Physica Medica 130 (2025) 104911

Contents lists available at ScienceDirect

Physica Medica

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

# Development and validation of fully automated robust deep learning models for multi-organ segmentation from whole-body CT images

Yazdan Salimi [a,1], Isaac Shiri [a,b,1], Zahra Mansouri [a], Habib Zaidi [a,c,d,e,*]

[a] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital CH-1211 Geneva, Switzerland
[b] Department of Cardiology, Inselspital, Bern University Hospital, University of Bern, Switzerland
[c] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[d] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark
[e] University Research and Innovation Center, Óbuda University, Budapest, Hungary

ARTICLE INFO

*Keywords:*
Computed Tomography
Segmentation
Organs at Risk
Deep Learning
Computational Models

ABSTRACT

*Purpose:* This study aimed to develop a deep-learning framework to generate multi-organ masks from CT images in adult and pediatric patients.
*Methods:* A dataset consisting of 4082 CT images and ground-truth manual segmentation from various databases, including 300 pediatric cases, were collected. In strategy#1, the manual segmentation masks provided by public databases were split into training (90%) and testing (10% of each database named subset #1) cohort. The training set was used to train multiple nnU-Net networks in five-fold cross-validation (CV) for 26 separate organs. In the next step, the trained models from strategy #1 were used to generate missing organs for the entire dataset. This generated data was then used to train a multi-organ nnU-Net segmentation model in a five-fold CV (strategy#2). Models' performance were evaluated in terms of Dice coefficient (DSC) and other well-established image segmentation metrics.
*Results:* The lowest CV DSC for strategy#1 was $0.804 \pm 0.094$ for adrenal glands while average DSC $> 0.90$ were achieved for 17/26 organs. The lowest DSC for strategy#2 ($0.833 \pm 0.177$) was obtained for the pancreas, whereas DSC $> 0.90$ was achieved for 13/19 of the organs. For all mutual organs included in subset #1 and subset #2, our model outperformed the TotalSegmentator models in both strategies. In addition, our models outperformed the TotalSegmentator models on subset #3.
*Conclusions:* Our model was trained on images with significant variability from different databases, producing acceptable results on both pediatric and adult cases, making it well-suited for implementation in clinical setting.

## 1. Introduction

Segmentation of healthy organs from Computed Tomography (CT) images is critical and beneficial in a number of applications, including the generation of anthropomorphic computational models, delineation of organs at risk in radiation therapy (RT) treatment planning [1–4], and other computer-assisted applications, such as pathologic detection [5,6], prognosis and outcome prediction [7–10], image quantification [11–13], and radiation dosimetry calculations [14–17]. The manual slice-by-slice segmentation of organs is labor-intensive and time-consuming, in addition to the high inter- and intra-observer variability reported for the segmentation of healthy organs and malignant lesions [18,19]. Since the emergence of machine learning and deep learning (DL) algorithms in medical imaging research, especially medical image segmentation, a number of studies focused on the automatic segmentation of structures from CT images and other imaging modalities [20–23]. Most published studies attempted to improve segmentation accuracy (commonly quantified by the Dice coefficient), robustness, and generalizability on new unseen datasets acquired with different imaging settings on disparate patient characteristics and including a large number of organs [24–26].

Newly developed neural network architectures, loss functions, and image processing algorithms contributed to the improvement of the performance of image segmentation models [23]. Yet, the number of datasets and their diversity remains the bottleneck for successful implementation of DL-based algorithms [27]. Most studies conveyed the

* Corresponding author at: Geneva University Hospital, Division of Nuclear Medicine and Molecular Imaging, CH-1211 Geneva, Switzerland.
*E-mail address:* habib.zaidi@hcuge.ch (H. Zaidi).
[1] Yazdan Salmi and Isaac Shiri contributed equally to this work.

https://doi.org/10.1016/j.ejmp.2025.104911
Received 13 August 2024; Received in revised form 2 December 2024; Accepted 24 January 2025
Available online 2 February 2025

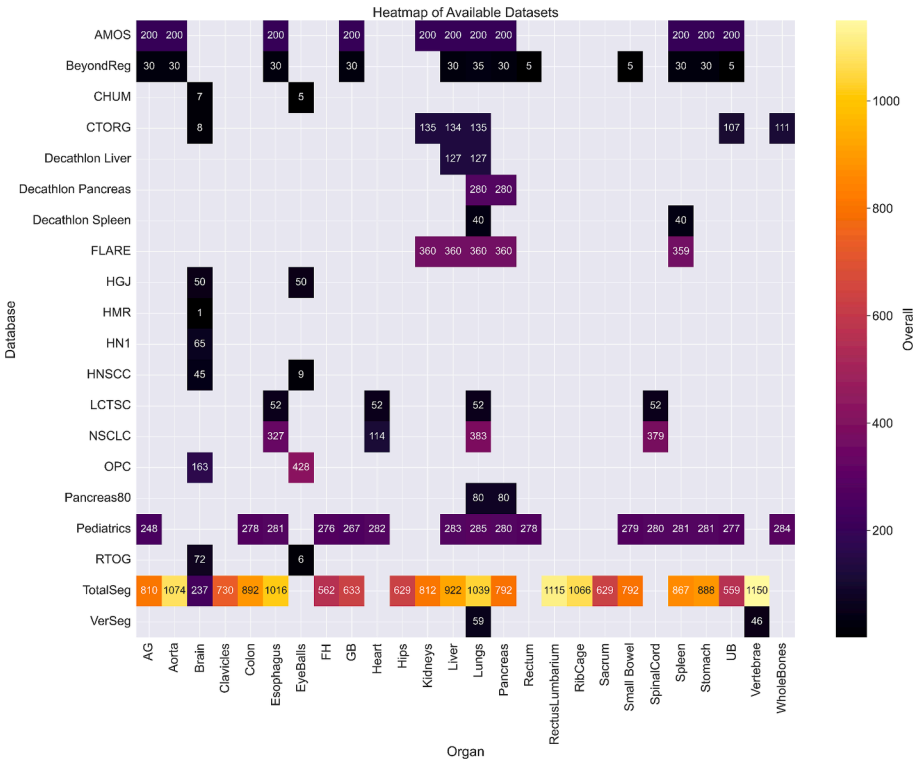


**Fig. 1.** Number of 3D CT images extracted from the clinical studies included in all databases used for training and testing of the models. UB: Urinary Bladder, GB: Gall Bladder, AG: Adrenal Gland, FH: femoral heads. The color bar is presented on right side.

performance of the developed models on a test set excluded from the training set, thus reaching very high Dice coefficients (DSCs) as reported in a few challenges held on multiple organ segmentations [28]. Yet, the majority of these studies didn't investigate models' performance on unseen external datasets. Xu et al. [29] focused on the occurrence of outliers during image segmentation and how to solve this problem. Since the emergence of machine learning and deep learning (DL) algorithms in medical imaging research, especially medical image segmentation, several studies focused on the automatic segmentation of organs/tissues from CT images and other imaging modalities [20–23,30]. Bordigoni et al. [31] demonstrated the potential of automated segmentation models to save time in delineating pelvic area lesions and organs at risk. Tong et al. [32] highlighted the potential of achieving higher segmentation accuracy by using a two-stage model for localization and segmentation. Recent studies addressed the limitations and benefits of DL-based organ segmentation in real-life clinical scenarios [18,33,34]. The comparison of the results achieved by different techniques using private/local databases is not straightforward given that the used datasets are not publicly available. Besides, it's well established that acquisition, scanner, and demographic parameters can affect the performance of a model on external unseen datasets from other centers [18,35,36]. Ma *et al.* [24] described the low performance of segmentation models trained and inferenced on different databases for abdominal organs segmentation task. In this context, a segmentation model trained on a dataset presenting with large variability and tested on an unseen dataset may be beneficial in estimating the performance in real clinical scenarios. Pediatric organ segmentation may be challenging because of differences in organs' shape, size, and texture compared to adults as well as lower image quality in low-dose pediatric protocols. To the best of our knowledge, there is no multiple organ segmentation tool available dedicated to pediatric patients.

