Biomedical Physics & Engineering Express

Endorsed by IPEM Institute of Physics and Engineering in Medicine

ACCEPTED MANUSCRIPT

Multi-organ segmentation of CT via convolutional neural network: Impact of training setting and scanner manufacturer

To cite this article before publication: Amy J Weisman *et al* 2023 *Biomed. Phys. Eng. Express* in press https://doi.org/10.1088/2057-1976/acfb06

Multi-organ segmentation of CT via convolutional neural network: Impact of training setting and scanner manufacturer

Amy J Weisman[1], Daniel T Huff[1], Rajkumar Munian Govindan[1], Song Chen[2], Timothy G Perk[1]

[1] AIQ Solutions, Madison, WI, USA
[2] Department of Nuclear Medicine, The First Hospital of China Medical University, Shenyang, Liaoning, China
E-mail: amy.weisman@aiq-solutions.com

Abstract (300/300 words)

Objective: Automated organ segmentation on CT images can enable the clinical use of advanced quantitative software devices, but model performance sensitivities must be understood before widespread adoption can occur. The goal of this study was to investigate performance differences between Convolutional Neural Networks (CNNs) trained to segment one (single-class) versus multiple (multi-class) organs, and between CNNs trained on scans from a single manufacturer versus multiple manufacturers.

Methods: The multi-class CNN was trained on CT images obtained from 455 whole-body PET/CT scans (413 for training, 42 for testing) taken with Siemens, GE, and Phillips PET/CT scanners where 16 organs were segmented. The multi-class CNN was compared to 16 smaller single-class CNNs trained using the same data, but with segmentations of only one organ per model. In addition, CNNs trained on Siemens-only (N=186) and GE-only (N=219) scans (manufacturer-specific) were compared with CNNs trained on data from both Siemens and GE scanners (manufacturer-mixed). Segmentation performance was quantified using five performance metrics, including the Dice Similarity Coefficient (DSC).

Results: The multi-class CNN performed well compared to previous studies, even in organs usually considered difficult auto-segmentation targets (*e.g.*, pancreas, bowel). Segmentations from the multi-class CNN were significantly superior to those from smaller single-class CNNs in most organs, and the 16 single-class models took, on average, six times longer to segment all 16 organs compared to the single multi-class model. The manufacturer-mixed approach achieved minimally higher performance over the manufacturer-specific approach.

Significance: A CNN trained on contours of multiple organs and CT data from multiple manufacturers yielded high-quality segmentations. Such a model is an essential enabler of image processing in a software device that quantifies and analyzes such data to determine a patient's treatment response. To date, this activity of whole organ segmentation has not been adopted due to the intense manual workload and time required.

Keywords: Multi-site, organ segmentation, whole-body, multi-scanner, CNN

1. Introduction

Segmentation of organs on Computed Tomography (CT), Positron Emission Tomography (PET), or Magnetic Resonance Imaging (MRI) scans has proven useful for

many medical image analysis tasks. Examples include diagnosis *(Diaconis and Rao 1980, Mahmood* et al *2019)*, treatment response monitoring *(Padhani and Koh 2011)*, radiotherapy treatment planning *(Thorwarth 2015, Stieb* et al *2019)*, and treatment-related toxicity detection *(Frelau* et al *2021, Hribernik* et al *2022)*. However, when performed manually, organ segmentation is so time-consuming that it is not feasible to integrate into a clinical setting *(Vaassen* et al *2020, van der Veen* et al *2020)* and subject to substantial inter-observer variations *(Hansen* et al *2018, Lorenzen* et al *2021)*. Convolutional Neural Networks (CNNs) have demonstrated the ability to perform automated segmentation of multiple anatomical sites *(Kavur* et al *2021, Liu* et al *2018, 2020)*, and have been shown to reduce both time spent on segmentation and inter-observer variability *(Vaassen* et al *2020, van der Veen* et al *2020, Gooding* et al *2018, Trimpl* et al *2022).*

Implementing CNN-based segmentation in a clinical setting requires careful consideration of several practical issues. Commercially available scanners vary both in terms of imaging hardware and reconstruction software. Options for segmentation algorithm architecture also differ widely, each having their own specific advantages and drawbacks *(Moeskops* et al *2016, Minaee* et al *2022)*. Finally, the variety and complexity of the organs to be segmented vary depending on the clinical use case. These factors may hamper the implementation of CNN-based segmentation workflows. For clinical use, it is crucial that automated medical image processing steps, such as segmentation by CNN, be trained and evaluated in large, heterogeneous datasets representing a realistic variety of imaging hardware they are likely to encounter in the clinic.

Although CNN-based segmentation is becoming more widely adopted in some radiotherapy workflows *(Cha* et al *2021, Chang* et al *2021, Schreier* et al *2020)*, it is not widely used in settings aiming to assess response to therapy in patients with advanced cancers. In addition, a number of uncertainties and unknowns remain to be sufficiently addressed. Among these are possible scanner- or manufacturer-specific effects (e.g., noise patterns) as well as the question of model performance when data from multiple institutions are used *(Ng* et al *2021, Roth* et al *2020)*. Additionally, for the segmentation of multiple organs, it is unknown whether features learned for the segmentation of one organ are optimal for the segmentation of others, and thus whether optimal performance is achieved by training one multi-class model, or by training multiple single-class models *(Amjad* et al *2022)*.

Our aims in this study were: 1) to assess the performance of CNN-based multi-class segmentation model in a large, diverse data set containing CT scans from multiple PET/CT scanner manufacturers, and 2) to evaluate the sensitivity of segmentation performance to the training setting (multi- vs. single-class training) and to the scanner manufacturer used for training and testing.

2. Methods

2.1 Data Set

The imaging data used in this study consisted of 455 retrospectively collected whole-body CT scans either from public sources or obtained by AIQ Solutions as part of research collaborations with academic medical centers. These cohorts were selected for their range of patient sex and disease burden, which can impact the presentation of certain organs.

Scans were acquired on either Siemens Healthineers (186 scans, 11 scanner models), GE Medical Systems scanners (219 scans, 8 scanner models), or Phillips Medical Systems machines (26 scans, 5 scanner models) between 2005 and 2021. For 24 scans, scanner information was unavailable. As scans were acquired retrospectively, scans were reconstructed according to each sites clinical workflow and thus encompassed a variety of reconstruction settings. Details of the patient demographics and scanner information is outlined in *Table 1*.

In each scan, sixteen structures were manually contoured by either an experienced nuclear medicine physician with 15 years' experience (author SC) or a radiographer with over 10 years' experience: liver, spleen, lungs, thyroid, kidneys, pancreas, bladder,

aorta, adrenal glands, bowel, stomach, heart, eyes, salivary glands (consisting of parotid and submandibular glands), pituitary gland, and choroid plexus. Note that while not all structures are organs, the term "organ" is used in this work for brevity. Image contouring and review was completed using 3D Slicer *(Kikinis* et al *2014)*.

2.2 CNN model architecture and training

For the multi-class segmentation model, 393 CT scans were used for training (86%), 20 scans were used for monitoring training progress (4%), and 42 scans were held out as an external test set (10%).

A deep learning model with a fully 3D U-net architecture was trained for organ segmentation, outlined as follows. As in *(Çiçek* et al *2016)*, the U-net architecture involves an "analysis" path and a "synthesis" path, with skip connections that allows the network to learn features at multiple resolutions. A single input channel was used for the CT image, which was resampled to a grid size of 2.0×2.0×2.0 mm and normalized such that CT values within the patient had a mean of 0 and standard deviation of 1.

Patches of size 128×128×128 voxels were extracted from the training images using class balancing to ensure an equal number of patches were sampled from each target organ. In total, 34 patches per patient (2 per class, including background) were extracted before training, resulting in 15,470 total patches. Data augmentation including random Gaussian noise, random rotations, random flips, and random elastic transformations, were randomly chosen and applied to 70% of the training patches on the fly. The loss was the average of the cross entropy and dice similarity coefficient (DSC), which was optimized using stochastic gradient descent with a learning rate of 0.01 decreased by a factor of 2 every 20 epochs for a total of 150 epochs. One epoch was defined as the process of all 15,470 patches undergoing one forward pass through the model exactly once.

