scientific reports

OPEN

Evaluation of a hybrid pipeline for automated segmentation of solid lesions based on mathematical algorithms and deep learning

Liam Burrows[1]✉, Ke Chen[1]✉, Weihong Guo[2], Martin Hossack[4], Richard G. McWilliams[3] & Francesco Torella[4]

We evaluate the accuracy of an original hybrid segmentation pipeline, combining variational and deep learning methods, in the segmentation of CT scans of stented aortic aneurysms, abdominal organs and brain lesions. The hybrid pipeline is trained on 50 aortic CT scans and tested on 10. Additionally, we trained and tested the hybrid pipeline on publicly available datasets of CT scans of abdominal organs and MR scans of brain tumours. We tested the accuracy of the hybrid pipeline against a gold standard (manual segmentation) and compared its performance to that of a standard automated segmentation method with commonly used metrics, including the DICE and JACCARD and volumetric similarity (VS) coefficients, and the Hausdorff Distance (HD). Results. The hybrid pipeline produced very accurate segmentations of the aorta, with mean DICE, JACCARD and VS coefficients of: 0.909, 0.837 and 0.972 in thrombus segmentation and 0.937, 0.884 and 0.970 for stent and lumen segmentation. It consistently outperformed the standard automated method. Similar results were observed when the hybrid pipeline was trained and tested on publicly available datasets, with mean DICE scores of: 0.832 on brain tumour segmentation, and 0.894/0.841/0.853/0.847/0.941 on left kidney/right kidney/spleen/aorta/liver organ segmentation.

Volumetric assessment of solid lesions has been revolutionised by modern cross-sectional imaging. Presently, such assessment is performed by radiologists using manual or semi-automated segmentation tools provided by various software packages. This process can be time-consuming, particularly when comparative analysis of serial scans is required. Fully automated segmentation could, potentially, greatly accelerate this process and improve its accuracy by removing the bias associated with operator dependency. Recent work into image segmentation has focused on two main types: model-driven[1–3] and data-driven methods[4–6]. Model-driven methods utilise mathematical (variational) formulations, which have been shown to provide a powerful framework for segmenting images. A variational method for segmentation can be composed of a number of terms, each hand crafted to achieve a desired result in the output; essentially image features desirable to a particular segmentation task can be built into the model (for example a desired segmented object may be found using a combination of a particular image intensity, shape, and/or location in the image), and the resultant algorithm is run independently on each image (for example: a computed tomography slice).

Data-driven methods have drawn attention in recent years due to the rise of deep learning and powerful computing hardware. Segmentation methods using deep learning methods such as convolutional neural networks (CNNs), are considered the gold standard, if provided with enough data.

One challenge for deep learning segmentation is the acquisition of an appropriately large dataset with ground truth labels, which can be both an expensive and time-consuming process. Ground truth labels are usually manual segmentations of the region of interest performed by an expert. They are necessary for training deep learning

[1]Centre for Mathematical Imaging Techniques and Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZL, UK. [2]Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University, Cleveland, OH 44106, USA. [3]Department of Radiology, Liverpool University Hospitals, Liverpool, UK. [4]Liverpool Vascular and Endovascular Service, Liverpool University Hospitals NHS Foundation Trust, Liverpool, UK. ✉email: liam.burrows@liv.ac.uk; k.chen@liv.ac.uk

algorithms and for evaluating their performance. To address this problem, hybrid approaches to segmentation have been proposed, whereby variational methods are used to supplement manual ground truth labels[7–9], thus reducing the demand for manual hand drawn segmentation.

We have developed our own fully automated hybrid approach (a hybrid segmentation pipeline) for volumetric segmentation, which combines a recently developed variational model[10] with a deep learning algorithm. In this work, we tested the accuracy of this hybrid segmentation pipeline.