In this study, we aimed to develop deep neural network models to segment multiple healthy organs from total-body CT images targeting improvement of the accuracy and generalizability compared to previously developed models. We also compared the performance of our models with existing methods. We included a large data set, incorporating as many publicly available datasets as possible to train and test our models. However, each database provided segmentation masks for only a limited number of organs. To address this limitation, we generated a comprehensive, unified dataset containing segmentations for all organs. Using this dataset, we trained a new, fast inference model aimed at achieving robust models with superior performance on both internal and external test sets. Additionally, we provided a large versatile dataset which includes multiple organ segmentation masks of both adult and pediatric patients with various pathologies, that can be used for further research.

## 2. Materials and methods

### 2.1. Patient population

After excluding cases with segmentation errors or data conversion issues through visual assessment, this study included 4082 CT images (971,361 axial 2D CT slices, 51,058 3D image/segment pairs) collected from multiple online available datasets [37–43]. Of the 4082 cases, 300 cases were pediatric patients with 18.9 ± 4.13 cm effective diameter, while the rest were adult cases with 27.53 ± 5.35 cm effective diameter as defined by the AAPM #204 Report [44]. The average age was 6.32 ± 4.34 years for pediatric patients and 66.98 ± 9.84 years for adult patients. The age, gender, and acquisition parameters were available only in a limited group of datasets; the rest were either anonymized or in NIFTI format without additional information. Fig. 1 presents the number of images for each organ from different databases. These datasets included segmentation masks for 26 different organs, including adrenal glands (AG), aorta, brain, clavicles, colon, esophagus, eyeballs, femoral heads (FH), gall bladder (GB), hips, sacrum, kidneys, liver, lungs, pancreas, rectum, rectus lumbarium, ribs, small bowel, spinal cord, spleen, stomach, urinary bladder (UB), vertebrae, whole bones (including all bones in the field-of-view), and heart. Paired organs, such as kidneys were combined and considered as a single segmentation

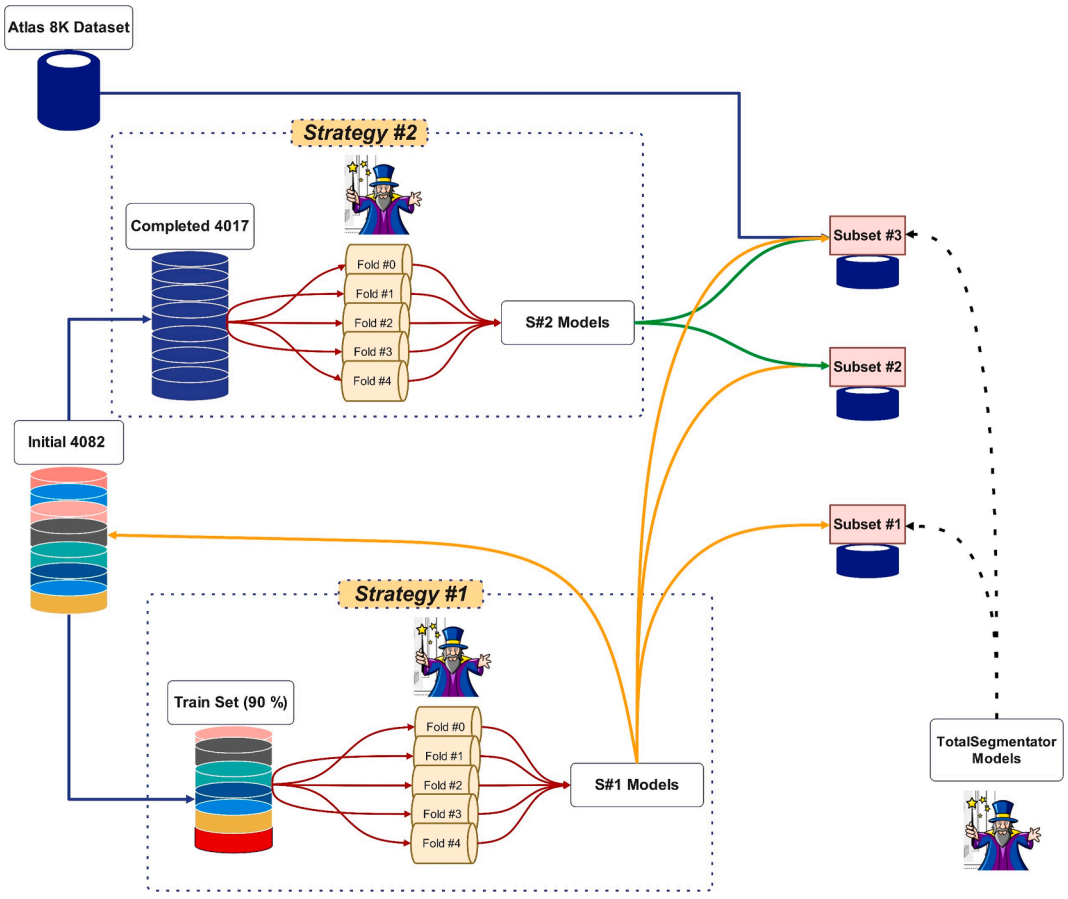


**Fig. 2.** Dataflow and data split strategies adopted in this study. Blue lines represent data transfer, orange lines show strategy#1 inference, and green lines indicate strategy#2 inference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mask.

Two strategies were implemented in this study:

1. **Strategy #1:** This strategy included manual segmentation masks provided by the public dataset as standard of reference. Every database has a limited number of organs delineated manually. Hence, the number of training data for each organ varies as shown in Fig. 1. First, the test portion of the dataset provided by Wasserthal *et al.* [43] (named Subset #2, containing 65 CT images) was excluded from the training data. the remainder of the TotalSegmentator dataset was used for training, following the data split reported by Wasserthal *et al.* [43] to ensure a fair comparison with their study. The remaining data from the other databases were randomly split into training (cross-validation) and testing sets, with 90 % allocated to training and 10 % to testing (10 % of each database, denoted as Subset #1). A single-organ, the five-fold nnU-Net model, was then trained for each organ. At the end, the five-fold trained models were tested on Subset #1 test set, performing external validation by ensembling all five folds for each test image. It should be noted that TotalSegmentator images were not included in the Subset #1 test set. Hence, organs for which training data were only available from the TotalSegmentator database were absent in Subset #1. Additionally, the number of images available for each organ varied, resulting in a different number of tests for each organ in Subset #1. The trained five-fold nnU-Net models were used to complete the dataset by generating the missing organ segmentations on all 4017 CT images (4082 images minus the 65 test sets from the TotalSegmentator database). This completed dataset was then used to train a unified nnU-Net model, as described below in strategy#2.
2. **Strategy#2:** The entire dataset generated by trained models from strategy#1 was used to train a single nnU-Net model for a select number of organs with reliable performance. These organs included the aorta, brain, clavicles, colon, esophagus, eyeballs, femoral heads, gall bladder, hips, sacrum, liver, lungs, pancreas, ribs, spleen, stomach, urinary bladder, vertebrae, and heart. First, all segmentation masks were combined to create a multi-label mask without any overlaps. In cases where a voxel was segmented by multiple models, the model with the higher DSC in strategy#1 was selected. For example, if a single voxel was segmented by both liver and colon models, it was considered as liver mask. The training data for this strategy was larger than that for strategy#1, as it included all 4017 images (from a total of 4082 images, excluding Subset #2). Finally, models trained in strategy#2 were tested on Subset #2, consisting of 65 CT images, by ensembling all five models on each test image.

In summary, strategy #1 models were tested on both the Subset #1 test set and the Subset #2 images as indicated by Wasserthal *et al.* [43]. Conversely, models from strategy#2 were trained using all available data with cross-validation and were tested only on Subset #2.