For the testing dataset, inference was performed on processed images (cubic voxel size and normalized CT) using overlapping patches with a step size of 64 voxels (half of the patch size). Voxels at the center of each patch were weighted with higher confidence using a 3D Gaussian function. After patch inference, the U-net probability maps were resampled back to the natural CT resolution using linear interpolation. Final segmentation maps were then generated by taking the maximum probability in each segmentation class.

A step involving the largest connected component analysis was taken to remove extraneous segmentations. For the liver, spleen, pancreas, aorta, bladder, bowel, stomach, and heart, the single largest connected component was taken. For the lungs, thyroid, and kidneys, the largest two connected components were taken. This step was applied to the outputs of all trained models.

Models were trained using an NVIDIA 3090 RTX GPU with 24 GB of VRAM.

2.2 Sub-study I: Single- vs multi-class model

To investigate the impact of single- versus multi-class training on segmentation performance, sixteen organ-specific models were trained for each organ using the same training data set as the mutli-class model, but with contours of only one organ as targets. All training parameters (train/validation/test split, optimizer, loss, learning rate) were kept identical to the multi-class model. However, single class models were made smaller by reducing the number of feature maps by a factor of 4. This was done to reduce the inference time of the single-class models, as maintaining the same large model size for 16 individual single-class models would result in an inference time rougly 16 times that of the multi-class model. The total inference time for each patient was extracted and compared between the multi-class and single-class models.

2.3 Sub-study II: Manufacturer-specific vs. manufacturer-mixed models

To determine the effect of scanner manufacturer on CNN segmentation performance, four models were trained. Two manufacturer-specific models were trained: a GE-only model trained on 186 CT scans from GE scanners, and a Siemens-only model trained on 186

CT scans from Siemens scanners. Each model was tested on all scans available from the other manufacturer. Two manufacturer-mixed models were trained with the same data as used for the manufacturer-specific models using an approach similar to 2-fold cross validation: each model was trained with 186 CT scans split evenly between GE and Siemens scanners, and evaluated on the remaining scans. Due to the small number of CT scans acquired on Phillips scanners, these data were excluded from this substudy. In addition, the 24 scans for which scanner information was not available were excluded from this substudy.

2.4 Metrics for segmentation evaluation

Model performance was quantified using a combination of overlap, surface distance, and voxel-wise metrics:

The Dice Similarity Coefficient (DSC)

$$DSC(A,B) = 2\frac{|A \cap B|}{|A| + |B|}$$

where A and B are the evaluated and reference segmentations, respectively.

The average symmetric surface distance (ASSD)

$$ASSD(A,B) = \frac{1}{|S_A + S_B|}\left(\sum_{s\in S_A} d(s,S_B) + \sum_{s\in S_B} d(s,S_A)\right)$$

where S_A and S_B are the surfaces of the evaluated and reference segmentations, respectively, and d is the minimum distance between a voxel s and a set of boundary voxels S':

$$d(s,S') = \min_{s'\in S'} \|s - s'\|_2 .$$

The 95% Hausdorff distance (HSD95), which is the 95th percentile P_{95} of the set of surface distances between the evaluated and reference segmentations:

$$HSD_{95}(A,B) = P_{95}(d_{s\in S_B}(s,S_A), d_{s\in S_A}(s,S_B)) .$$

The voxel-wise sensitivity:

$$Sensitivity = \frac{TP}{TP + FN} .$$

The voxel-wise positive predictive value (PPV):

$$PPV = \frac{TP}{TP + FP} .$$

2.5 Statistical analysis

The multi-class segmentation model was assessed for bias in the test dataset by calculating Spearman correlation of DSC with patient age and weight. A Wilcoxon rank sum test was used to assess differences between patient sex. Differences in segmentation performance by single-class versus multi-class training setting were assessed using paired Wilcoxon Signed-Rank tests. Differences in segmentation performance across the manufacturer-specific and manufacturer-mixed models were assessed with paired Wilcoxon rank sum test. P-values were corrected for multiple hypotheses using the Bonferroni method. Following correction, p-values below 0.05 were considered statistically significant.

3. Results

3.1 Multi-class segmentation

The model trained on multi-class segmentation data sets performed well in delineating all of the investigated organs (*Table 2*). Large, visceral organs (*e.g.*, liver, lungs) and small, well-defined structures (*e.g.*, aorta, kidneys) achieved excellent performance in the evaluated metrics. Acceptable performance was achieved in smaller organs which have traditionally proven difficult for auto-segmentation models, such as the pancreas (median DSC 0.788, median ASSD 1.9 mm) and thyroid gland (median DSC 0.748, median ASSD 1.3 mm). CNN segmentations produced by the multi-class model for an example patient in the external test set are shown in *Figure 1*.

Examples of cases with poor performance are shown in Figure 2. A large number of cases with poor performance can be attributed to CT image artifacts,

abnormal pathology, or imperfections in the manual ground truth contours.

In the test dataset, the median [range] of patient age and weight were 66 years [39, 83] and 77 kg [49, 121]. The patient sex distribution was 14 female, 25 male, and 3 unknown. No Bonferroni corrected p-values were statistically significant for the correlation between DSC and patient age, patient weight, or patient sex.

3.2 Sub-study I: Single- vs. multi-class model

The multi-class model outperformed the single-class models for DSC, ASSD, HSD95, PPV, and Sensitivity in 16/16, 14/16, 13/16, 12/16, and 4/16 target organs, respectively (Wilcoxon paired test $p<0.05$). A comparison of DSC and ASSD between the multi-class and single-class models is shown in Figure 3. All performance metrics are summarized in *Table 2*.

The multi-class model's superior performance was especially pronounced in the pancreas (median DSC 0.788 vs. 0.670, median ASSD 1.93 mm vs. 3.48 mm) and the stomach (median DSC 0.902 vs. 0.824, median ASSD 1.71 mm vs. 2.88 mm), as shown in Figure 3b. Visual analysis indicates that improved performance of the multi-class model is especially pronounced in areas where two organs are touching, such as the liver and stomach boundary. An example case comparing the multi-class and single-class models to the ground truth is shown in Figure 4.

Inference time for the single multi-class model was substantially faster than the total inference time for the 16 single-class models. Across the external test set, inference time for the multi-class model was 79 ± 29 seconds (mean ± sd). The total inference time for all 16 single-class models was 537 ± 224 seconds.

3.3 Sub-study II: Manufacturer-specific vs. manufacturer-mixed models

For the 219 GE images, the manufacturer-mixed approach had overall better performance (Figure 5a). The results from the manufacturer-mixed models had significantly higher DSC for 9 of the 16 organs compared to the model trained on only Siemens data. Median improvements in DSC ranged from +0.0007 to +0.05. In the remaining 7 organs, no significant differences in DSC were found. Similar results indicating superior performance of the manufacturer-mixed model were found for ASSD, HSD95, and PPV (Supplemental Material **Error! Reference source not found.**).

In the 186 Siemens images, results varied more widely by organ (Figure 5b). The manufacturer-mixed models achieved significantly higher DSC compared to the model trained with GE only images in the spleen, lung, bladder, and bowel, but achieved significantly lower DSC in the adrenals and pituitary gland. Median differences in DSC for the significantly different organs ranged from -0.04 to +0.01. Other organs showed non-significant differences in DSC performance. Similar mixed results was found for ASSD, HSD95, PPV, and Sensitivity (Supplemental Material **Error! Reference source not found.**).

No significant differences between patient age or patient weight was found across the GE and Siemens cohorts via Wilcoxon rank sum testing. The cohorts had a similar number of female patients (37% for GE data, 44% for Siemens data).

4. Discussion

In this study, we trained a CNN to segment sixteen organs in a large, diverse dataset of whole-body CT images. Our 3D U-Net model trained in a multi-class setting was capable of segmenting the target organs with excellent performance across a wide set of patient demographics, indicating the model should generalize well to other patient cohorts. We investigated the impact of multi-class versus single-class training and observed that the multi-class model outperformed smaller single-class models for a majority of organs and performance metrics. We also investigated manufacturer-specific versus manufacturer-mixed training and found segmentation quality to be largely independent of scanner manufacturer.