Methods

Reference standards. To test our pipeline, we chose contrast-enhanced computed tomography (CT) angiograms of patients being followed up after endovascular repair (EVAR) of abdominal aortic aneurysms (AAAs), as the stented aorta is an example of a solid lesion that requires serial follow-up with cross-sectional imaging. We chose post-EVAR CTs because serial comparison of scans is a common diagnostic problem, which can be time-consuming and labour intensive. The choice was also due, in part, with familiarity (three authors are vascular interventionists), and to the fact that our group had previously used and tested a reproducible manual segmentation technique on these scans[11,12]. CTs were performed with a 64-slice Siemens Somatom scanner (Siemens Healthcare, Frimley, UK). Manual segmentation was performed by one of the authors (MH) on reconstructed 2 mm thick slices with intervals of 2 mm, according to a previously described technique[11–13].

We acquired a total of 70 fully anonymised postoperative CTs, of which 50 were manually segmented to provide a "ground truth" for training the deep learning algorithm, and to provide useful evaluation metrics. The manual segmentation ("ground truth labelling") was conducted between the lowermost renal artery and the aortic bifurcation, by hand, using an open source application called ITK-SNAP[14]. This segmentation was considered the reference standard.

We used a typical 60:20:20 ratio for training, validation, and testing the images: 30 sets to train the deep learning part, 10 sets to validate and prevent overfitting, and reserved the final 10 sets for evaluation purposes. Our pipeline also used 20 unlabelled datasets during the training phase, providing us with 50 volumes in total during training.

Additionally, we evaluated the pipeline on two publicly available datasets: The Brain Tumour Segmentation challenge (BraTS)[15–17], and the Abdomen data from the Multi-Atlas Labelling Beyond the Cranial Vault challenge[18]. The BraTS dataset contains a range of MR modalities, but for our purposes we considered only the fluid attenuated inversion recovery (FLAIR) volumes. All sets have been labelled by one to four raters following the same protocol, and their annotations approved by experienced neuro-radiologists. We segmented only the tumour region in each volume. We used a total of 200 volumes for the BraTS data: 120 during training, 40 for validation and 40 for evaluation purposes. In addition to the BraTS, we used the Abdomen dataset, a collection of CT scans of the abdomen in which 13 organs have been segmented by two experienced undergraduate students and the segmentation verified by a radiologist. We evaluated our pipeline on five of the 13 organs: the spleen, the right kidney, the left kidney, the liver and the aorta. The Abdomen dataset contains 30 scans; of these, we used 15 for training, 5 for validation and 10 for evaluation.

Pipeline tests. After windowing and selecting the uppermost and lowermost slice for segmentation on either a post-EVAR aneurysm or a selected organ from the BraTS or Abdomen datasets, we ran our variational model[10]. This original model uses an enhanced method of edge detection, which allows for images containing low contrast to be segmented effectively. Following edge detection, the region of interest is segmented based on image intensity and pixel location in the image[10]. The variational method provided us with a good but not perfect initial segmentation, as some regions may contain no contrast at the boundary. Furthermore, artifacts may result in poor definition in certain areas.

Although it is possible to obtain accurate segmentation by using the variational method only, this is a time-consuming process, taking up to 20 min for a large volume. Further, each new volume would require a user to manually insert a set of markers to indicate the region of interest. We eliminated this step entirely by using image registration. In practice, we ran the variational segmentation model for one 3D volume (aneurysm, tumour or organ), obtaining an initial segmentation. For each subsequent scan we simply superimposed this segmentation to the new scan, thereby registering the saved segmentation onto the new image, rather than re-running the variational method on each new image. The overlapped images were then registered by a previously trained network[19], which, although not entirely accurate, produced an estimate that could be fed to the CNN, to produce an accurate final result. This image with the initial segmentation was not excluded, and maintained a place in the training set. This registration step removed the need for user interaction, making the method automatic.