### 2.2. Network architecture

The same network architecture and training hyperparameters were used for strategy#1 and strategy#2. The self-configuring nnU-Net [45] pipeline was used with a five-fold cross-validation data split, employing the default 3D-fullres hyperparameters. However, the training length was increased from 1000 epochs to 2000 epochs to achieve an improved accuracy. The default hyperparameters are an initial learning rate of 1e-2 decreased at every epoch by the decay of 3e-5 and Dice cross-entropy

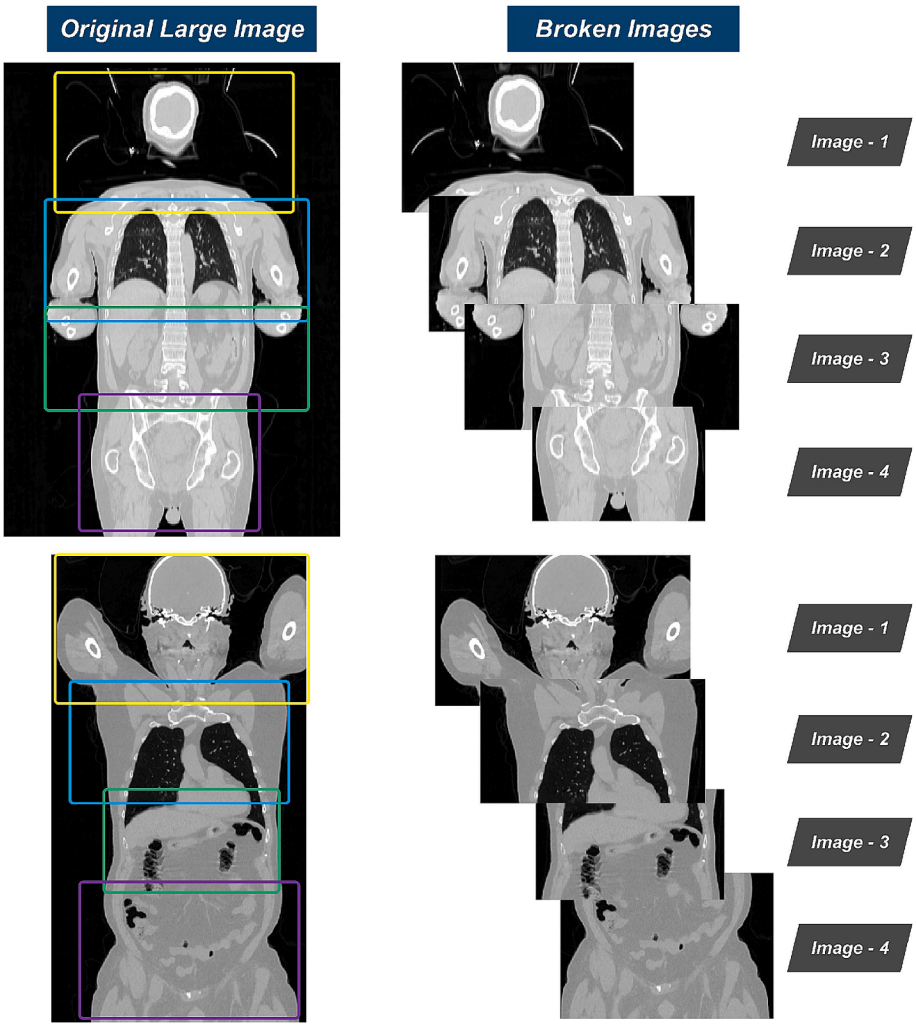


**Fig. 3.** Illustration of breaking down process. Each colored box shows the coverage of a single broken image. Note the overlap between the images and the smaller cropped area based on the body shape and arms position.

loss function. All CT images and segmentation masks were cropped to the foreground using automated body contour detection. This process employed analytical object detection algorithms previously utilized in our study [36,46]. This step reduces the image loading burden and time as well as the training and inference time. Fig. 2 shows the flowchart and dataflow adopted in this study.

### 2.3. Performance comparison and benchmarking

To evaluate the performance of our model in real clinical scenarios and compare it with previously reported DL models, we tested both strategy#1 and strategy#2 models on Subset #2, which was unseen during training for both strategies. Additionally, the models published by Wasserthal *et al.* [43] were collected from the TotalSegmentator (GitHub page on July 27th, 2024) and tested on Subset #1. We compared the performance of our models on their dataset (subset #2) with the performance of their models on our test set (subset #1) to ensure a fair comparison. It should be mentioned that this comparison focused on training data and preprocessing and data cleaning steps, not on the training algorithm and network architecture, which were similar in both studies. Finally, we tested both strategy#1 and strategy#2 models as well as models from Wasserthal *et al.* [43] on a limited sample of data from Qu *et al.* [47] study, referred to as Subset #3. We compared the models' outputs to the reference segmentation provided. For practical reasons, such as inference time and computational limits and visual assessment time, we used only the first 200 images from Qu *et al.* [47]. Some cases with errors in the reference segmentation were excluded from the comparison. Supplementary Fig. 1 shows a few examples of three excluded cases with errors in the reference segmentation. Subset #3 includes segmentation masks for seven organs, including the aorta, GB, kidneys, liver, pancreas, spleen, and stomach. The right and left kidney masks were combined into a single kidneys' segmentation.

### 2.4. Inference facilitation

The nnU-Net 3D-Fulress configuration model uses 3D image patches for training and applies sliding windows inference using patches of the same size during training. Depending on the size of test images, the required amount of device random access memory (RAM) could vary, potentially slowing down the inference process on PCs with limited RAM. To address this issue, in addition to crop to foreground process, we proposed a solution that involves breaking the test images into smaller images along the cranio-caudal (Z) axis and feeding these smaller segments to the nnU-Net model. We included some overlap between the images to prevent a possible performance drop due to lack of information about the voxel neighborhood at the borders.

The body contour cropping is performed for each broken axial image, further reducing the image size and speeding up the inference time. In the end, the effectiveness of this approach was compared with the usual inference method. Fig. 3 shows the schematic explanation of breaking

**Table 1**
Cross-validation Dices coefficients for both strategy#1 and strategy#2. NI: Not included in strategy#2.

| Organ | strategy#1 Dice | strategy#2 Dice |
|---|---|---|
| AG | 0.804 ± 0.094 | NI |
| Aorta | 0.95 ± 0.086 | 0.953 ± 0.088 |
| Brain | 0.964 ± 0.114 | 0.857 ± 0.132 |
| Clavicles | 0.962 ± 0.078 | 0.942 ± 0.088 |
| Colon | 0.888 ± 0.114 | 0.908 ± 0.103 |
| Esophagus | 0.867 ± 0.108 | 0.895 ± 0.094 |
| Eyeballs | 0.928 ± 0.058 | 0.908 ± 0.062 |
| FH | 0.939 ± 0.117 | 0.941 ± 0.124 |
| GB | 0.817 ± 0.192 | 0.87 ± 0.172 |
| Hips | 0.976 ± 0.089 | 0.949 ± 0.116 |
| Sacrum | 0.937 ± 0.135 | 0.894 ± 0.195 |
| Kidneys | 0.937 ± 0.067 | NI |
| Liver | 0.966 ± 0.074 | 0.968 ± 0.070 |
| Lungs | 0.955 ± 0.068 | 0.975 ± 0.083 |
| Pancreas | 0.882 ± 0.164 | 0.833 ± 0.177 |
| Rectum | 0.85 ± 0.148 | NI |
| Rectus Lumbarium | 0.966 ± 0.057 | NI |
| Ribs | 0.963 ± 0.054 | 0.920 ± 0.077 |
| Small Bowel | 0.871 ± 0.123 | NI |
| Spinal Cord | 0.883 ± 0.067 | NI |
| Spleen | 0.944 ± 0.105 | 0.956 ± 0.110 |
| Stomach | 0.927 ± 0.103 | 0.940 ± 0.096 |
| UB | 0.899 ± 0.184 | 0.853 ± 0.196 |
| Vertebrae | 0.966 ± 0.074 | 0.953 ± 0.08 |
| Whole Bones | 0.974 ± 0.046 | NI |
| Heart | 0.952 ± 0.117 | 0.909 ± 0.154 |