Automated methods for organ segmentation, especially through the use of a multi-class model, can

significantly decrease the time required for any clinical task requiring whole organ segmentation *(Men* et al *2017)*. However, this time reduction does not necessarily impact the overall time a clinican spends assessing patients. Instead, it would enable the use of whole-organ assessment in a clinical setting where it is currently not in use. This is because manual segmentation of large organs is so time consuming that it is not feasible to be added to a clinical workflow. With a processing time of fewer than two minutes per scan, organ contours can be generated before clinicans begin their workflow and can be used to quantify useful imaging features, such as organ size and shape metrics, or location-specific PET tracer uptake (*e.g.*, liver SUV_{mean}). These imaging features have been shown to be useful in many applications, with examples ranging from quantification of uptake or organ volume for assessment of systemic disease *(Martin* et al *2022)* or immune response *(Frelau* et al *2021, Hribernik* et al *2022)*, categorizing regions of interest based on their location to complement radiological reads, and segmenting organs at risk for radiation treatment planning *(Chung* et al *2021)*. Models trained on large, heterogeneous datasets which segment a large number of structures such as the one presented in this work would enable a single CNN to be used across many different tasks.

The performance of our large, multi-class model is similar to the results of several past studies (*Table* 3). Organs which are large and which demonstrate contrast between the neighbouring organs and the background tissue such as lung, liver, and spleen have traditionally shown the highest segmentation performance; our results demonstrate high performance in lung, liver, and spleen (*Table 2*). Similarly, organs that are small or have poor contrast with neighbouring organs such as pancreas, thyroid, and adrenal glands showed lower to intermediate performance in our study. Additionally, organs that have inherent shape and surface complexity or variable appearance on CT such as the bowel are also difficult to segment and have shown poor segmentation performance *(Men* et al *2017)*; however, our model achieved a median bowel DSC of 0.90, which is higher than past literature, and demonstrates the robustness of CNN-based bowel segmentation when trained with a large, heterogeneous dataset. A thorough review of CNN-based organ segmentation for radiotherapy treatment planning can be found in *(Samarasinghe* et al *2021)*.

The model trained on delineations of multiple organs outperformed single-class models for a majority of target organs and evaluated performance metrics. This may be due to the single-class model architectures having a reduced number of feature maps (by a factor of 4) compared to the multi-class models. Single-class models were made smaller to reduce the total time needed for inference: larger models are more computationally expensive thus taking longer to perform inference. Hence, the single-class models showed reduced performance and increased overall time to run. Despite the potential performance disadvantages, single-class models may offer more flexibility in training: curating datasets with all needed organs is difficult, while many public datasets are available with varying organs segmented and could be combined for single-class model training.

It is possible that single-class models would have superior performance to multi-class models if CNN architecture is kept sufficiently large, or if additional hyperparameter and architecture tuning were investigated such as through model self configuration *(Isensee* et al *2021)*. For example, single-class models allow for (and may achieve better performance with) softmax activation, which was not investigated in this work. However, improved performance of the multi-class CNN may also be achieved with additional architecture tuning, or through the use of more sophisticated CNN techniques such as transformer models *(Hatamizadeh* et al *2022)*. Architecture tuning or self configuration may also reveal that the optimal architecture for data from a single manufacturer or image quality may differ from that of another manufacturer. This assessment is outside the scope of the current paper, but is of interest for future work.

In our investigation of manufacturer-specific versus manufacturer-mixed segmentation models, our results indicated that the manufacturer-mixed approach achieved only minimally higher performance despite

being statistically significant. This indicates that CT-based segmentation models may achieve good performance on images acquired on scanners from manufacturers not included in the training dataset. The ability of manufacturer-specific models to generalize well to other scanner manufacturers in this study may be due to the wide range of scanner models present in the training set from each manufacturer, and due to reconstruction protocols varying by imaging site. Additionally, ll CT convolution kernels for the GE and Siemens scans were smooth kernels that have similar image quality *(Mackin* et al *2019)*, as is expected for CT scans acquired for PET attenuation correction. Thus, it is likely that the matching of image quality across the training and testing datasets allows for minimal differences in performance to be found across the manufacturers. Further research is needed to determine whether generalizability extends across CT image quality (e.g., CT dose, kernel sharpness, contrast agents). In those scenarios, image standardization algorithms may allow for improved generalization across CT image quality.

This study had several important limitations that should be discussed. Our training data set was restricted to CT scans of adults imaged at centers in the USA. We conducted the study using one CNN architecture, U-Net, which has demonstrated excellent performance in similar segmentation tasks. Future research regarding our findings should be to validate using additional CNN hyperparameters. Finally, our analysis of scanner manufacturer focused on GE and Siemens, currently two of the largest manufacturers of CT scanners. However, generalizability to CT images from other vendors should also be investigated.

5. Conclusion

A 3D U-Net model produced high-quality segmentations in a multi-class setting. A single multi-class model outperformed multiple smaller, single-class models, and performance was consistent across models trained using data from multiple vendors. The multi-class organ segmentation model is a component of the software device, *TRAQinform IQ*, a software-based medical device developed by AIQ Solutions for the analysis of PET and PET/CT data regarding identified regions of interest and the quantification of change during treatment. CNNs for segmenting organs trained on large imaging datasets with characteristics similar to real-world clinical data have potential for immediate clinical translation within these types of software devices.

Acknowledgements

We kindly thank Laura Patricia Kaplan for her help in drafting this manuscript.

Conflicts of interest

Authors AJW, DTH, RMG, and TGP are employees of AIQ Solutions, Madison, WI, USA, which funded the study. Author SC is a contractor of AIQ Solutions, Madison, WI, USA.

References

Amjad A, Xu J, Thill D, Lawton C, Hall W, Awan M J, Shukla M, Erickson B A and Li X A 2022 General and custom deep learning autosegmentation models for organs in head and neck, abdomen, and male pelvis *Med. Phys.* **49** 1686–700

Cha E, Elguindi S, Onochie I, Gorovets D, Deasy J O, Zelefsky M and Gillespie E F 2021 Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **159** 1–7

Chang Y, Wang Z, Peng Z, Zhou J, Pi Y, Xu X G and Pei X 2021 Clinical application and improvement of a CNN-based autosegmentation model for clinical target volumes in cervical cancer radiotherapy *J. Appl. Clin. Med. Phys.* **22** 115–25

Chung S Y, Chang J S, Choi M S, Chang Y, Choi B S, Chun J, Keum K C, Kim J S and Kim Y B 2021 Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery *Radiat. Oncol. Lond. Engl.* **16** 44

Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O 2016 3D U-net: Learning dense volumetric segmentation from sparse annotation *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol 9901 LNCS

Diaconis J N and Rao K C 1980 CT in head trauma: a review *J. Comput. Tomogr.* **4** 261–70

Frelau A, Palard-Novello X, Jali E, Boussemart L, Dupuy A, James P, Devillers A, Le Jeune F, Edeline J and Lesimple T 2021 Increased thyroid uptake on 18F-FDG PET/CT is associated with the development of permanent hypothyroidism in stage IV melanoma patients treated with anti-PD-1 antibodies *Cancer Immunol. Immunother.* **70** 679–87

Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, Davidson B, Pereira S P, Clarkson M J and Barratt D C 2018 Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks *IEEE Trans. Med. Imaging* **37** 1822–34

Gonzalez Y, Shen C, Jung H, Nguyen D, Jiang S B, Albuquerque K and Jia X 2021 Semi-automatic sigmoid colon segmentation in CT for radiation therapy treatment planning via an iterative 2.5-D deep learning approach *Med. Image Anal.* **68** 101896

Gooding M J, Smith A J, Tariq M, Aljabar P, Peressutti D, van der Stoep J, Reymen B, Emans D, Hattu D, van Loon J, de Rooy M, Wanders R, Peeters S, Lustberg T, van Soest J, Dekker A and van Elmpt W 2018 Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test *Med. Phys.* **45** 5105–15

