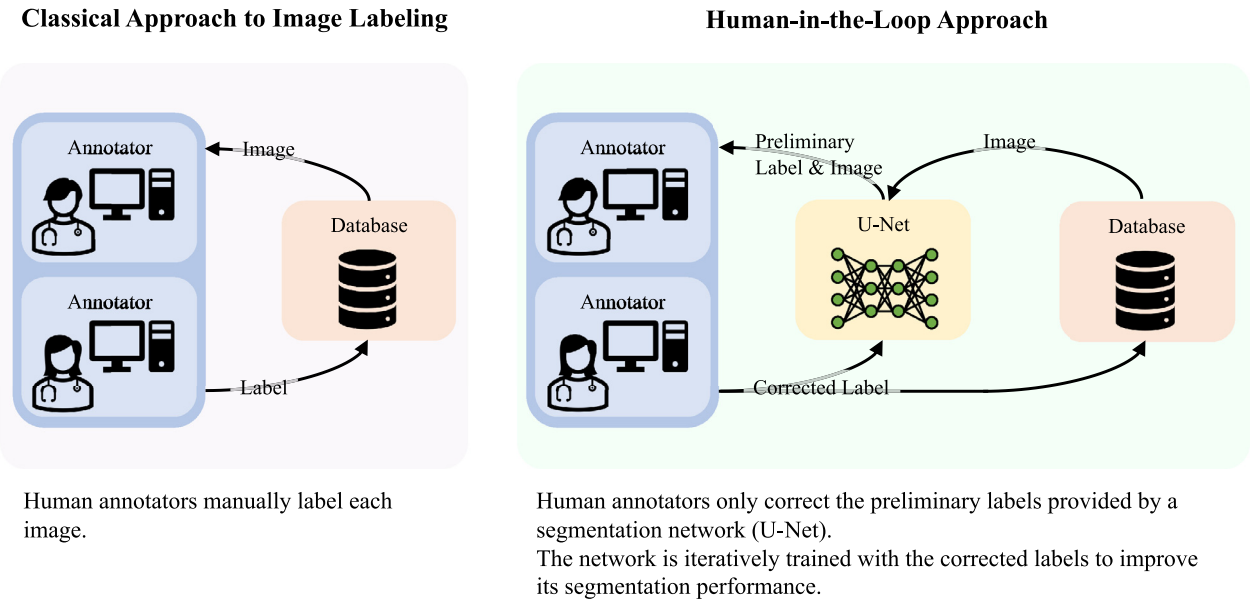

Fig. 1. Diagram of the human-in-the-loop approach used in our study to increase the speed and effectiveness of the human annotators.

anatomical mask was finally reviewed by a board-certified radiologist with twelve or 20 years of experience and adjusted if necessary (JLV or SMN).

2.3. Computer-assisted anatomical segmentation

To improve the efficiency of the segmentation process, a computer-aided segmentation workflow was implemented using MONAILabel (v. 0.3.2) [19]. After segmentation of 500 images, a U-Net model was trained to produce preliminary segmentations. The human annotators only corrected the preliminary segmentations of the model. After 750, 1000, 1500, and 2000 segmentations, the model was re-trained to improve segmentation quality. After 1500 segmentations, the model already achieved subjectively excellent performance on CXRs without pathological findings. Therefore, an active learning approach was adopted to identify and prioritize complex images in the dataset for annotation. To this end, the model made repeated predictions for individual CXR images with random dropout applied to model weights. Based on the assumption that complex CXRs have a higher individual prediction variance, this allowed us to compute epistemic uncertainty and its use to identify potentially difficult CXRs.

Fifty of the 500 most difficult CXRs were randomly selected as the test dataset, and the remaining images were included in the labeling workflow. Without this workflow, segmentation of a single CXR took b 2–8 min (mean ± standard deviation (SD): 4:23 ± 2:04 min), depending on the difficulty level. After implementation of this workflow, the mean annotation time decreased to 2:58 ± 1:23 min. Appendix A (Figure A.1) provides further details on annotation times. The validation dataset consisted of 50 randomly selected CXRs from the whole dataset.

Fig. 1 provides a schematic overview of our human-in-the-loop workflow.

2.4. Model training and post-processing

Model training was performed using MONAI (v. 0.8.1) and PyTorch (v. 1.11.0). A 5-layer U-ResNet was trained for 150 epochs with early stopping using a combination of Dice similarity coefficient (DSC) and cross-entropy as loss function and monitoring the DSC as a key metric [20]. During training, images were resized to 512 × 512 pixels, and a moving window of size 384 × 384 pixels was used to extract image subregions for training. A further increase in image resolution did not improve the key metric. Several image modifications, such as cropping, rotating, contrast, and brightness changes, were randomly applied to the images during training. The resolution of the segmentation masks created by the model was smaller than that of the original images, so resizing had to be applied to create overlays. As this resulted in stair-step artifacts, median smoothing was applied to the resized segmentation masks.

Training was performed on an Ubuntu 20.04 Workstation with AMD Ryzen™ Threadripper™ 2970WX Processor (Advanced Micro Devices, Santa Clara, California, United States), 64 GB of RAM, and 24 GB VRAM Nvidia GeForce RTX 3090 (Nvidia, Santa Clara, California, United States). For training, the learning rate was kept at 1e-3. A weight decay of 0.001 and a dropout of 0.1 were used during training to regularize the model.

For an exemplary illustration of the predicted segmentation masks, labels within the top and bottom 5% Dice scores were randomly selected. Semi-transparent contoured masks of the predicted and associated manually labeled masks were then created and superimposed on the original image (see Appendix B).

2.5. Metrics and statistical analysis

Python (v. 3.8.13) with MONAI (v. 0.8.1) and PyTorch (v. 1.11.0) were used to read and process the radiographs and labels for evaluation. Furthermore, R (v. 4.1.2), including the tidyverse (v. 1.3.1) package, was used for statistical analysis and the creation of boxplots and images. DSC, Jaccard Index (JI), Hausdorff distance in mm (HD), and average symmetric surface distance in mm (ASSD) were taken as metrics to report the performance achieved for each anatomical structure [20,21]. Mean and SD, as well as median and interquartile range, were reported for each metric. Please refer to Appendix B for a detailed explanation of the evaluation, including all necessary code.