down images during RAM-friendly inference. Users can adjust parameters, such as broken image size and the overlap between the broken images to optimize models' performance based on available RAM. Besides, cropping to the foreground process occurs independently for each image part, resulting in more precise cropping and reduced background area as shown in Fig. 3. To evaluate the performance drops due to incomplete coverage and loss of neighborhood information, both strategy#1 and strategy#2 models were tested on subset #1 data. Finally, we compared the dice coefficients of nnU-Net models with and without breaking down images, testing with no overlap and with 5 cm overlap. We used in house developed body contouring analytic algorithm to generate body contour on CT images. The input image is first oriented to ensure that the 2D extracted images correspond to axial images. Then the algorithm goes through each axial slice and through multiple adaptive thresholding methods based on CT image Hounsfield units extracts the area inside the patient's body. In the next step, this algorithm fills the holes inside the body segmented area to fill the smaller areas inside body containing air, such as the trachea. Then, it removes the object islands in the segmented body having a volume smaller than a certain threshold to keep the arms and legs segmentation in the segmented area and remove other objects. The final step consists in concatenating the 2D axial segmentations and keeping the largest connected component (island) to ensure removing every external object, such as respiratory device or blanket. The algorithm is available on GitHub in MatLab and Python programming languages (https://github.com/YazdanSalimi/Organ-Segmentation).

### 2.5. Evaluation metrics

Common segmentation evaluation metrics including Dice and Jaccard coefficients, mean surface distance, Hausdorff distance, and the segment volume difference were used to compare the nnU-Net generated outputs versus the standard of reference segmentations.

### 2.6. Statistical analysis

First, the normality of data distribution was evaluated through the Kolmogorov Smirnov test. We then used the Wilcoxon rank *t*-test to compare the performance metrics between different groups and models. Two-tailed P-values less than 0.05 were considered statistically significant.

## 3. Results

### 3.1. Cross-validation

The lowest cross-validation DSC for strategy #1 was 80.41 ± 9.4 for adrenal glands, while an average DSC greater than 0.90 was achieved for 17 out of 26 organs. For strategy#2, the lowest DSC was 83.3 ± 17.7 for the pancreas, with a DSC greater than 0.90 achieved for 13 out of 19 organs. The P-value comparing the DSC and Jaccard indices between strategy#1 and strategy#2 during cross-validation training was less than 0.05. In strategy#2, the DSC was higher than in strategy#1 for 9 out of 19 organs but lower for 10 out of 19 organs. It should be noted that the reference segmentations for strategy#2 were generated by DL and were not verified visually. Detailed DSC for each organ for both strategies are summarized in Table 1. The inference time was reduced

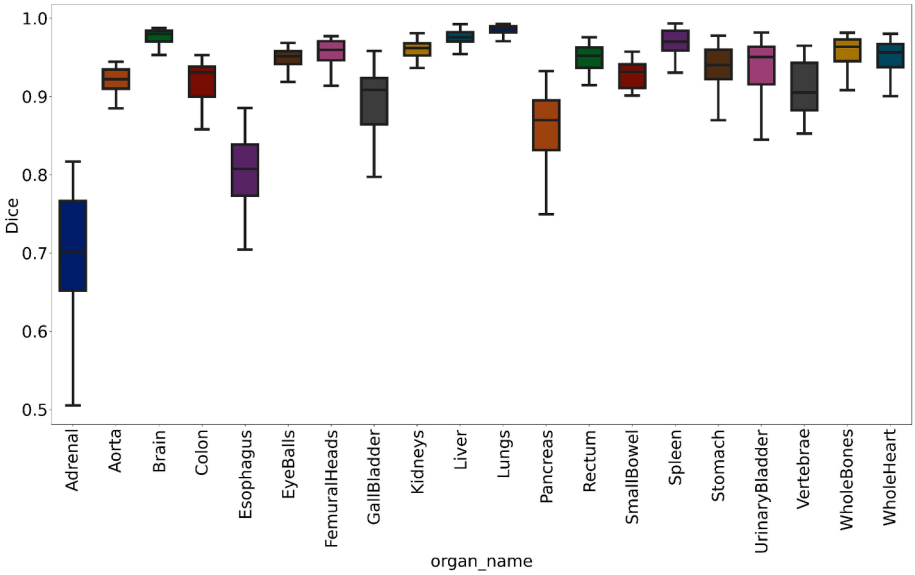


**Fig. 4.** Box plot of Dice coefficients for strategy#1 model tested on Test-10% internal test set.

**Table 2**
Internal test on 10% separate test set performance metrics for strategy #1. MSD: Mean Surface Distance, HD: Hausdorff Distance, VD: Volume Difference.

| Organ | # of tests | Dice coefficient | Jaccard | MSD (mm) | HD (mm) | VD (ml) |
|---|---|---|---|---|---|---|
| AG | 47 | 0.698 ± 0.083 | 0.542 ± 0.095 | 3.508 ± 13.69 | 24.153 ± 88.521 | −1.171 ± 1.3 |
| Aorta | 23 | 0.922 ± 0.016 | 0.856 ± 0.028 | 0.433 ± 0.41 | 3.144 ± 6.09 | −2.14 ± 6.17 |
| Brain | 37 | 0.961 ± 0.067 | 0.931 ± 0.099 | 0.666 ± 0.99 | 3.785 ± 6.826 | −4.07 ± 48.999 |
| Colon | 27 | 0.913 ± 0.05 | 0.843 ± 0.075 | 1.044 ± 1.736 | 7.546 ± 13.694 | −15.554 ± 29.59 |
| Esophagus | 88 | 0.791 ± 0.077 | 0.66 ± 0.094 | 1.094 ± 0.931 | 7.851 ± 9.09 | −1.116 ± 5.721 |
| Eyeballs | 47 | 0.947 ± 0.018 | 0.901 ± 0.03 | 0.181 ± 0.065 | 1.03 ± 0.293 | −0.119 ± 0.886 |
| FH | 27 | 0.956 ± 0.018 | 0.915 ± 0.032 | 0.235 ± 0.129 | 1.155 ± 0.78 | −2.351 ± 5.529 |
| GB | 48 | 0.885 ± 0.061 | 0.799 ± 0.092 | 1.616 ± 2.697 | 19.873 ± 37.667 | −1.833 ± 2.89 |
| Kidneys | 69 | 0.947 ± 0.055 | 0.903 ± 0.081 | 0.887 ± 2.24 | 6.174 ± 16.935 | −20.163 ± 115.32 |
| Liver | 112 | 0.972 ± 0.031 | 0.947 ± 0.048 | 1.837 ± 11.967 | 14.587 ± 81.769 | 16.018 ± 67.801 |
| Lungs | 227 | 0.966 ± 0.06 | 0.94 ± 0.094 | 0.455 ± 1.28 | 4.073 ± 17.362 | 19.85 ± 128.435 |
| Pancreas | 123 | 0.851 ± 0.066 | 0.746 ± 0.091 | 1.04 ± 2.213 | 7.192 ± 25.134 | −6.674 ± 16.281 |
| Rectum | 27 | 0.945 ± 0.028 | 0.896 ± 0.048 | 0.304 ± 0.291 | 2.128 ± 4.8 | −0.68 ± 1.537 |
| Small Bowel | 27 | 0.924 ± 0.033 | 0.86 ± 0.052 | 0.816 ± 0.463 | 5.227 ± 3.242 | 2.168 ± 32.223 |
| Spinal Cord | 71 | 0.892 ± 0.066 | 0.81 ± 0.097 | 0.887 ± 1.974 | 5.985 ± 11.519 | 2.717 ± 11.647 |
| Spleen | 90 | 0.965 ± 0.048 | 0.935 ± 0.067 | 2.79 ± 24.457 | 12.727 ± 109.96 | 9.895 ± 93.143 |
| Stomach | 50 | 0.933 ± 0.044 | 0.877 ± 0.068 | 0.676 ± 0.601 | 4.761 ± 7.418 | −8.356 ± 13.326 |
| UB | 56 | 0.936 ± 0.04 | 0.882 ± 0.068 | 0.423 ± 0.374 | 1.816 ± 2.259 | −1.628 ± 12.622 |
| Vertebrae | 24 | 0.904 ± 0.045 | 0.828 ± 0.072 | 0.557 ± 0.673 | 5.443 ± 7.02 | 8.352 ± 24.02 |
| Whole Bones | 39 | 0.949 ± 0.035 | 0.905 ± 0.06 | 0.597 ± 0.893 | 5.608 ± 10.135 | 4.135 ± 147.958 |
| Heart | 44 | 0.944 ± 0.046 | 0.897 ± 0.072 | 0.768 ± 0.794 | 3.03 ± 2.529 | 3.312 ± 35.962 |