Hänsch A, Schwier M, Gass T, Morgas T, Haas B, Dicken V, Meine H, Klein J and Hahn H K 2019 Evaluation of deep learning methods for parotid gland segmentation from CT images *J. Med. Imaging Bellingham Wash* **6** 011005

Hansen C R, Johansen J, Samsøe E, Andersen E, Petersen J B B, Jensen K, Andersen L J, Sand H M B, Bertelsen A S and Grau C 2018 Consequences of introducing geometric GTV to CTV margin expansion in DAHANCA contouring guidelines for head and neck radiotherapy *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **126** 43–7

Haq R, Hotca A, Apte A, Rimner A, Deasy J O and Thor M 2020 Cardio-pulmonary substructure segmentation of radiotherapy computed tomography images using convolutional neural networks for clinical outcomes analysis *Phys. Imaging Radiat. Oncol.* **14** 61–6

Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H R and Xu D 2022 Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* Lecture Notes in Computer Science vol 12962, ed A Crimi and S Bakas (Cham: Springer International Publishing) pp 272–84 Online: https://link.springer.com/10.1007/978-3-031-08999-2_22

Hribernik N, Huff D T, Studen A, Zevnik K, Klanečček Ž, Emamekhoo H, Škalic K, Jeraj R and Reberšek M 2022 Quantitative imaging biomarkers of immune-related adverse events in immune-checkpoint blockade-treated metastatic melanoma patients: a pilot study *Eur. J. Nucl. Med. Mol. Imaging* **49** 1857–69

Isensee F, Jaeger P F, Kohl S A A, Petersen J and Maier-Hein K H 2021 nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation *Nat. Methods* **18** 203–11

Jackson P, Hardcastle N, Dawe N, Kron T, Hofman M S and Hicks R J 2018 Deep Learning Renal Segmentation for Fully Automated Radiation Dose Estimation in Unsealed Source Therapy *Front. Oncol.* **8** 215

Kavur A E, Gezer N S, Barış M, Aslan S, Conze P-H, Groza V, Pham D D, Chatterjee S, Ernst P and Özkan S 2021 CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation *Med. Image Anal.* **69** 101950

Kikinis R, Pieper S D and Vosburgh K G 2014 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support *Intraoperative Imaging and Image-Guided Therapy* ed F A Jolesz (New York, NY: Springer New York) pp 277–89 Online: http://link.springer.com/10.1007/978-1-4614-7657-3_19

Lamba N, Wan H, Kruzer A, Platt E and Nelson A 2019 Clinical utility of a 3D convolutional neural network kidney segmentation method for radionuclide dosimetry *J. Nucl. Med.* **60** 267–267

Liu X, Guo S, Yang B, Ma S, Zhang H, Li J, Sun C, Jin L, Li X, Yang Q and Fu Y 2018 Automatic Organ Segmentation for CT Scans Based on Super-Pixel and Convolutional Neural Networks *J. Digit. Imaging* **31** 748–60

Liu Y, Lei Y, Fu Y, Wang T, Tang X, Jiang X, Curran W J, Liu T, Patel P and Yang X 2020 CT-based multi-organ segmentation using a 3D self-attention U-net network for pancreatic radiotherapy *Med. Phys.* **47** 4316–24

Lorenzen E L, Kallehauge J F, Byskov C S, Dahlrot R H, Haslund C A, Guldberg T L, Lassen-Ramshad Y,

Lukacova S, Muhic A, Witt Nyström P, Haldbo-Classen L, Bahij I, Larsen L, Weber B and Hansen C R 2021 A national study on the inter-observer variability in the delineation of organs at risk in the brain *Acta Oncol. Stockh. Swed.* **60** 1548–54

Mackin D, Ger R, Gay S, Dodge C, Zhang L, Yang J, Jones A K and Court L 2019 Matching and Homogenizing Convolution Kernels for Quantitative Studies in Computed Tomography: *Invest. Radiol.* **54** 288–95

Mahmood M, Kendi A T, Ajmal S, Farid S, O'Horo J C, Chareonthaitawee P, Baddour L M and Sohail M R 2019 Meta-analysis of 18F-FDG PET/CT in the diagnosis of infective endocarditis *J. Nucl. Cardiol. Off. Publ. Am. Soc. Nucl. Cardiol.* **26** 922–35

Martin E, Stuckey A, Heidel R E, Kennel S, Weisman A and Wall J 2022 Repeat PET/CT Imaging of a Patient with Systemic Amyloidosis Using iodine (124I) evuzamitide (124I-p5+ 14) Identifies Organ-Specific Amyloid Regression

Men K, Dai J and Li Y 2017 Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks *Med. Phys.* **44** 6377–89

Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N and Terzopoulos D 2022 Image Segmentation Using Deep Learning: A Survey *IEEE Trans. Pattern Anal. Mach. Intell.* **44** 3523–42

Mirando D, Saiprasad M, Pirozzi S, Kruzer A and Nelson A 2018 Evaluation of an automated lung segmentation method using an iterative thresholding and processing technique *J. Nucl. Med.* **59** 1756–1756

Moeskops P, Wolterink J M, van der Velden B H M, Gilhuijs K G A, Leiner T, Viergever M A and Išgum I 2016 Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* Lecture Notes in Computer Science ed S Ourselin, L Joskowicz, M R Sabuncu, G Unal and W Wells (Cham: Springer International Publishing) pp 478–86

Ng D, Lan X, Yao M M-S, Chan W P and Feng M 2021 Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets *Quant. Imaging Med. Surg.* **11** 852–7

Padhani A R and Koh D-M 2011 Diffusion MR Imaging for Monitoring of Treatment Response *Magn. Reson. Imaging Clin.* **19** 181–209

Park J, Lee J S, Oh D, Ryoo H G, Han J H and Lee W W 2021 Quantitative salivary gland SPECT/CT using deep convolutional neural networks *Sci. Rep.* **11** 7842

Rister B, Yi D, Shivakumar K, Nobashi T and Rubin D L 2020 CT-ORG, a new dataset for multiple organ segmentation in computed tomography *Sci. Data* **7** 381

Robinson-Weiss C, Patel J, Bizzo B C, Glazer D I, Bridge C P, Andriole K P, Dabiri B, Chin J K, Dreyer K, Kalpathy-Cramer J and Mayo-Smith W W 2023 Machine Learning for Adrenal Gland Segmentation and Classification of Normal and Adrenal Masses at CT *Radiology* **306** e220101

Roth H R, Chang K, Singh P, Neumark N, Li W, Gupta V, Gupta S, Qu L, Ihsani A, Bizzo B C, Wen Y, Buch V, Shah M, Kitamura F, Mendonça M, Lavor V, Harouni A, Compas C, Tetreault J, Dogra P, Cheng Y, Erdal S, White R, Hashemian B, Schultz T, Zhang M, McCarthy A, Yun B M, Sharaf E, Hoebel K V, Patel J B, Chen B, Ko S, Leibovitz E, Pisano E D, Coombs L, Xu D, Dreyer K J, Dayan I, Naidu R C, Flores M, Rubin D and Kalpathy-Cramer J 2020 Federated Learning for Breast Density Classification: A Real-World Implementation *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* Lecture Notes in Computer Science ed S Albarqouni, S Bakas, K Kamnitsas, M J Cardoso, B Landman, W Li, F Milletari, N Rieke, H Roth, D Xu and Z Xu (Cham: Springer International Publishing) pp 181–91

Samarasinghe G, Jameson M, Vinod S, Field M, Dowling J, Sowmya A and Holloway L 2021 Deep learning for segmentation in radiation therapy planning: a review *J. Med. Imaging Radiat. Oncol.* **65** 578–95

Schreier J, Genghi A, Laaksonen H, Morgas T and Haas B 2020 Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **145** 1–6

Stieb S, McDonald B, Gronberg M, Engeseth G M, He R and Fuller C D 2019 Imaging for Target Delineation and Treatment Planning in Radiation Oncology: Current and Emerging Techniques *Hematol. Oncol. Clin. North Am.* **33** 963–75