2.6. Validation on an independent external dataset

Our model was validated on an independent external test dataset consisting of 20 randomly selected bedside CXRs from the

Table 2
Overview results of the median and mean Dice similarity coefficient (DSC), Jaccard index (JI), Hausdorff distance (HD), and average symmetric surface distance (ASSD) with interquartile range (IQR) or standard deviation (SD) for the predicted segmentation labels of the heart, lungs, mediastinum, trachea, and clavicles.

Label	DSC		JI		HD (mm)		ASSD (mm)	
	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)
Heart	0.91 (0.07)	0.88 (0.07)	0.83 (0.12)	0.8 (0.1)	27.14 (22.46)	31.74 (16.62)	6.91 (5.4)	8.73 (4.79)
Lungs	0.93 (0.03)	0.93 (0.02)	0.88 (0.05)	0.88 (0.04)	27.29 (15.72)	32.1 (17.57)	5.35 (2.04)	5.8 (2.78)
Mediastinum	0.93 (0.05)	0.92 (0.05)	0.87 (0.09)	0.86 (0.08)	16.49 (15.02)	21.65 (13.14)	4.29 (3.65)	4.85 (3.7)
Trachea	0.97 (0.09)	0.9 (0.16)	0.94 (0.15)	0.85 (0.19)	1.1 (7.86)	9.6 (16.67)	0.5 (1.34)	2.19 (4.48)
Clavicles	0.92 (0.04)	0.91 (0.06)	0.86 (0.07)	0.84 (0.08)	9.07 (4.7)	11.83 (14.15)	1.18 (0.68)	1.35 (0.76)

level 1 medical center at University Hospital Aachen. The imaging devices used were four mobile X-ray systems (Mobilett MIRA, Siemens Healthcare, Erlangen, Germany), and images were acquired in seven surgical and three internal medicine ICUs. All CXR images in this test set were 16-bit grayscale images with a spatial resolution of 2140 × 1760 pixels and 0.2 × 0.2 mm spacing. The images were labeled by a local radiologist with nine years of experience in chest radiology (DT). Metrics described above were applied to evaluate performance.

3. Results

The final training dataset consisted of 1,900 bedside CXRs with manually segmented heart, trachea, lungs, mediastinum, and clavicles. The validation and test datasets each included 50 labeled CXRs.

Of the 2,000 CXRs in this study, 812 (40.6%) were from female patients, 1,170 (58.5%) were from male patients, and 18 (0.9%) from patients who did not wish to provide information. The mean age (± SD) of female patients in our study was 63.5 ± 17.1 years at the time of CXR examination. Males had a mean age of 63.0 ± 16.2 years. Patients who chose not to report their sex had a mean age of 56.4 ± 18.1 years.

3.1. Performance on the test dataset

In the test dataset, the highest mean DSC and JI values for lung segmentation were 0.93/0.88, followed by 0.92/0.86 for the mediastinum, 0.91/0.84 for the clavicles, 0.9/0.85 for the trachea, and 0.88/0.8 for the heart. There were outliers in segmentation performance, particularly for the trachea and clavicles. When median DSC and JI instead of means were considered, segmentation performance was best for the trachea at 0.97/0.94, followed by 0.93/0.88 for the lung, 0.93/0.87 for the mediastinum, 0.92/0.86 for the clavicle, and 0.91/0.83 for the heart.

Mean HD and ASSD was 9.6/2.19 for the trachea, 11.83/1.35 for the clavicles, 21.65/4.85 for the mediastinum, 31.74/8.73 for the heart, and 32.1/5.8 for the lungs. Again, for medians instead of means, the best segmentation performance was obtained for the trachea with an HD/ASSD of 1.1/0.5, followed by the clavicles with 9.07/1.18, the mediastinum with 16.49/4.29, the heart with 27.14/6.91, and the lung with 27.29/5.35.

An overview of all metrics and their distributions is provided in Table 2 and Fig. 2. Fig. 3 shows examples of the predicted labels within the top 5% of the Dice score. In contrast, Fig. 4 shows radiographs with a significant deviation of the predicted labels from the manually created segmentation masks. Additional image examples with corresponding original and predicted masks can be found in the supplementary material (Appendix A, Figure A.2).

3.2. Performance on the external test dataset

Validating our model on an independent external test dataset of 20 randomly selected ICU AP bedside CXRs resulted in a mean DSC/JI/HD/ASSD of 0.92/0.86/25.04/5.77 for the lungs, 0.87/0.77/29.91/9.29 for heart, 0.86/0.76/27.53/8.57 for mediastinum, 0.83/0.72/18.06/1.96 for clavicles, and 0.72/0.58/23.4/4.98 for the trachea. Again, due to the high variance of the achieved performance, it is advisable to consider medians, which were as follows: DSC/JI/HD/ASSD of 0.93/0.87/24.22/5.0 for the lungs, 0.88/0.78/29.14/8.99 for heart, 0.87/0.77/25.92/8.38 for mediastinum, 0.85/0.74/13.47/1.86 for clavicles, and 0.73/0.57/23.05/3.83 for the trachea. Overall, satisfactory performance was achieved for each anatomical structure. Notably, tracheal segmentations deviated slightly from the results for our internal test dataset, while all other results were highly comparable.

Table 3 summarizes the results of the median and mean metrics for each anatomical structure. Fig. 5 displays the distribution of each metric for the external test dataset.

4. Discussion

We propose a multi-label segmentation model for identification of the true anatomical extent of the heart, lungs, trachea, clavicles, and mediastinum on bedside CXRs. The solid performance on an independent external dataset underscores the model's generalizability. Furthermore, this study provides evidence for the effectiveness of combining computational human-in-the-loop approaches with active learning.