from 20 s per organ per image for five-fold inference in strategy #1 to 45 s per image for strategy #2, segmenting 19 organs on a PC with a Corei913900 KF CPU and NVIDIA RTX 4090 GPU with 32 GB of RAM. These times are dependent on PC configurations.

### 3.2. External evaluation on test-10 % (strategy#1)

An average DSC of 0.69, 0.92, 0.961, 0.913, 0.791, 0.947, 0.956, 0.885, 0.947, 0.972, 0.966, 0.851, 0.945, 0.924, 0.892, 0.965, 0.933, 0.936, 0.904, 0.949, and 0.944 were achieved on our 10 % separate test set for the AG, aorta, brain, colon, esophagus, eyeballs, FH, GB, kidneys, liver, lungs, pancreas, rectum, small bowel, spinal cord, spleen, stomach, UB, vertebrae, whole bones, and heart, respectively. Fig. 4 shows the box plot of DSCs for the mentioned organs. Supplementary Table 1 shows the comparison of segmentation metrics between adult and pediatric groups in subset #1.

The detailed performance metrics for the internal test set is summarized in Table 2.

### 3.3. Evaluation on external datasets

Table 3 summarizes the detailed DSC and Jaccard values comparing our model to Wasserthal *et al.* [43] models on three external test sets of Subset #1, #2, and #3. The DSC of our models tested on subset #2 were higher than those of TotalSegmentator models tested on subset #1 for all 16 mutual organs. This pattern was the same for both strategy#1 and strategy#2 models, with the differences being statistically significant for all organs except the liver. strategy#2 where DSC were higher than strategy#1 only for the AG, GB, and spleen, while strategy#1 DSC were higher for the remaining organs.

For Subset #3, the DSC of both strategy#1 and strategy#2 models were higher than those of Wasserthal *et al.* [43] models for all seven included organs. This improvement was statistically significant for all organs, although it was less than 2 % DSC for 4 out of 7 organs, including the aorta, liver, spleen and stomach. The DSCs between strategy#1 and strategy#2 were comparable, with no statistically significant differences.

### 3.4. Breaking down images

There was a small difference in the nnU-Net output when breaking down the images with zero overlap, leading to different segmentation results. This drop resulted in different segmentation outputs with a DSC of 97.3 and 98.5 for strategy#1 and strategy#2 models, respectively. This change in the segmentation output was more significant in smaller organs, such as the AG. However, adding 5 cm overlap between the image regions resolved this issue, resulting in DSCs of more than 99.6 for both strategy#1 and strategy#2 models. It should be emphasized that the nnU-Net output with the same inference configuration but without breaking down images into parts were considered as the reference segmentation in this evaluation. The aim of this step was to evaluate the models' output reproducibility by breaking down the images. Python multiprocessing library can use parallel processing and decrease the calculation time using multiple threads or processes at the same time with a number of workers larger than 1. However, an additional number of workers requires more random access memory to perform well.

On the other hand, the inference time with nnU-Net pre-processing and saving a number of multiprocessing workers equal to one was the same. However, performing inference on smaller image parts allowed to select higher numbers of multiprocessing workers, resulting in inference times 20 % shorter than using the original image. The 20 % saved time was almost the same as the time spent for breaking down images into parts on a solid-state disk.

### 3.5. Merits of strategy#2 over strategy#1 on pediatric cases

Fig. 4 illustrates examples of pediatric images from the test set, segmented using models from strategy#1 and strategy#2. For older patients, both strategy#1 and strategy#2 produce nearly identical segmentations for mutual organs. However, for younger patients, such as those around one and three years old, depicted in Fig. 4, strategy#2 outperforms strategy#1 models, particularly in organs where manual segmentations are lacking. strategy#2 benefits from a large training dataset generated by strategy#1, although the initial training data for organs like the sacrum, hip, and clavicles came from a single online

**Table 3**
Summary of Dice coefficients and Jaccard indices comparing our models trained in strategy#1 and strategy#2 using Wasserthal et al. (44) models on multiple databases. Excluded: Organs with manual segmentations. NI: Not included or provided.