Sundar L K S, Yu J, Muzik O, Kulterer O C, Fueger B, Kifjak D, Nakuz T, Shin H M, Sima A K, Kitzmantl D, Badawi R D, Nardo L, Cherry S R, Spencer B A, Hacker M and Beyer T 2022 Fully Automated, Semantic Segmentation of Whole-Body 18F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence *J. Nucl. Med. Off. Publ. Soc. Nucl. Med.* **63** 1941–8

Thorwarth D 2015 Functional imaging for radiotherapy treatment planning: current status and future directions-a review *Br. J. Radiol.* **88** 20150056

Trimpl M J, Primakov S, Lambin P, Stride E P J, Vallis K A and Gooding M J 2022 Beyond automatic medical image segmentation-the spectrum between fully manual and fully automatic delineation *Phys. Med. Biol.* **67**

Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R and van Elmpt W 2020 Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy *Phys. Imaging Radiat. Oncol.* **13** 1–6

van der Veen J, Willems S, Bollen H, Maes F and Nuyts S 2020 Deep learning for elective neck delineation: More consistent and time efficient *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **153** 180–8

Weston A D, Korfiatis P, Philbrick K A, Conte G M, Kostandy P, Sakinis T, Zeinoddini A, Boonrod A, Moynagh M, Takahashi N and Erickson B J 2020 Complete abdomen and pelvis segmentation using U-net variant architecture *Med. Phys.* **47** 5609–18

Zhu J, Zhang J, Qiu B, Liu Y, Liu X and Chen L 2019 Comparison of the automatic segmentation of multiple organs at risk in CT images of lung cancer between deep convolutional neural network-based and atlas-based techniques *Acta Oncol. Stockh. Swed.* **58** 257–64

Table 1: Patient and scan information for all patients, the patients scanned on Siemens scanners, and for patients scanned on GE scanners. For cases where data was lost during scan transfer or anonymization, "Unknown" is listed.

	Siemens *(n=186)*	**GE *(n=219)***	**Philips *(n=26)***
Patient sex, n ***Female / Male / Unknown***	81 / 99 / 6	82 / 130 / 7	14 / 12 / 0
Patient age, years **Median [range]**	64 *[22, 91]*	65 *[33, 88]*	60.5 *[33, 86]*
Patient weight, kg **Median [range]**	76.7 *[42.2, 125.2]*	79.0 *[36.0, 141.0]*	77.4 *[47.4, 121.0]*
Scanner model	Biograph Models 1023, 1024 *(n=45)* Biograph HiRes Model 1080 *(n=40)* Biograph 16 *(n=39)* Biograph TrueV Models 1093, 1094 *(n=22)* Biograph64_mCT 4R *(n=14)* Biograph64 *(n=14)* Biograph128 *(n=5)* Biograph 6 *(n=4)* Emotion Duo Model 1062 *(n=3)*	Discovery STE *(n=71)* Discovery ST *(n=60)* Discovery LS *(n=29)* Discovery 710 *(n=27)* Discovery MI *(n=19)* Discovery MI DR *(n=6)* Discovery RX *(n=5)* Discovery 690 *(n=2)*	GEMINI TF TOF 16 *(n=9)* GEMINI TF Big Bore *(n=8)* Guardian Body *(n=4)* Allegro Body *(n=3)* Vereos PET/CT *(n=2)*
Slice thickness, mm	5.0 *(n=121)* 4.0 *(n=57)* 3.0 *(n=8)*	3.75 *(n=87)* 5.00 *(n=78)* 3.27 *(n=47)* 2.50 *(n=7)*	5.0 *(n=14)* 4.0 *(n=12)*
Convolution Kernel	B30f *(n=47)* B31s *(n=33)* B30s *(n=31)* B31f *(n=25)* B40s *(n=23)* Br38f *(n=14)* B19f *(n=10)* B40f *(n=2)* B41f *(n=1)*	STANDARD *(n=111)* SOFT *(n=62)* Unknown *(n=46)*	B *(n=24)* Unknown *(n=2)*

Table 2: Segmentation performance of the multi-organ model and single-organ models. Multi-organ versus single-organ model performance was assessed with Wilcoxon paired tests. P values were Bonferroni corrected for the number of target organs (16) and the number of performance metrics (5). Significantly better performance are bolded. DSC: Dice Similarity Coefficient, ASSD: Average Symmetric Surface Distance, HSD95, 95th percentile Hausdorff distance, PPV: Positive Predictive Value.

Organ	**DSC**	**ASSD [mm]**	**HSD95 [mm]**	**PPV**	**Sensitivity**
Liver					
Multi	**0.961 [0.956, 0.966]**	**1.07 [0.98, 1.36]**	**3.63 [3.27, 4.31]**	**0.959 [0.954, 0.969]**	0.963 [0.953, 0.973]
Single	0.952 [0.947, 0.956]	1.37 [1.25, 1.79]	4.24 [3.81, 5.83]	0.953 [0.942, 0.963]	0.953 [0.942, 0.965]
p-value	<0.001	<0.001	<0.001	<0.001	0.21
Spleen					
Multi	**0.944 [0.931, 0.951]**	**0.85 [0.72, 1.05]**	**3.09 [2.5, 3.87]**	**0.944 [0.927, 0.959]**	0.944 [0.933, 0.957]
Single	0.93 [0.896, 0.941]	0.98 [0.85, 1.62]	3.34 [2.93, 4.98]	0.939 [0.914, 0.957]	0.93 [0.91, 0.946]
p-value	<0.001	<0.001	0.037	0.003	1.0
Lung					
Multi	**0.969 [0.961, 0.975]**	**0.82 [0.63, 1.17]**	2.85 [2.01, 3.37]	0.967 [0.959, 0.975]	0.974 [0.965, 0.979]
Single	0.968 [0.955, 0.974]	0.89 [0.62, 1.33]	2.83 [2.18, 4]	**0.969 [0.958, 0.975]**	0.969 [0.954, 0.976]
p-value	<0.001	0.002	0.18	<0.001	1.0
Thyroid					
Multi	**0.748 [0.705, 0.812]**	**1.31 [1.07, 1.81]**	**4.69 [3.27, 5.93]**	0.805 [0.744, 0.854]	0.715 [0.64, 0.814]
Single	0.675 [0.603, 0.725]	1.66 [1.43, 2.24]	4.89 [4.12, 7.84]	**0.827 [0.738, 0.87]**	0.604 [0.528, 0.657]
p-value	<0.001	<0.001	0.014	<0.001	1.0
Kidney					
Multi	**0.924 [0.912, 0.937]**	**0.88 [0.77, 1.12]**	**3.07 [2.52, 4]**	**0.92 [0.899, 0.94]**	0.935 [0.905, 0.951]
Single	0.904 [0.865, 0.924]	1.1 [0.9, 1.77]	3.4 [2.76, 5.84]	0.914 [0.882, 0.938]	0.919 [0.863, 0.933]
p-value	<0.001	<0.001	0.001	<0.001	1.0