The best mean DSC/JI was obtained for the lungs, followed by the mediastinum, clavicles, trachea, and heart. Of note, we here used more difficult-to-interpret bedside CXRs from different level 1 medical centers [6]. In addition, we aimed to approximate the true size of the chest organs, rather than to segment only the non-overlapping parts, to obtain a more realistic architecture for anatomical prediction. See Appendix A (Figure A.3) for an example of a bedside CXR image delineating the complete lung versus its visible portion only. Validation of our model on an independent, external test dataset of bedside CXRs yielded overall convincing results, with only minor performance discrepancy in tracheal prediction (mean DSC/JI of 0.72/0.57 versus 0.9/0.85). What may have contributed to this comparatively lower DSC value is that segmentation of the trachea is particularly challenging due to low image contrast in the mediastinum. Furthermore, there is no clear anatomical boundary where the segmentation of the trachea should end proximally and distally, which further contributes to different segmentations by human readers. The latter may have led to observer bias in the test dataset, as the labeling technique may have been different.

We compared our results with the current state-of-the-art multi-label segmentation approaches for at least three anatomical structures in CXRs (please refer to Table 4).

Notably, most approaches included segmentation of the heart, lungs, and clavicles [24,26-32] while only one model performed additional segmentation of the trachea [22]. There is no comparable multi-label model for mediastinum segmentation. Of all approaches, the UNet_ResNeXt50_Masks+Contours model by Kholi-

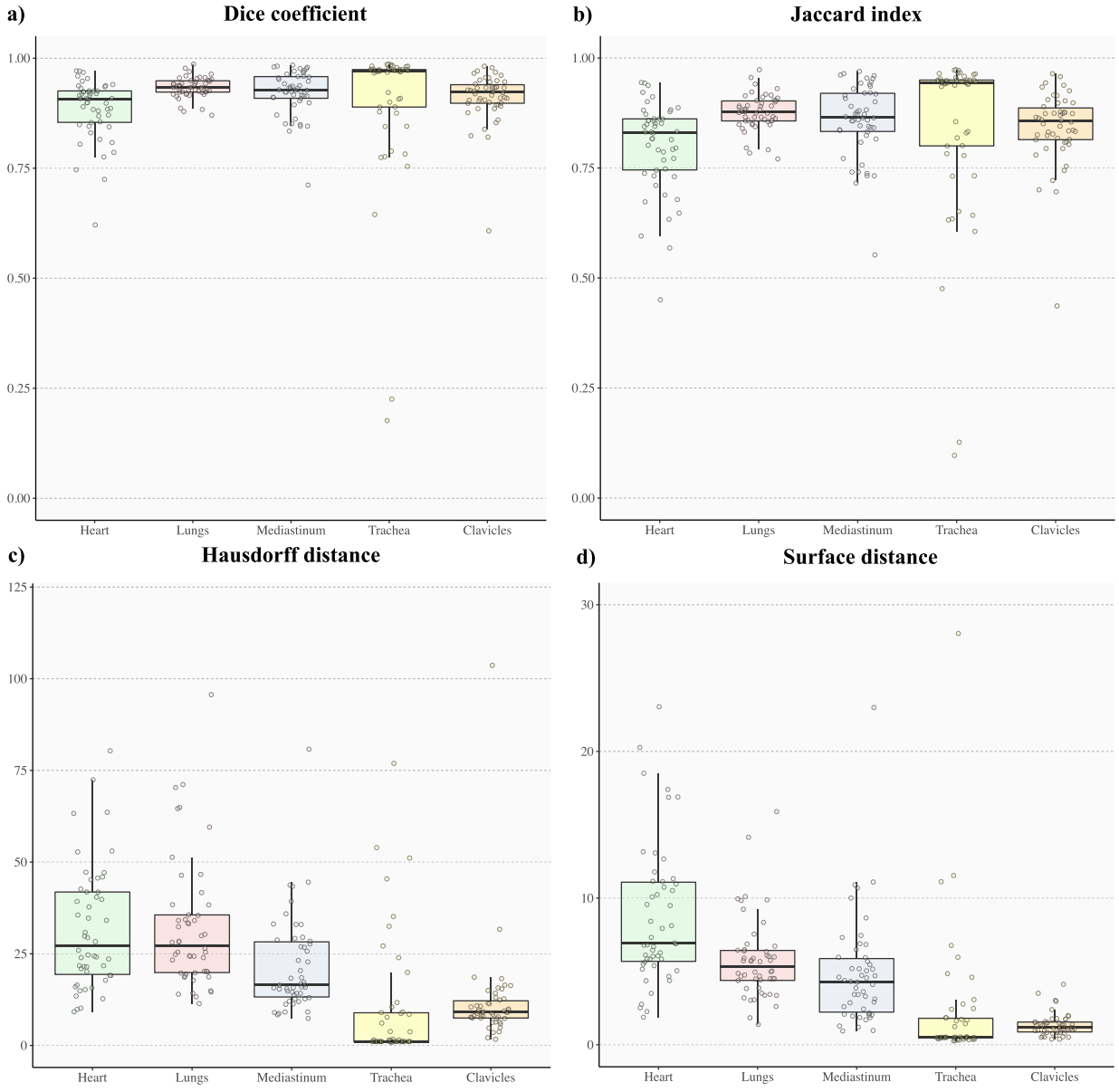

Fig. 2. Boxplots displaying the distribution of the Dice similarity coefficient (**a**), Jaccard index (**b**), Hausdorff distance in mm (**c**), and average symmetric surface distance in mm (**d**) for the predicted segmentation labels of the heart (green plots), lungs (red plots), mediastinum (blue plots), trachea (yellow plots), and clavicles (orange plots).

avchenko et al. showed the highest DSC/JI for segmentation of the heart (0.97/0.93), lungs (0.99/0.97), and clavicles (0.95/0.90) [24]. For clavicles, the nnU-Net approach of Gaggion et al. achieved a similar performance (DSC: 0.95/JI: 0.90) [32]. These results surpass our metrics for the heart (DSC: 0.88/JI: 0.80), lungs (DSC: 0.93/JI: 0.88), and clavicles (DSC: 0.91/JI: 0.84).