| | Strategy#1 on subset#2 | | Strategy#2 on subset#2 | | Wasserthal et al. on subset#1 | | Strategy#1 on subset#3 | | Strategy#2 on subset#3 | | Wasserthal et al. on subset #3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Organ** | **Dice** | **Jaccard** | **Dice** | **Jaccard** | **Dice** | **Jaccard** | **Dice** | **Jaccard** | **Dice** | **Jaccard** | **Dice** | **Jaccard** |
| AG | 0.907 ± 0.104 | 0.842 ± 0.133 | 0.969 ± 0.024 | 0.940 ± 0.042 | 0.531 ± 0.207 | 0.387 ± 0.188 | NI | NI | NI | NI | NI | NI |
| Aorta | 0.975 ± 0.018 | 0.952 ± 0.033 | 0.957 ± 0.046 | 0.921 ± 0.077 | 0.910 ± 0.017 | 0.835 ± 0.028 | 0.844 ± 0.113 | 0.665 ± 0.261 | 0.845 ± 0.113 | 0.685 ± 0.241 | 0.828 ± 0.118 | 0.723 ± 0.159 |
| Brain | 0.959 ± 0.037 | 0.923 ± 0.065 | 0.958 ± 0.011 | 0.919 ± 0.020 | 0.740 ± 0.365 | 0.693 ± 0.367 | NI | NI | NI | NI | NI | NI |
| Clavicles | 0.978 ± 0.010 | 0.957 ± 0.018 | 0.895 ± 0.169 | 0.836 ± 0.173 | Excluded | Excluded | NI | NI | NI | NI | NI | NI |
| Colon | 0.946 ± 0.036 | 0.899 ± 0.059 | 0.941 ± 0.034 | 0.889 ± 0.054 | 0.600 ± 0.220 | 0.460 ± 0.202 | NI | NI | NI | NI | NI | NI |
| Esophagus | 0.956 ± 0.019 | 0.917 ± 0.034 | 0.928 ± 0.171 | 0.898 ± 0.197 | 0.713 ± 0.114 | 0.565 ± 0.125 | NI | NI | NI | NI | NI | NI |
| FH | 0.954 ± 0.121 | 0.929 ± 0.153 | 0.906 ± 0.125 | 0.846 ± 0.161 | Excluded | Excluded | NI | NI | NI | NI | NI | NI |
| GB | 0.900 ± 0.155 | 0.844 ± 0.187 | 0.974 ± 0.058 | 0.953 ± 0.083 | 0.816 ± 0.089 | 0.697 ± 0.113 | 0.879 ± 0.085 | 0.637 ± 0.328 | 0.881 ± 0.084 | 0.662 ± 0.311 | 0.851 ± 0.095 | 0.751 ± 0.127 |
| Hips | 0.987 ± 0.011 | 0.974 ± 0.021 | 0.963 ± 0.015 | 0.934 ± 0.027 | Excluded | Excluded | NI | NI | NI | NI | NI | NI |
| Sacrum | 0.986 ± 0.005 | 0.972 ± 0.009 | 0.970 ± 0.017 | 0.943 ± 0.029 | Excluded | Excluded | NI | NI | NI | NI | NI | NI |
| Kidneys | 0.951 ± 0.092 | 0.917 ± 0.126 | NI | NI | 0.916 ± 0.101 | 0.855 ± 0.117 | 0.930 ± 0.073 | 0.848 ± 0.185 | NI | NI | 0.716 ± 0.156 | 0.582 ± 0.206 |
| Liver | 0.977 ± 0.036 | 0.958 ± 0.058 | 0.974 ± 0.058 | 0.953 ± 0.083 | 0.964 ± 0.012 | 0.931 ± 0.023 | 0.972 ± 0.024 | 0.936 ± 0.105 | 0.975 ± 0.011 | 0.951 ± 0.021 | 0.967 ± 0.015 | 0.937 ± 0.026 |
| Lungs | 0.992 ± 0.012 | 0.984 ± 0.022 | 0.992 ± 0.011 | 0.985 ± 0.022 | 0.944 ± 0.113 | 0.909 ± 0.140 | NI | NI | NI | NI | NI | NI |
| Pancreas | 0.881 ± 0.145 | 0.808 ± 0.167 | 0.856 ± 0.181 | 0.779 ± 0.200 | 0.748 ± 0.117 | 0.609 ± 0.134 | 0.873 ± 0.083 | 0.736 ± 0.195 | 0.868 ± 0.079 | 0.736 ± 0.182 | 0.839 ± 0.097 | 0.733 ± 0.121 |
| Rectus Lumbarium | 0.975 ± 0.038 | 0.953 ± 0.058 | NI | NI | Excluded | Excluded | NI | NI | NI | NI | NI | NI |
| Ribs | 0.980 ± 0.015 | 0.960 ± 0.028 | 0.934 ± 0.018 | 0.876 ± 0.031 | Excluded | Excluded | NI | NI | NI | NI | NI | NI |
| Small Bowel | 0.942 ± 0.040 | 0.892 ± 0.066 | NI | NI | 0.615 ± 0.18 | 0.467 ± 0.180 | NI | NI | NI | NI | NI | NI |
| Spleen | 0.975 ± 0.034 | 0.953 ± 0.056 | 0.979 ± 0.020 | 0.960 ± 0.037 | 0.941 ± 0.060 | 0.893 ± 0.088 | 0.959 ± 0.058 | 0.891 ± 0.193 | 0.963 ± 0.038 | 0.925 ± 0.092 | 0.956 ± 0.039 | 0.919 ± 0.061 |
| Stomach | 0.953 ± 0.049 | 0.914 ± 0.080 | 0.940 ± 0.087 | 0.896 ± 0.121 | 0.883 ± 0.084 | 0.799 ± 0.118 | 0.895 ± 0.112 | 0.751 ± 0.263 | 0.903 ± 0.105 | 0.780 ± 0.232 | 0.882 ± 0.134 | 0.809 ± 0.174 |
| UB | 0.957 ± 0.023 | 0.919 ± 0.040 | 0.92 ± 0.06 | 0.858 ± 0.093 | 0.855 ± 0.098 | 0.759 ± 0.137 | NI | NI | NI | NI | NI | NI |
| Vertebrae | 0.990 ± 0.006 | 0.980 ± 0.011 | 0.974 ± 0.008 | 0.949 ± 0.016 | 0.852 ± 0.052 | 0.746 ± 0.071 | NI | NI | NI | NI | NI | NI |

dataset that predominantly includes adults. However, after using the expanded dataset, strategy#2 models showed improved segmentation results, as detailed in Fig. 4. These cases were derived from the study by Jordan *et al.* [42] where manual segmentation was limited to a few numbers of organs. In other words, the noisy segmentations produced by strategy#1 models served to improve the performance of strategy#2 models in pediatric cases. Unfortunately, there were no manual segmentations available for these images to report the DSCs. The pediatric dataset [42] lacked manual segmentations of the aorta, brain, clavicles, eyeballs, hips, sacrum, ribs, and vertebrae.

## 4. Discussion

Automated multi-organ segmentation is a critical step in a wide range of clinical applications, including personalized radiation dosimetry, computational modeling, image quantification and radiation treatment planning. The availability of a fast and reliable organ segmentation tool can facilitate the automation of these procedures and their adoption/deployment in clinical setting. In this work, we developed DL-based models to segment multiple organs from total-body CT images using state-of-the-art nnU-Net [45] pipeline and compared the performance of our models with previous algorithms reported in the literature. Our model was trained on images presenting high variability using large datasets, including adults, pediatrics, and patients presenting with a wide range of pathologies and anatomic variations. The proposed models showed acceptable performance robust in both pediatric and adult images as shown in supplementary Table 1. The public databases we used are from multiple repositories, each intended for different purposes, such as RT treatment planning and image quantification. Each database provided a limited number of delineated organs. To complete the missing organ masks, we used our models trained in strategy#1 to infer and generate segmentations for the entire 4017 dataset, resulting in a large, uniform dataset containing multiple organs. Trained models in strategy#2 showed better performance in pediatrics, especially for those organs where the training data were available only for adults. These results show that strategy#2 models successfully learned the image's important patterns and features to segment those organs, even if inaccurate reference segmentations were generated by strategy#1 models. We will make the entire dataset publicly available as a useful resource for scientists active in the field. Strategy#2 models can generate a unified segmentation mask with much less inference time compared to strategy#1 models because it involves only a single inference. However, it should be noted that there is a performance drop for a few specific organs.

To evaluate the performance of our proposed model on real-world, unseen external datasets, we tested it on the test dataset provided by