Pancreas					
Multi	**0.788 [0.721, 0.84]**	**1.93 [1.52, 3.03]**	**6.3 [4.62, 12.48]**	**0.84 [0.749, 0.89]**	0.78 [0.738, 0.845]
Single	0.67 [0.584, 0.741]	3.48 [2.78, 6.37]	13.1 [7.9, 27.68]	0.777 [0.634, 0.88]	0.603 [0.512, 0.709]
p-value	<0.001	<0.001	0.013	<0.001	0.5
Bladder					
Multi	**0.871 [0.791, 0.923]**	**1.36 [1.07, 2.04]**	**4.25 [3.27, 6.51]**	**0.892 [0.752, 0.946]**	**0.901 [0.787, 0.942]**
Single	0.794 [0.697, 0.878]	2.19 [1.72, 2.79]	7.09 [4.66, 9.03]	0.839 [0.67, 0.916]	0.822 [0.69, 0.894]
p-value	<0.001	<0.001	0.017	0.004	0.034
Aorta					
Multi	**0.919 [0.91, 0.928]**	**1.06 [0.94, 1.13]**	**3.29 [2.81, 4.19]**	**0.915 [0.896, 0.93]**	0.926 [0.906, 0.941]
Single	0.889 [0.873, 0.907]	1.44 [1.24, 1.63]	4.27 [3.91, 5]	0.909 [0.868, 0.924]	0.878 [0.862, 0.905]
p-value	<0.001	<0.001	0.015	<0.001	0.75
Adrenals					
Multi	**0.67 [0.609, 0.725]**	**1.36 [1.08, 2.33]**	**5.02 [3.52, 8.6]**	0.748 [0.676, 0.814]	0.619 [0.555, 0.685]
Single	0.551 [0.446, 0.614]	2.2 [1.71, 3.5]	7.7 [5.89, 12.84]	**0.788 [0.7, 0.828]**	0.418 [0.326, 0.501]
p-value	<0.001	<0.001	0.001	<0.001	1.0
Bowel					
Multi	**0.903 [0.88, 0.922]**	**1.47 [1.16, 2.02]**	**4.51 [3.53, 6.96]**	**0.907 [0.865, 0.93]**	**0.922 [0.888, 0.934]**
Single	0.851 [0.807, 0.874]	2.44 [1.96, 3.31]	9.22 [6.94, 13.49]	0.88 [0.832, 0.903]	0.864 [0.789, 0.889]
p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Stomach					
Multi	**0.902 [0.873, 0.928]**	**1.71 [1.17, 2.26]**	**4.8 [3.87, 9.56]**	**0.924 [0.886, 0.944]**	0.91 [0.842, 0.937]
Single	0.824 [0.739, 0.877]	2.88 [2.01, 5.52]	11.69 [5.74, 23.06]	0.905 [0.837, 0.942]	0.804 [0.621, 0.858]
p-value	<0.001	<0.001	<0.001	<0.001	1.0
Heart					
Multi	**0.941 [0.927, 0.948]**	**1.46 [1.23, 1.99]**	**4.86 [3.52, 6.38]**	**0.944 [0.929, 0.961]**	0.94 [0.922, 0.956]
Single	0.925 [0.904, 0.937]	1.88 [1.47, 2.51]	5.84 [4.25, 7.74]	0.938 [0.906, 0.957]	0.927 [0.896, 0.95]
p-value	<0.001	<0.001	<0.001	0.01	0.67
Eyes					
Multi	**0.854 [0.835, 0.878]**	1.08 [0.87, 1.31]	3.27 [2.71, 3.92]	**0.882 [0.828, 0.911]**	0.848 [0.829, 0.865]
Single	0.84 [0.817, 0.855]	1.23 [0.97, 1.43]	3.27 [2.86, 4.01]	0.872 [0.837, 0.91]	0.815 [0.783, 0.837]
p-value	0.001	0.57	0.23	0.001	1.0
Salivary					
Multi	**0.818 [0.79, 0.847]**	**1.48 [1.37, 1.68]**	**4.13 [3.82, 5.15]**	**0.833 [0.782, 0.874]**	**0.823 [0.781, 0.848]**
Single	0.759 [0.722, 0.783]	2.28 [1.95, 2.55]	6.11 [5.09, 7.03]	0.77 [0.686, 0.821]	0.757 [0.714, 0.795]
p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Pituitary					
Multi	**0.459 [0.362, 0.596]**	**1.66 [1.18, 2.18]**	**4.17 [3.27, 4.89]**	0.57 [0.461, 0.775]	**0.424 [0.34, 0.628]**
Single	0.229 [0.116, 0.352]	2.68 [1.66, 3.43]	4.84 [3.77, 6.93]	**1 [0.804, 1]**	0.132 [0.062, 0.219]
p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Choroid Plexus					
Multi	**0.537 [0.445, 0.637]**	2.05 [1.63, 2.75]	5.49 [4.43, 9.57]	**0.608 [0.515, 0.676]**	0.522 [0.383, 0.645]
Single	0.336 [0.276, 0.444]	3.33 [2.49, 4.25]	8.45 [7.6, 12.18]	0.418 [0.336, 0.6]	0.284 [0.208, 0.39]
p-value	<0.001	0.32	1.0	<0.001	0.43

Table 3 Comparison of segmentation performance between previous studies and the current study literature as quantified by Dice similarity coefficient (DSC). Note that while past studies typically report mean DSC, we report Median DSC, as distributions of DSC were not normally distributed in our study.

Organ	Mean DSC values in past studies	Median DSC values in current study
Liver	**0.89** (Zhu *et al* 2019) **0.95** (Weston *et al* 2020) **0.95** (Gibson *et al* 2018) **0.952** (Rister *et al* 2020)	0.96
Spleen	**0.96** (Gibson *et al* 2018) **0.97** (Isensee *et al* 2021)	0.94
Lung	**0.95** (Rister *et al* 2020) **0.95** (Zhu *et al* 2019) **0.958** (Mirando *et al* 2018)	0.97
Thyroid	**0.89** (Chung *et al* 2021)	0.75
Kidney	**0.91** (Jackson *et al* 2018) **0.918** (Rister *et al* 2020) **0.93** (Weston *et al* 2020) **0.93** (Lamba *et al* 2019) **0.95** (Gibson *et al* 2018)	0.92
Pancreas	**0.78** (Gibson *et al* 2018) **0.79** (Weston *et al* 2020) **0.82** (Isensee *et al* 2021)	0.79

	0.85 (Sundar *et al* 2022)	
Bladder	**0.77**(Rister *et al* 2020) **0.86** (Sundar *et al* 2022) **0.932** (Schreier *et al* 2020)	0.87
Aorta	**0.92** (Haq *et al* 2020)	0.92
Adrenals	**0.69** (Weston *et al* 2020) **0.72** (Sundar *et al* 2022) **0.84** (Robinson-Weiss *et al* 2023)	0.67
Bowel	**0.65** (Men *et al* 2017) **0.88** (Gonzalez *et al* 2021)	0.90
Stomach	**0.89** (Gibson *et al* 2018)	0.90
Heart	**0.95** (Chung *et al* 2021)	0.94
Salivary Glands	**0.81**(Park *et al* 2021) **0.86** (Hänsch *et al* 2019)	0.82