However, regarding these performance metrics for anatomical heart and lung segmentation, it is important to point out that our model approximates the true anatomical size compared to previous approaches, which focused on the borders of projection radiography. In addition, the performance of other published works depends on the dataset examined, including variations in pathologies

Table 3
Overview of model performance on the external test dataset. Median and mean Dice similarity coefficient (DSC), Jaccard index (JI), Hausdorff distance (HD), and average symmetric surface distance (ASSD) with interquartile range (IQR) or standard deviation (SD) are displayed for the predicted segmentation labels of the heart, lungs, mediastinum, trachea, and clavicles.

Label	DSC		JI		HD (mm)		ASSD (mm)	
	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)
Heart	0.88 (0.06)	0.87 (0.05)	0.78 (0.09)	0.77 (0.07)	29.14 (13.98)	29.91 (10.45)	8.99 (3.48)	9.29 (3.32)
Lungs	0.93 (0.03)	0.92 (0.03)	0.87 (0.05)	0.86 (0.04)	24.22 (9.17)	25.04 (8.0)	5.0 (2.35)	5.77 (1.71)
Mediastinum	0.87 (0.04)	0.86 (0.05)	0.77 (0.07)	0.76 (0.07)	25.92 (11.16)	27.53 (9.37)	8.38 (4.64)	8.57 (2.94)
Trachea	0.73 (0.15)	0.72 (0.11)	0.57 (0.19)	0.58 (0.13)	23.05 (14.88)	23.4 (12.99)	3.83 (2.96)	4.98 (3.48)
Clavicles	0.85 (0.04)	0.83 (0.05)	0.74 (0.07)	0.72 (0.07)	13.47 (9.76)	18.06 (11.51)	1.86 (0.43)	1.96 (0.51)

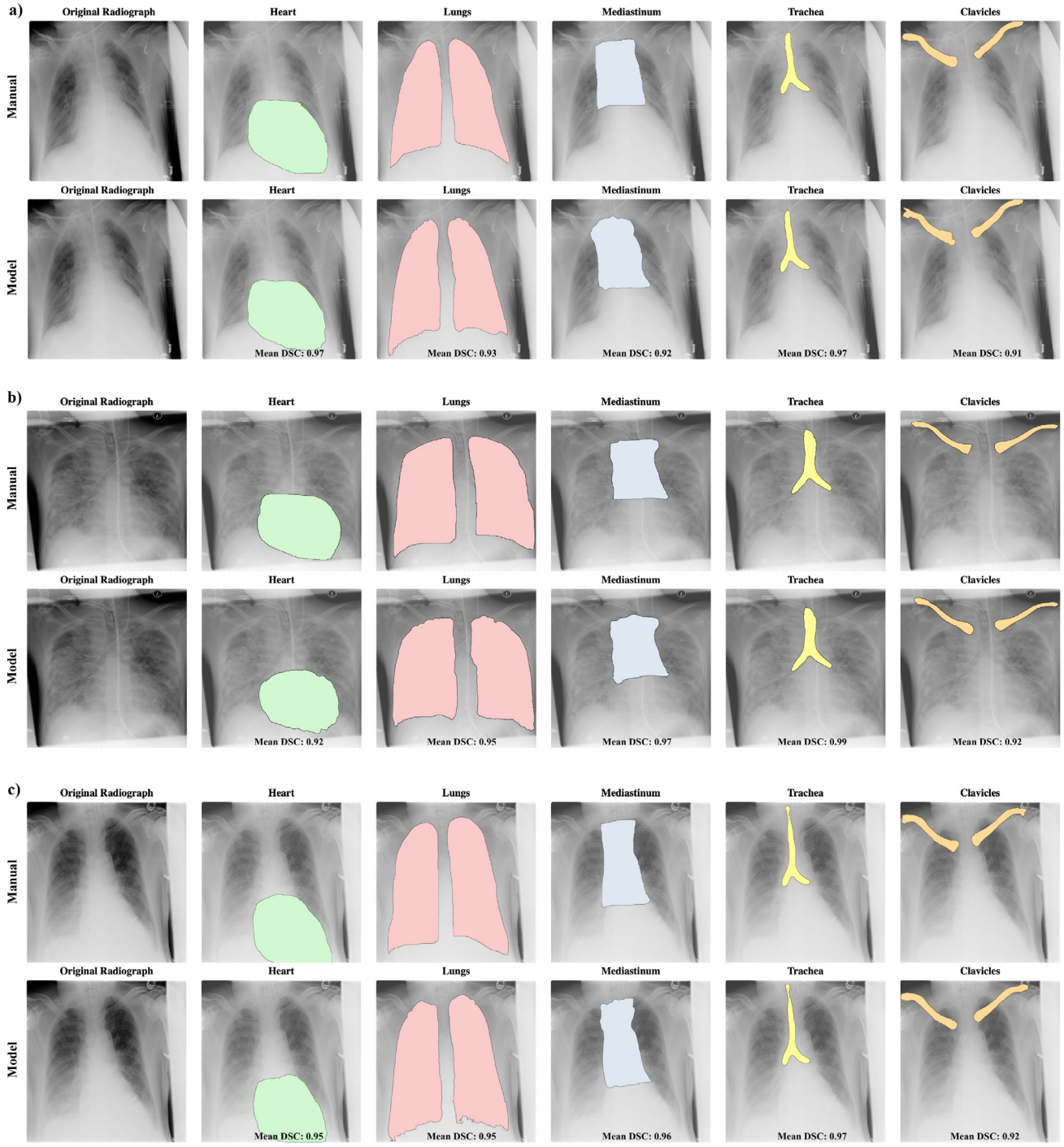

Fig. 3. Example series of chest radiographs within the top 5% of the Dice similarity coefficient (DSC) for at least one of the predicted segmentation masks. Notes: Upper rows in **a**), **b**), and **c**) show overlaid semi-transparent manually created segmentation masks compared to the model's predicted segmentations below each image for the heart (green), lungs (red), mediastinum (blue), trachea (yellow), and clavicles (orange). In the predicted masks, slight rim inhomogeneities can be occasionally observed.