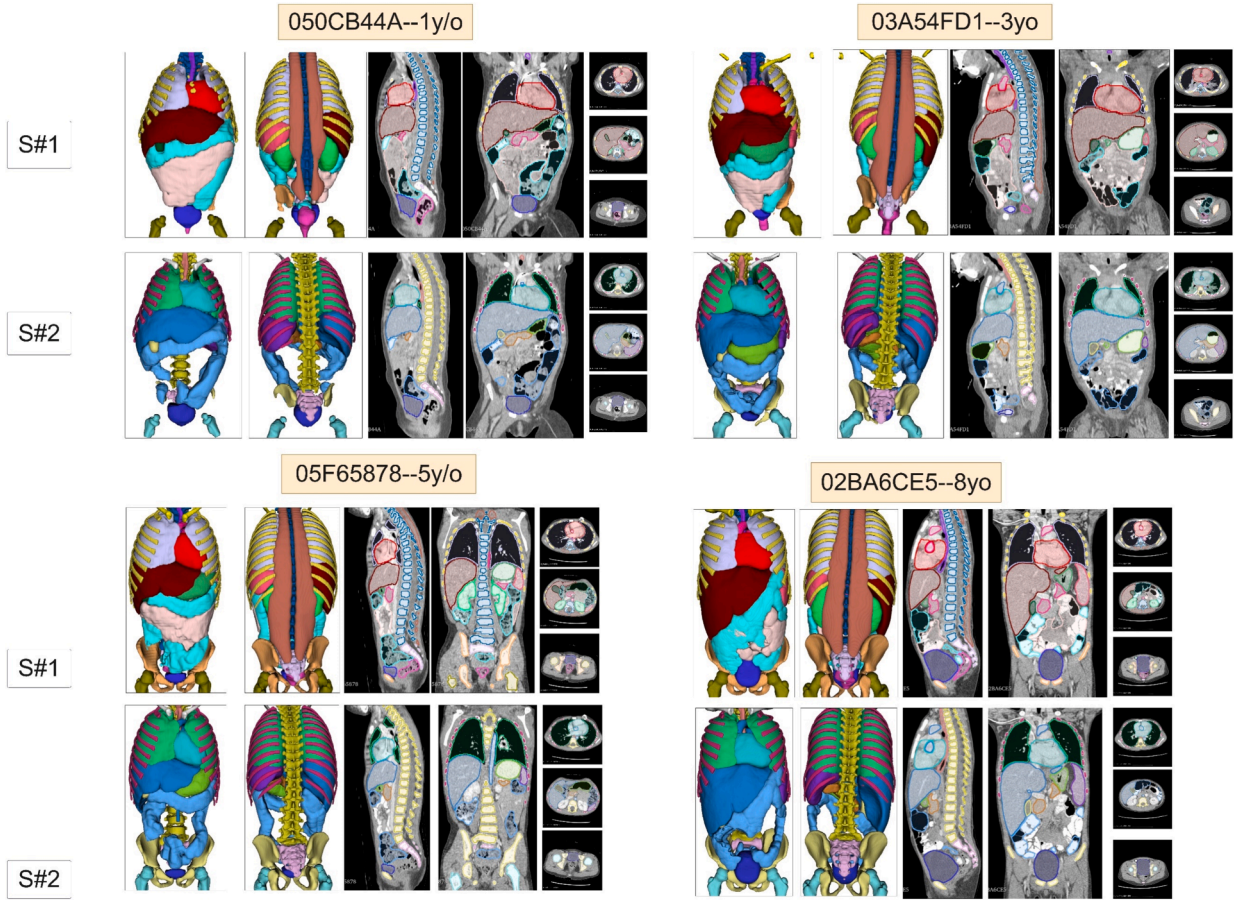


**Fig. 5.** Pediatric cases from subset#1 segmented using strategy#1 and strategy#2 models. The yellow box at the top of each image indicates the original image ID provided by Jordan et al. (43). Note the differences in the sacrum, hips, and aorta. The distinctions between strategy#1 and strategy#2 models are less noticeable in older pediatric patients, as seen in the bottom row for 5- and 8-years old patients. A high-resolution version of these images is available in supplementary material. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Wasserthal *et al.* [43], referred to as subset #2. Additionally, we tested their well-known and widely used models on our test set, referred to as subset #1. The number of valid cases in the testing group was limited, which could amplify the impact of statistical differences. In contrast, our test dataset was larger. Additionally, we tested Wasserthal *et al.* [43] as well as our strategy#1 and strategy#2 models on a selected sample from Qu *et al.* [47] study, designated as subset #3, to provide a fair comparison. We also visually checked subset #3 images and found that some provided reference segmentations were not accurate and excluded them from the evaluations. Both strategy#1 and strategy#2 models outperformed Wasserthal *et al.* [43] models on subset #3. Our models tested on subset #2 achieved superior DSC compared to their models tested on subset #1. This highlights the importance of training nnU-Net with larger and more diverse training data. The entire 4000 CT images were used to train strategy #2 models and the trained models in strategy #2 performance was evaluated on two external sets consisting of subsets # 2 and #3 where manual segmentations were available. As shown in Fig. 5, strategy #2 models which generate a group of organ segmentations using a single model could generate better segmentations in paediatric cases. This improvement may be due to two reasons. The first reason could the multi-label prediction and training where the network can be optimised according to the localization and the information in the location of the different organs (labels in the segmentation mask). The second reason which is probably more important is the larger number of training images in strategy #2 compared to strategy #1. The larger dataset even with minimal error in the training segmentation masks generated by models in strategy #1 can result in a more robust and generalizable model, especially on the external unseen dataset, such as subsets #2 and #3. For example, the sacrum and hips segmentations were not available for the training images in our 300 paediatric CT images for models in strategy #1, but the CT and imperfect segmentations were available during training in strategy #2 models.

In summary, we developed organ segmentation models in two different strategies: one per organ and another for organ group generation models trained and tested on a big cohort containing adult and paediatric CT images. Our models showed superior performance in a fair comparison with widely used models developed by Wasserthal et al. [43]. In addition, we made our models and the inference instruction as well as 4000 CT/ segment datasets publicly available.

One limitation of using nnU-Net 3Dfullres configuration on large images is RAM occupation which limits the inference speed and the user experience to use standard PCs for inference. We proposed two approaches to tackle this issue: (i) cropping the patient's body contour and (ii) breaking down images into parts along the cranio-caudal axis. It should be noted that our body contouring method differs from the crop-to-foreground algorithms provided by image processing libraries. Our proposed body contouring algorithm effectively removes any additional object, such as blankets or the CT table using a combination of 2D and 3D image processing. Besides, we considered an overlap between the image parts to ensure that the results are consistent with those obtained using the original image. This overlap provides the neighborhood information necessary for the model during sliding window inference, resulting in nearly identical outcomes. Although the overall inference time on a PC with sufficient RAM remains the same, the reduced RAM usage on PCs with limited RAM allows for a higher number of multi-processing workers, thereby speeding up the inference process.

We trained different DL models to segment 26 organs from total-body CT images which can be beneficial in various clinical tasks. We evaluated our models on an external dataset. The number of cases was limited to a few organs. The segmentation criteria varied across the manual segmentations available from the online databases, inherently causing inter-observer variability. For instance, some databases provided fully segmented kidneys, while others excluded pelvicalyceal systems. These differences could have misled our models and affected their performance.

## 5. Conclusion

We have developed a fully automated deep learning-based algorithm capable of generating accurate masks for multiple organs from CT images in an affordable computing time. We provided two different approaches and tested our model on an external dataset with excellent results. This tool should enables the implementation of many applications in clinical and research setting. Trained models for both strategies as well as inference instruction are avaiable on our GitHub page at: https://github.com/YazdanSalimi/Organ-Segmentation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Euratom research and training programme 2019-2020 Sinfonia project under grant agreement No 945196 and the Swiss National Science Foundation under grant SNSF 320030_231742.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2025.104911.