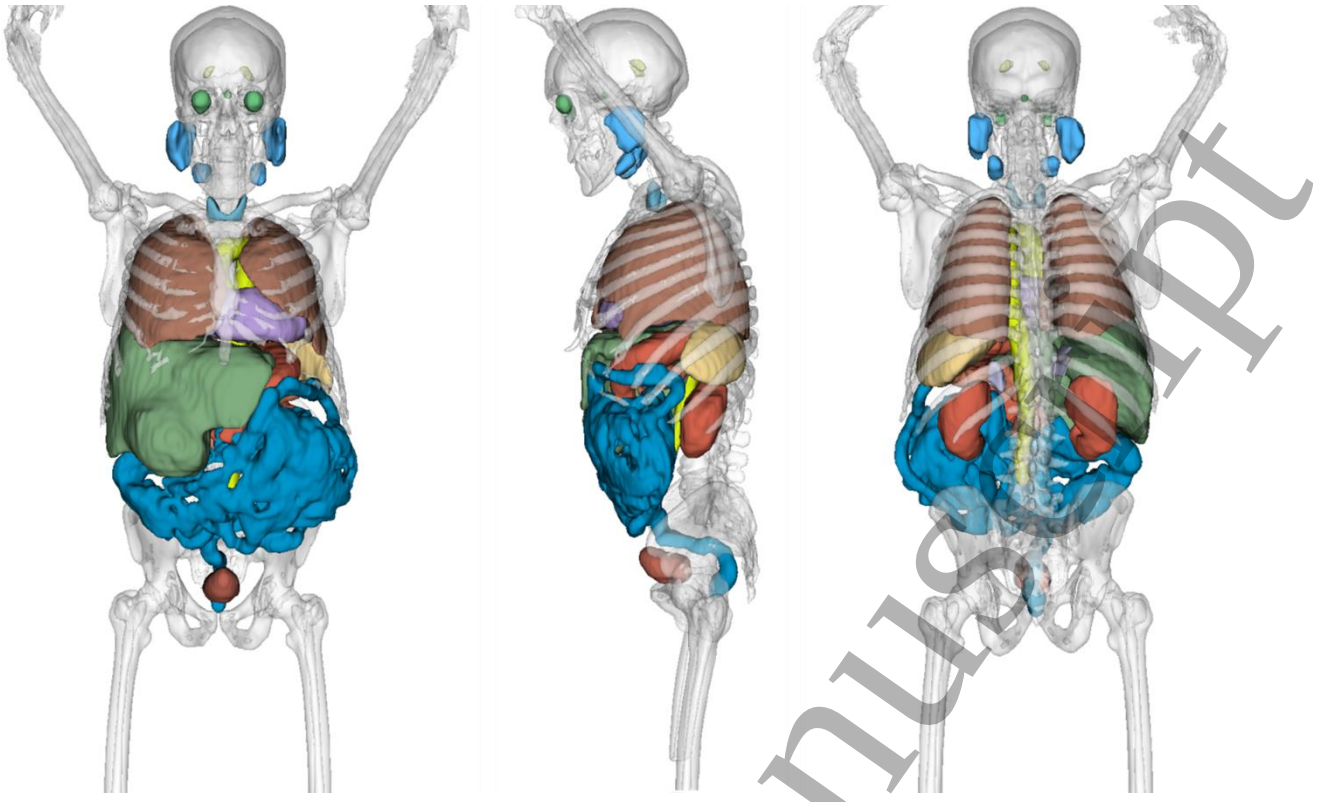

Figure 1: Example organ segmentations produced by the multi-organ CNN on a randomly selected patient from the test set. The CNN produced all colored organ renderings. A 150 HU isocontour was produced and is shown for visualization purposes in grey.

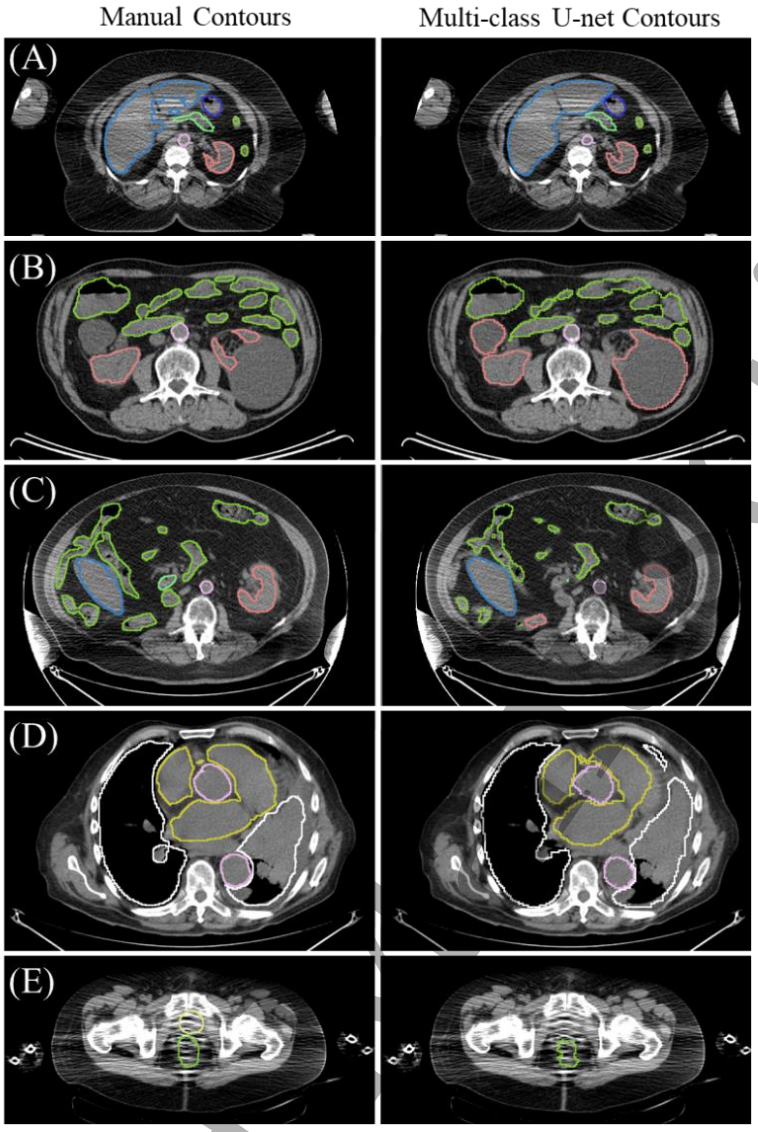

Figure 2: Examples of cases with poor performance comparing the manual ground truth contours (left) with the multi-class U-net contours (right). (A) The case with the lowest liver DSC performance due to imperfections of the manual contours. (B) A case with poor kidney performance due to a large cyst on the left kidney. (C) Poor bowel performance is found in this example case which may be due to the abnormal presence of omental/peritoneal fat and the collection of abnormal peritoneal fluids. (D) A case with poor lung segmentation due to lung disease. (E) Severe CT artifacts result in the U-net not contouring the bladder, while the manual contourer may have relied on the corresponding PET image to identify the bladder in this difficult case.

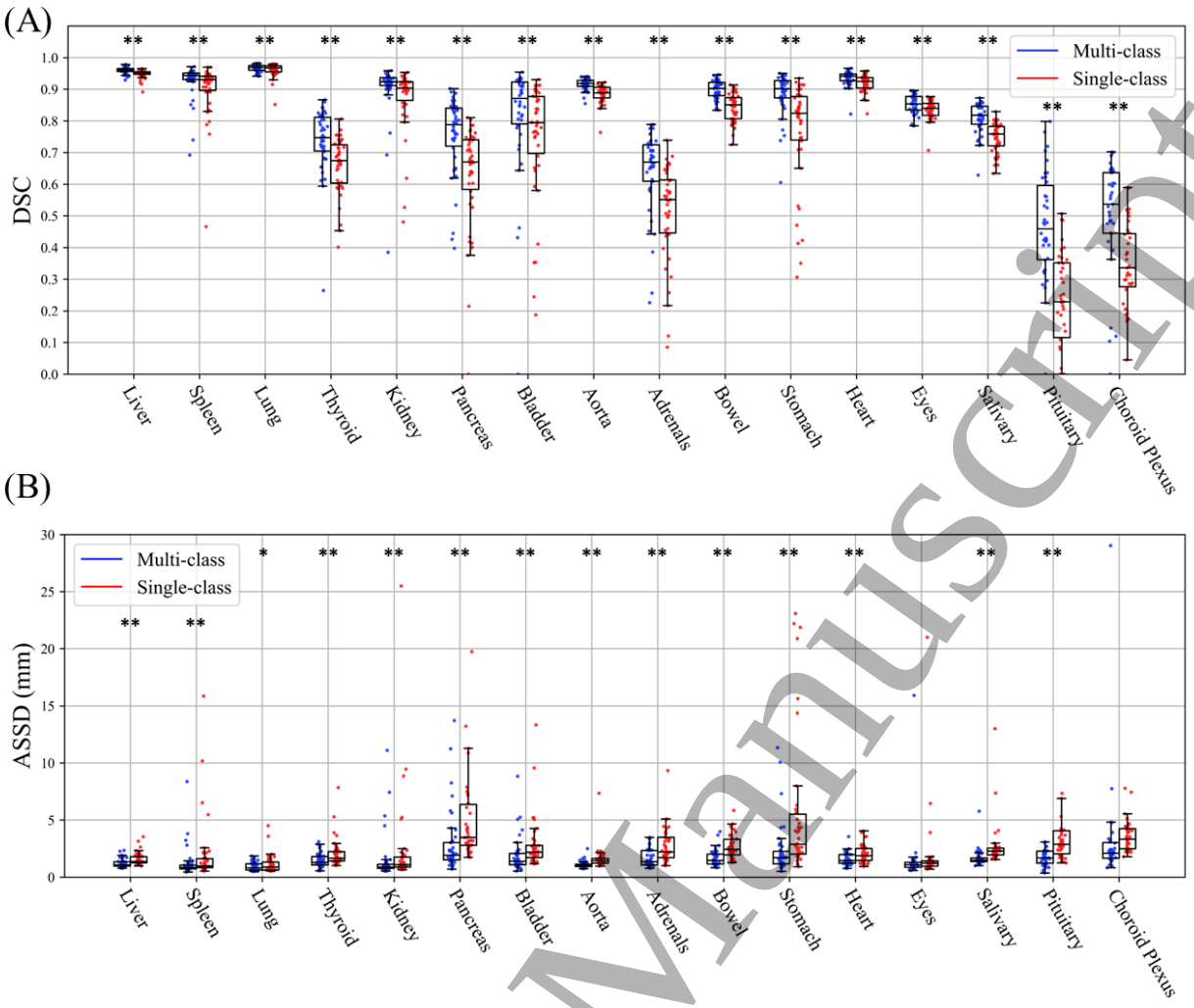

Figure 3: Comparison of performance of the multi-class versus the single-class organ segmentation models in the held-out test dataset of 42 images. Displayed performance metrics are A) Dice Similarity Coefficient (DSC) and B) Average Symmetric Surface Distance (ASSD). Significant differences between DSC values are indicated by ** for $p<0.001$, and * for $p<0.05$.