and image characteristics, making it difficult to compare different segmentation models in detail and identify state-of-the-art models for each anatomical structure [24]. Kholiavchenko et al. and Gaggion et al. reported their metrics on 247 CXRs of the publicly available JSRT database [24,25,32]. The JSRT database contains 100 PA view standing CXRs of malignant nodules, 54 images of benign lung nodules, and 93 images without pathology (2048 $\times$ 2048 pixel size and 0.175 mm pixel spacing). By comparison, our U-ResNet is based on a different dataset consisting of ICU bedside CXRs, which are generally considered more challenging to interpret than standing CXRs because of the higher proportion of overlapping soft tissue, the higher frequency of pathologies, as well as poorer patient compliance [6]. Moreover, our model was trained on CXRs from different centers acquired with different x-ray machines with consecutively heterogeneous pixel sizes and spacing as well as different gray levels, which may have degraded model performance. Finally, patient misplacement/rotation can also influence segmentation accuracy [24]. In support of this argument, other authors reported a decrease in their lung segmentation performance when they applied their segmentation models to the public Mont-

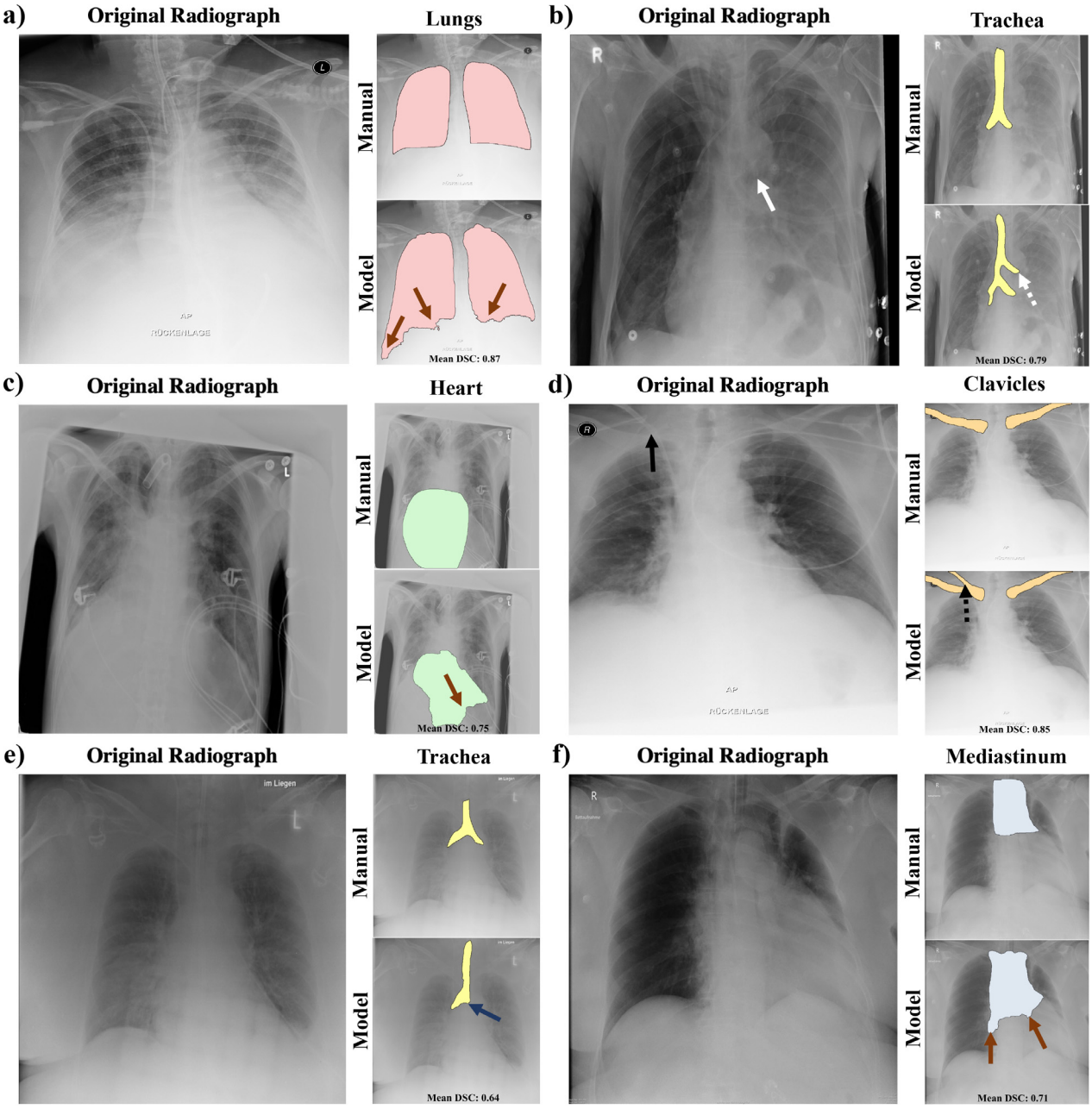

Fig. 4. Selection of chest radiographs with a significant deviation of the predicted labels from the manually created segmentation masks. Notes: The brown arrows in **a**), **c**), and **f**) indicate poorly delineated lungs, heart, and mediastinal borders in areas with compaction of soft tissue in the model-derived segmentation masks compared to the corresponding manual segmentations shown above. Image series **b**) shows a dilated esophagus in the original radiograph (white arrow), which was predicted as part of the trachea by our model (white dotted arrow). The black arrow in **d**) marks a peripherally inserted central catheter, which was predicted as part of the right clavicle by our model (black dotted arrow). In **e**), especially the course of the upper trachea was more accurately predicted than initially labeled in the manual mask above. In contrast, the outlet of the left main bronchus is incomplete (blue arrow). The complete image series of the selected radiographs and additional images within the top and bottom 5% of the Dice similarity coefficient (DSC) are provided in Appendix A (Figure A.2).

gomery dataset, which contains 58 of 138 CXRs with tuberculosis in standing PA view (4020 × 4892 or 4892 × 4020 pixel size and 0.0875 mm pixel spacing), instead of the JSRT dataset, where lung nodules have little effect on the delineation of anatomical structures [24,33,34].