## References

[1] Shi F, Hu W, Wu J, Han M, Wang J, Zhang W, et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. Nat Commun 2022;13(1):6566.
[2] Mohammadi R, Shokatian I, Salehi M, Arabi H, Shiri I, Zaidi H. Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. Radiother Oncol 2021;159:231–40.
[3] Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 2017;44(2):547–57.
[4] Akhavanallaf A, Fayad H, Salimi Y, Aly A, Kharita H, Al Naemi H, et al. An update on computational anthropomorphic anatomical models. Digit Health 2022;8:20552076221111941.
[5] Tang Y, Huo Y, Xiong Y, Moon H, Assad A, Moyo T, et al. Improving splenomegaly segmentation by learning from heterogeneous multi-source labels. SPIE Medical Imaging: SPIE; 2019.
[6] Yang Y, Tang Y, Gao R, Bao S, Huo Y, McKenna MT, et al. Validation and estimation of spleen volume via computer-assisted segmentation on clinically acquired CT scans. J Med Imaging (Bellingham) 2021;8(1):014004.
[7] Hernandez-Boussard T, Macklin P, Greenspan EJ, Gryshuk AL, Stahlberg E, Syeda-Mahmood T, et al. Digital twins for predictive oncology will be a paradigm shift for precision cancer care. Nat Med 2021;27(12):2065–6.
[8] Salimi Y, Hajianfar G, Mansouri Z, Sanaat A, Amini M, Shiri I, et al. Organomics: a concept reflecting the importance of PET/CT healthy organ radiomics in non-small cell lung cancer prognosis prediction using machine learning. medRxiv. 2024: 2024.05.15.24307393.
[9] Mansouri Z, Salimi Y, Amini M, Hajianfar G, Oveisi M, Shiri I, et al. Development and validation of survival prognostic models for head and neck cancer patients using machine learning and dosiomics and CT radiomics features: a multicentric study. Radiat Oncol 2024;19(1):12.
[10] Mansouri Z, Salimi Y, Hajianfar G, Wolf NB, Knappe L, Xhepa G, et al. The role of biomarkers and dosimetry parameters in overall and progression free survival prediction for patients treated with personalized 90Y glass microspheres SIRT: a preliminary machine learning study. Eur J Nucl Med Mol Imaging 2024.
[11] Lindgren Belal S, Sadik M, Kaboteh R, Enqvist O, Ulen J, Poulsen MH, et al. Deep learning for segmentation of 49 selected bones in CT scans: first step in automated PET/CT-based 3D quantification of skeletal metastases. Eur J Radiol 2019;113:89–95.
[12] van Sluis J, Noordzij W, de Vries EGE, Kok IC, de Groot DJA, Jalving M, et al. Manual versus artificial intelligence-based segmentations as a pre-processing step in whole-body PET dosimetry calculations. Mol Imaging Biol 2023;25(2):435–41.
[13] Salimi Y, Mansouri Z, Hajianfar G, Sanaat A, Shiri I, Zaidi H. Fully automated explainable abdominal CT contrast media phase classification using organ segmentation and machine learning. Med Phys 2024;51(6):4095–104.
[14] Xie T, Zaidi H. Estimation of the radiation dose in pregnancy: an automated patient-specific model using convolutional neural networks. Eur Radiol 2019;29(12):6805–15.
[15] Fu W, Sharma S, Abadi E, Iliopoulos AS, Wang Q, Lo JY, et al. iPhantom: a framework for automated creation of individualized computational phantoms and its application to CT organ dosimetry. IEEE J Biomed Health Inform 2021;25(8):3061–72.
[16] Salimi Y, Akhavanallaf A, Mansouri Z, Shiri I, Zaidi H. Real-time, acquisition parameter-free voxel-wise patient-specific Monte Carlo dose reconstruction in whole-body CT scanning using deep neural networks. Eur Radiol 2023;33(12):9411–24.
[17] Mansouri Z, Salimi Y, Akhavanallaf A, Shiri I, Teixeira EPA, Hou X, et al. Deep transformer-based personalized dosimetry from SPECT/CT images: a hybrid approach for [(177)Lu]Lu-DOTATATE radiopharmaceutical therapy. Eur J Nucl Med Mol Imaging 2024;51(6):1516–29.
[18] Hobbis D, Yu NY, Mund KW, Duan J, Rwigema JM, Wong WW, et al. First report on physician assessment and clinical acceptability of custom-retrained artificial intelligence models for clinical target volume and organs-at-risk auto-delineation for postprostatectomy patients. Pract Radiat Oncol 2023;13(4):351–62.
[19] Liao W, Luo X, He Y, Dong Y, Li C, Li K, et al. Comprehensive evaluation of a deep learning model for automatic organs at risk segmentation on heterogeneous computed tomography images for abdominal radiation therapy. Int J Radiat Oncol Biol Phys 2023.
[20] Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. Phys Med 2021;85:107–22.
[21] Vrtovec T, Mocnik D, Strojan P, Pernus F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. Med Phys 2020;47(9):e929–50.
[22] Xiaoyu Liu LQ, Ziyue X, Jiayue Z, Yonghong S, Zhijian S. Towards More Precise Automatic Analysis: A Comprehensive Survey of Deep Learning-based Multi-organ Segmentation. Arxive. 2023.
[23] Liu X, Qu L, Xie Z, Zhao J, Shi Y, Song Z. Towards more precise automatic analysis: a systematic review of deep learning-based multi-organ segmentation. Biomed Eng Online 2024;23(1):52.
[24] Ma J, Zhang Y, Gu S, Zhu C, Ge C, Zhang Y, et al. AbdomenCT-1K: is abdominal organ segmentation a solved problem? IEEE Trans Pattern Anal Mach Intell 2022;44(10):6695–714.
[25] Huang Z, Wang H, Deng Z, Ye J, Su Y, Sun H, et al. STU-Net: scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv pre-print server. 2023.
[26] Zhao Q, Zhong L, Xiao J, Zhang J, Chen Y, Liao W, et al. Efficient multi-organ segmentation from 3D abdominal CT images with lightweight network and knowledge distillation. IEEE Trans Med Imaging 2023;42(9):2513–23.
[27] Hadjiiski L, Cha K, Chan HP, Drukker K, Morra L, Nappi JJ, et al. AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med Phys 2023;50(2):e1–24.
[28] Ji Y, Bai H, Ge C, Yang J, Zhu Y, Zhang R, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Adv Neural Inf Proces Syst 2022;35:36722–32.
[29] Xu Y, Tang O, Tang Y, Lee HH, Chen Y, Gao D, et al. Outlier guided optimization of abdominal segmentation. Proc SPIE Int Soc Opt Eng 2020;11313:1131336.
[30] Fu H, Zhang J, Li B, Chen L, Zou J, Zhang Z, et al. Abdominal multi-organ segmentation in Multi-sequence MRIs based on visual attention guided network and knowledge distillation. Physica Medica: Eur J Med Phys 2024;122.
[31] Bordigoni B, Trivellato S, Pellegrini R, Meregalli S, Bonetto E, Belmonte M, et al. Automated segmentation in pelvic radiotherapy: a comprehensive evaluation of ATLAS-, machine learning-, and deep learning-based models. Physica Medica: Eur J Med Phys 2024;125.
[32] Tong N, Xu Y, Zhang J, Gou S, Li M. Robust and efficient abdominal CT segmentation using shape constrained multi-scale attention network. Physica Medica: Eur J Med Phys 2023;110.
[33] Lucido JJ, DeWees TA, Leavitt TR, Anand A, Beltran CJ, Brooke MD, et al. Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning. Front Oncol 2023;13:1137803.
[34] Julie A, Shiyam Sundar LK, Seban R-D, Luporsi M, Nioche C, Beyer T, et al. <strong>MOOSE vs TotalSegmentator: comparison of feature values of segmented anatomical regions in [18F]FDG PET/CT images</strong&gt. Journal of Nuclear Medicine. 2024;65(supplement 2):241948.
[35] Salimi Y, Shiri I, Akhavanallaf A, Mansouri Z, Saberi Manesh A, Sanaat A, et al. Deep learning-based fully automated Z-axis coverage range definition from scout scans to eliminate overscanning in chest CT imaging. Insights Imaging 2021;12(1):162.
[36] Salimi Y, Shiri I, Akhavanallaf A, Mansouri Z, Sanaat A, Pakbin M, et al. Deep Learning-based calculation of patient size and attenuation surrogates from localizer

Image: toward personalized chest CT protocol optimization. Eur J Radiol 2022; 157:110602.
[37] Patrick B, Eugene V, Grzegorz C, Hao C, Qi D, Chi-Wing F, et al. The Liver Tumor Segmentation Benchmark (LiTS). 2019.
[38] Heller N, Sathianathen N, Kalapara A, Walczak E, Moore K, Kaluzniak H, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:190400445. 2019.
[39] Rister B, Yi D, Shivakumar K, Nobashi T, Rubin DL. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. Sci Data 2020;7(1):381.
[40] Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The Medical Segmentation Decathlon Nature communications 2022;13(1):4128.
[41] Sekuboyina A, Husseini ME, Bayat A, Loffler M, Liebl H, Li H, et al. VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. Med Image Anal 2021;73:102166.
[42] Jordan P, Adamson PM, Bhattbhatt V, Beriwal S, Shen S, Radermecker O, et al. Pediatric chest-abdomen-pelvis and abdomen-pelvis CT images with expert organ contours. Med Phys 2022;49(5):3523–8.
[43] Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiology Artificial intelligence 2023;5(5):e230024.
[44] AAPM. Use of Water Equivalent Diameter for Calculating Patient Size and Size-Specific Dose Estimates (SSDE) in CT. AAPM; 2014. Report No.: The Report of AAPM Task Group 220.
[45] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–11.
[46] Salimi Y, Shiri I, Akavanallaf A, Mansouri Z, Arabi H, Zaidi H. Fully automated accurate patient positioning in computed tomography using anterior-posterior localizer images and a deep neural network: a dual-center study. Eur Radiol 2023; 33(5):3243–52.
[47] Qu C, Zhang T, Qiao H, Tang Y, Yuille AL, Zhou Z. Abdomenatlas-8k: annotating 8,000 CT volumes for multi-organ segmentation in three weeks. Adv Neural Inf Proces Syst 2024;36.