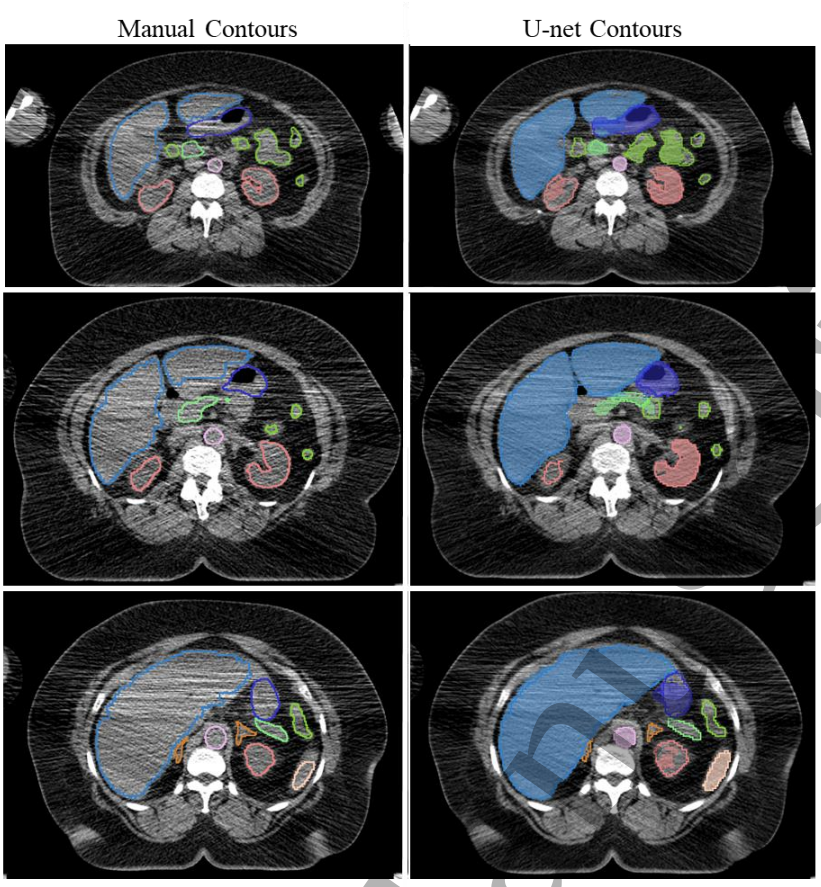

Figure 4: Comparison of manual ground truth contours (left) to both the multi-class CNN (outline) and the single-class CNN (filled in colors) for three slices in the abdomen of a single test patient. Areas of errors are often found along the border of two structures (e.g. liver and stomach). The multi-class CNN had substantially better segmentation of the adrenal glands for this case (DSC=0.09 for single-class model, DSC = 0.66 for multi-class model).

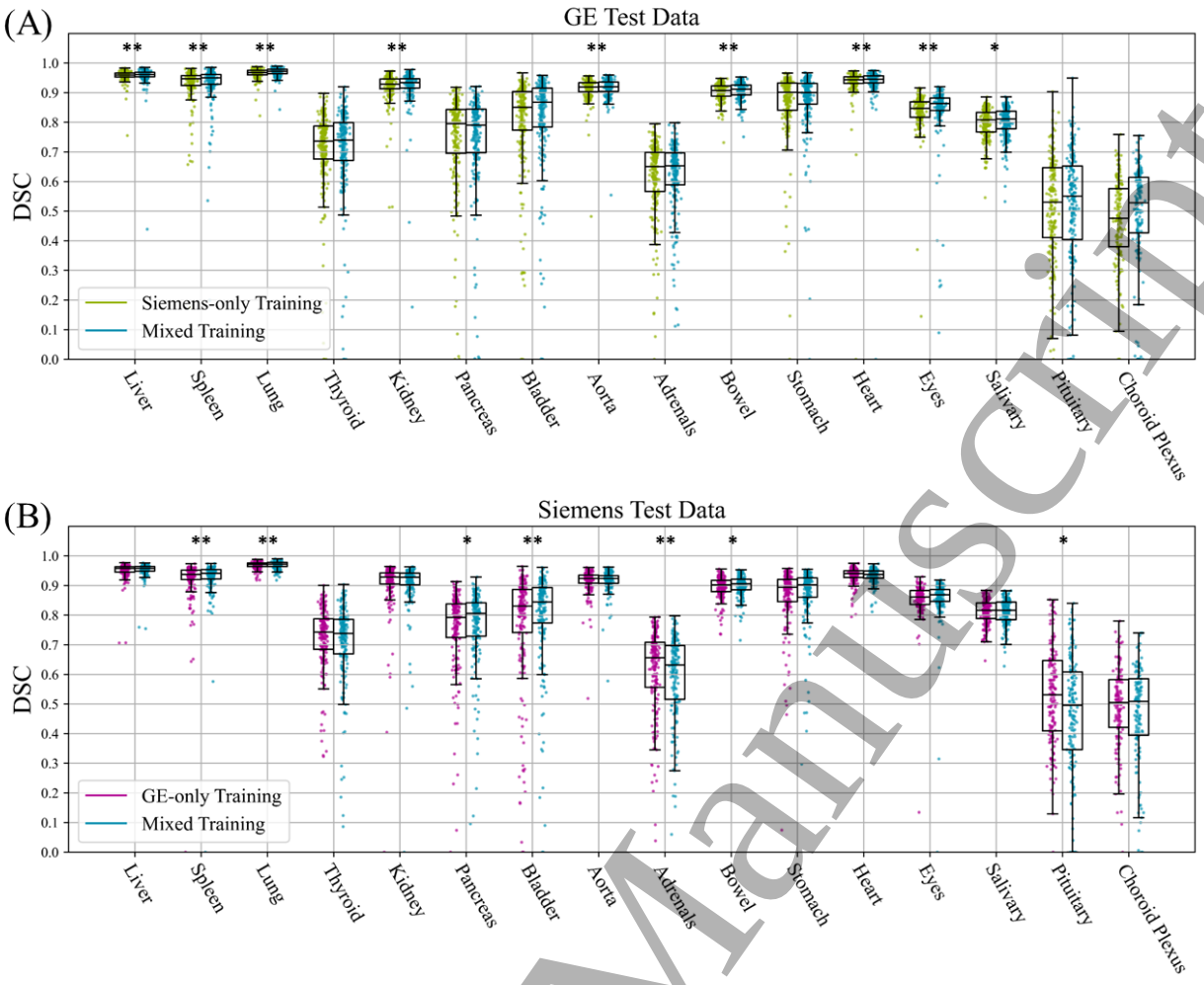

Figure 5: Comparison of manufacturer-specific versus manufacturer-mixed organ segmentation models in the (A) 219 images acquired on GE scanners and (B) 186 images acquired on Siemens scanners. The performance metric displayed is the Dice Similarity Coefficient (DSC). Significant differences between DSC values are indicated by ** for $p<0.001$, and * for $p<0.05$.