Remarkably, our approach outperformed the only other available multi-label segmentation approach with trachea segmentation proposed by Pal et al., although they validated their approach on PA standing CXRs of the ChestX-ray8 dataset instead of AP bedside images as we did [22,23]. Of note, it can be beneficial to include the prediction of tracheal or mediastinal anatomy in multi-label segmentation models, e.g., when evaluating anatomical variations of the respective structures in the presence of concurrent diseases or when overlapping invasive therapeutic devices complicate manual delineation. The additional prediction of anatomical structures is of particular importance for bedside CXRs, which are typically obtained under poor image acquisition conditions, with patients having various pathologies and invasive therapeutic devices, complicating the visualization of anatomy.

Our study has limitations. The different definition of anatomical areas used in this study limits the comparability with previous work. Especially the hidden areas of the lungs are chal-

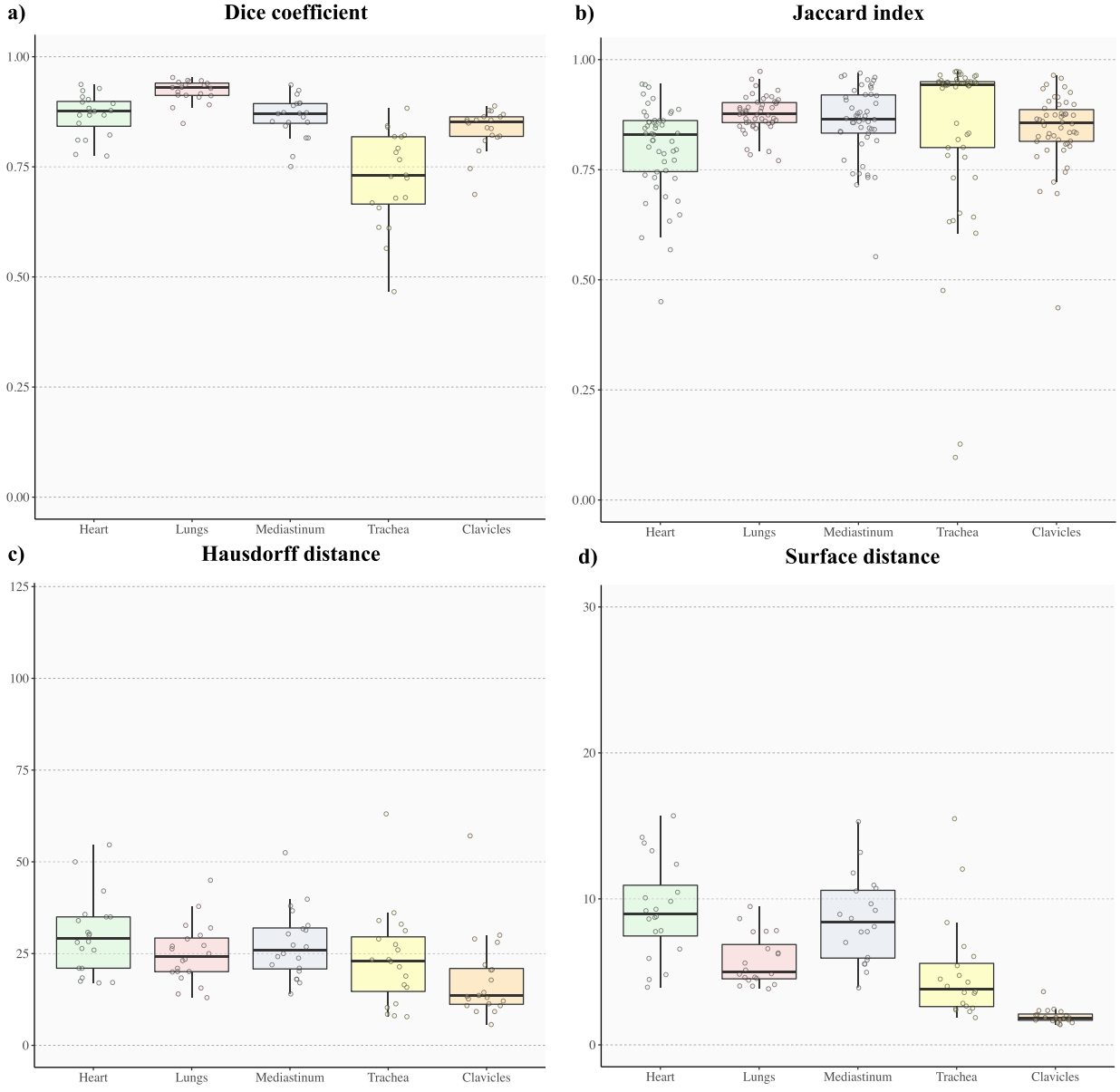

Fig. 5. Overview of the performance on the external test dataset. Boxplots displaying the distribution of the Dice similarity coefficient (a), Jaccard index (b), Hausdorff distance in mm (c), and average symmetric surface distance in mm (d) for the predicted segmentation labels of the heart (green plots), lungs (red plots), mediastinum (blue plots), trachea (yellow plots), and clavicles (orange plots).

Table 4
Performance of the proposed model compared to current state-of-the-art deep learning multi-label segmentation approaches for at least three anatomical structures in chest radiographs. Bold values show the highest performance for each structure.

Model	Dataset	Heart		Lungs		Mediastinum		Trachea		Clavicles	
		DSC	JI	DSC	JI	DSC	JI	DSC	JI	DSC	JI
U-ResNet-5 (proposed model)	Non-public[#]	0.88	0.80	0.93	0.88	**0.92**	**0.86**	**0.90**	**0.85**	0.91	0.84
Attention UW-Net [22]	ChestX-ray8[†] [23]	0.81	0.68[*]	0.96	0.92[*]	–	–	0.81	0.68[*]	0.78	0.63[*]
UNet_ResNeXt50_Masks+Contours [24]	JSTR[‡] [25]	**0.97**[*]	**0.93**	**0.99**[*]	**0.97**	–	–	–	–	**0.95**[*]	**0.90**
InvertedNet with ELU [26]	JSTR[‡] [25]	0.94	0.88	0.97	0.95	–	–	–	–	0.93	0.87
U-Net (VGG16) [27]	JSTR[‡] [25]	0.95	0.91	0.98	0.96	–	–	–	–	0.92	0.96
Multi-task FCN [28]	JSTR[‡] [25]	0.95[*]	0.90	0.98[*]	0.96	–	–	–	–	0.93[*]	0.86
X-Net+ single-class [29]	JSTR[‡] [25]	0.94	0.88	0.98	0.96	–	–	–	–	0.94	0.88
SegNet [30]	JSTR[‡] [25]	0.94	0.90	0.98	0.96	–	–	–	–	0.93	0.87
U-Net single-class [31]	JSTR[‡] [25]	0.95	0.90[*]	0.98	0.96[*]	–	–	–	–	0.94	0.88[*]
nnU-Net [32]	JSTR[‡] [25]	0.95	0.90[*]	0.98	0.96[*]	–	–	–	–	**0.95**	**0.90**[*]

Notes: For authors introducing multiple approaches, the model with the best overall performance was chosen. DSC = Dice similarity coefficient. JI = Jaccard index.
[*] = Values were calculated from the given metric.
[-] = Not included for segmentation.
[#] = Consisting of 2,000 anterior-posterior bedside chest radiographs.
[†] = A subset of 200 posterior-anterior standing chest radiographs was used.
[‡] = Consisting of 247 posterior-anterior standing chest radiographs.

lenging to segment for both human experts and the model, as no clear border is visible. This likely led to poorer performance metrics in our approach and limited comparability with published data. Furthermore, we exclusively used bedside CXRs, in which the anatomical structures are more often overlaid with pathologies, and patients are more commonly in atypical positions, further limiting the comparability with previous work. Third, there was just one ground-truth segmentation per image. However, multiple annotations by different readers would have been desirable to create more accurate segmentations and calculate inter-rater agreement. Lastly, our external test dataset of 20 images is relatively small, which might make results more prone to random variation.

5. Conclusions

In conclusion, the contribution of our model to existing architectures can be summarized as follows:

First, our approach simultaneously segments several key anatomical structures in bedside CXRs, including the heart, lungs, mediastinum, trachea, and clavicles, extending previous multi-label segmentation models. Second, our approach is almost similar in performance to existing state-of-the-art multi-label architectures, although it was developed on ICU AP bedside CXRs, including images acquired under challenging patient acquisition conditions, as well as heterogeneous imaging and disease characteristics. Third, our model approximates the true lung and heart size instead of only segmenting their non-overlapping portions, providing a more complex overall architecture for anatomical prediction. Finally, this study provides evidence for the effectiveness of combining computational human-in-the-loop approaches with active learning, allowing for time- and cost-efficient use of human annotators.

Data statement

Research data will be shared by the corresponding author upon reasonable request and in accordance with local data protection guidelines.

Funding statement

This study was funded by a grant from the Berlin Institute of Health (BIH) at Charité – Digital Health Accelerator. The funding did not affect the study design, the collection, analysis, and interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication.

Declaration of Competing Interest

SMN declares a relationship (honorary speaker) with the following companies: Canon Medical Systems, Bracco Imaging, and Teleflex. All other authors declare no conflict of interest.

Acknowledgments

LCA is grateful for her participation in the BIH Charité Junior Clinician and Clinician Scientist Program. KKB is grateful for his participation in the BIH Charité Digital Clinician Scientist Program, all funded by the Charité – Universitätsmedizin Berlin and the BIH.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107505.

References

[1] T. Bansal, R. Beese, Interpreting a chest X-ray, Br. J. Hosp. Med. (Lond) 80 (2019) C75–C79.
[2] W. Pezzotti, Chest X-ray interpretation: not just black and white, Nursing (Brux) 44 (2014) 40–47 quiz 47-48.
[3] R.P. Mathew, T. Alexander, V. Patel, G. Low, Chest radiographs of cardiac devices (Part 1): lines, tubes, non-cardiac medical devices and materials, SA J Radiol. 23 (2019) 1729.
[4] A.N. Rubinowitz, M.D. Siegel, I. Tocino, Thoracic imaging in the ICU, Crit. Care Clin. 23 (2007) 539–573.
[5] A. Ganapathy, N.K. Adhikari, J. Spiegelman, D.C. Scales, Routine chest x-rays in intensive care units: a systematic review and meta-analysis, Crit. Care 16 (2012) R68.
[6] C.I. Henschke, D.F. Yankelevitz, A. Wand, S.D. Davis, M. Shiau, Chest radiography in the ICU, Clin. Imaging 21 (1997) 90–103.
[7] J.G. Nam, M. Kim, J. Park, E.J. Hwang, J.H. Lee, J.H. Hong, J.M. Goo, C.M. Park, Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs, Eur. Respir. J. (2021) 57.
[8] Y. Khurana, U. Soni, Leveraging deep learning for COVID-19 diagnosis through chest imaging, Neural Comput. Appl. (2022) 1–10.
[9] M. Oloko-Oba, S. Viriri, A Systematic Review of Deep Learning Techniques for Tuberculosis Detection From Chest Radiograph, Front Med (Lausanne) 9 (2022) 830515.
[10] S.M. Niehues, L.C. Adams, R.A. Gaudin, C. Erxleben, S. Keller, M.R. Makowski, J.L. Vahldiek, K.K. Bressem, Deep-Learning-Based Diagnosis of Bedside Chest X-ray in Intensive Care and Emergency Medicine, Invest. Radiol. 56 (2021) 525–534.
[11] R.D.E. Henderson, X. Yi, S.J. Adams, P. Babyn, Automatic Detection and Classification of Multiple Catheters in Neonatal Radiographs with Deep Learning, J. Digit. Imaging 34 (2021) 888–897.
[12] B. van Ginneken, M.B. Stegmann, M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, Med. Image Anal. 10 (2006) 19–40.
[13] S. Soffer, A. Ben-Cohen, O. Shimon, M.M. Amitai, H. Greenspan, E. Klang, Convolutional Neural Networks for Radiologic Images: a Radiologist's Guide, Radiology 290 (2019) 590–606.
[14] M. Sarıgül, B.M. Ozyildirim, M. Avci, Differential convolutional neural network, Neural Netw. 116 (2019) 279–287.
[15] R. Hosch, L. Kroll, F. Nensa, S. Koitka, Differentiation Between Anteroposterior and Posteroanterior Chest X-Ray View Position With Convolutional Neural Networks, Rofo 193 (2021) 168–176.
[16] I.M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, A. Saalbach, Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification, Sci. Rep. 9 (2019) 6381.
[17] G. Chassagnon, M. Vakalopoulou, N. Paragios, M.P. Revel, Artificial intelligence applications for thoracic imaging, Eur. J. Radiol. 123 (2020) 108774.
[18] J.C.Y. Seah, C.H.M. Tang, Q.D. Buchlak, X.G. Holt, J.B. Wardman, A. Aimoldin, N. Esmaili, H. Ahmad, H. Pham, J.F. Lambert, B. Hachey, S.J.F. Hogg, B.P. Johnston, C. Bennett, L. Oakden-Rayner, P. Brotchie, C.M. Jones, Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study, Lancet Digit Health 3 (2021) e496–e506.
[19] A. Diaz-Pinto, S. Alle, A. Ihsani, M. Asad, V. Nath, F. Pérez-García, P. Mehta, W. Li, H. Roth, T. Vercauteren, D. Xu, P. Dogra, S. Ourselin, A. Feng, M.J. Cardoso, MONAI Label: a framework for AI-assisted Interactive Labeling of 3D Medical Images, 2022.
[20] S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, C. Kelly, A. Karthikesalingam, C. Chu, D. Carnell, C. Boon, D. D'Souza, S.A. Moinuddin, B. Garie, Y. McQuinlan, S. Ireland, K. Hampton, K. Fuller, H. Montgomery, G. Rees, M. Suleyman, T. Back, C.O. Hughes, J.R. Ledsam, O. Ronneberger, Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: deep Learning Algorithm Development and Validation Study, J. Med. Internet Res. 23 (2021) e26151.
[21] T.T. Tanimoto, An Elementary Mathematical Theory of Classification, Int. Business Mach. Corporation (1958) 10.
[22] D. Pal, P.B. Reddy, S. Roy, Attention UW-Net: a fully connected model for automatic segmentation and annotation of chest X-ray, Comput. Biol. Med. 150 (2022) 106083.
[23] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-Ray8: hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3462–3471.
[24] M. Kholiavchenko, I. Sirazitdinov, K. Kubrak, R. Badrutdinova, R. Kuleev, Y. Yuan, T. Vrtovec, B. Ibragimov, Contour-aware multi-label chest X-ray organ segmentation, Int. J. Comput. Assist. Radiol. Surg. 15 (2020) 425–436.
[25] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, 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, AJR Am. J. Roentgenol. 174 (2000) 71–74.
[26] A.A. Novikov, D. Lenis, D. Major, J. Hladůvka, M. Wimmer, K. Bühler, Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs, IEEE Trans Med Imaging 37 (2018) 1865–1876.

[27] M. Frid-Adar, A. Ben-Cohen, R. Amer, H. Greenspan, Improving the Segmentation of Anatomical Structures in Chest Radiographs Using U-Net with an ImageNet Pre-trained Encoder, in: D. Stoyanov, Z. Taylor, B. Kainz, G. Maicas, R.R. Beichel, A. Martel (Eds.), Image Analysis for Moving Organ, Breast, and Thoracic Images, Springer International Publishing, Cham, 2018, pp. 159–168.
[28] C. Wang, Segmentation of Multiple Structures in Chest Radiographs Using Multi-task Fully Convolutional Networks, in: P. Sharma, F.M. Bianchi (Eds.), Image Analysis, Springer International Publishing, Cham, 2017, pp. 282–289.
[29] O. Gómez, P. Mesejo, O. Ibáñez, A. Valsecchi, O. Cordón, Deep architectures for high-resolution multi-organ chest X-ray image segmentation, Neural Comput. Appl. 32 (2020) 15949–15963.
[30] H. Oliveira, J.d. Santos, Deep Transfer Learning for Segmentation of Anatomical Structures in Chest Radiographs, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2018, pp. 204–211.
[31] E.R.C.Q. Brioso, Anatomical Segmentation in Automated Chest Radiography Screening, Bioengineering, FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO, 2022, pp. 64.
[32] N. Gaggion, M. Vakalopoulou, D.H. Milone, E. Ferrante, Multi-center anatomical segmentation with heterogeneous labels via landmark-based models, ArXiv (2022) abs/2211.07395.
[33] S. Jaeger, S. Candemir, S. Antani, Y.X. Wáng, P.X. Lu, G. Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, Quant. Imaging Med. Surg. 4 (2014) 475–477.
[34] S. Candemir, S. Antani, A review on lung boundary detection in chest X-rays, Int. J. Comput. Assist. Radiol. Surg. 14 (2019) 563–576.