The final stage involved feeding each estimated segmented volume to the CNN. Unlike in commonly used methods of segmentations by CNNs[4,6], ours received both the scan and the estimated segmented volume (the output of the variational model). This provided the CNN with supplementary information to produce an accurate final result. The CNN was trained in a standard way using backpropagation (see appendix for details).

"Unlabelled" scans (i.e., image volumes without manual segmentation) were also used to train our pipeline. The estimated volume from the variational method was used here in place of the reference standard. This allowed us to expand our training set, exposing the network to more data without needing more time-consuming manual labeling. This is commonly known as a semi-supervised approach to learning (see appendix).

Data analysis and presentation. The accuracy of an automated segmentation pipeline depends on its ability to correctly identify all voxels, hence volume, belonging to an organ/lesion in a scan, as well as those that lie outside said organ/lesion. To evaluate the accuracy of our pipeline against manual segmentation we reported true positives (TP) as the number of voxels correctly identified by a segmentation method; false positives (FP) and false negatives (FN), as the number of voxels incorrectly identified/excluded by a segmentation method; sensitivity as the ability of a model to correctly identify all relevant voxels or volume, and false negative rate (the inverse of sensitivity). We did not report true negatives, as these are largely dependent on the total number of voxels included in a scan (usually a whole section of the body), hence values are always very high. Continuous variables were expressed as median and range, as they were generally not normally distributed. Correlation was evaluated by graphical methods and agreement with Bland–Altman plots[20].

We also used more commonly used segmentation metrics to evaluate the pipeline, including the DICE, the JACCARD and the volumetric similarity (VS) coefficients, which are defined as:

$$DICE = \frac{2TP}{2TP + FP + FN}, \quad JACCARD = \frac{TP}{TP + FP + FN}, \quad VS = 1 - \frac{|FN - FP|}{2TP + FP + FN}.$$

Finally, the Hausdorff Distance (HD) between a segmented volume X and a ground truth segmentation GT is defined as:

$$HD(X, GT) = \max\big(h(X, GT), h(GT, X)\big),$$

where $h(X, GT)$ is the directed Hausdorff distance given by:

$$h(X, GT) = \max_{x \in X} \min_{y \in GT} |x - y|,$$

where $|x - y|$ is the Euclidean distance between two points, where x is in X and y is in GT.

Both DICE and JACCARD scores range between 0 (where no overlap between output and reference occurs) and 1 (where the output is exactly the reference segmentation). VS represents how similar the volume of the segmented output is to the volume of the reference segmentation. This is not influenced by the overlap: two organs with the exact same volume but in different positions would result in a VS score of 1. Finally, HD describes the largest distance from one volume to the nearest point in the other. Unlike the other metrics described here, the ideal result of HD is a score of 0, as the voxels of an ideal segmentation would be in the same place as those of the reference standard.

In order to compare our pipeline to a more traditional method of automated segmentation, we trained a deep learning model in a typical fashion (using only the scan as input, without the input of our variational model). This method is referred as "standard" in our results.

As further experiments, we trained the hybrid method with a decreasing number of data in the training set in order to determine if the hybrid approach can perform better than the standard method when provided with less data. In addition we run the hybrid method on some simplified networks with lighter architectures (see appendix).

Ethical approval. The study was conducted in accordance with relevant institutional guidelines and regulations.

Results and discussion

Results.

Aortas. Figure 1 displays a 2D example output of both approaches, showing various cross-sections of a 3D volume. On manual segmentation, the median (range) thrombus volume was 185 (129–531) ml, corresponding to 179,303 (111,336–588,375) voxels, whereas the stent and lumen volume was 59 (45–88) ml, corresponding to 55,717 (36,117–102,904) voxels. The accuracy of the "standard" and hybrid pipelines is reported in Tables 1 and 2. Notably, the hybrid method had slightly more TPs, less FPs and more FNs, resulting in higher DICE, JACCARD and VS for the whole aneurysm and the thrombus. Correlation and agreement between the measurements of the two pipelines is displayed in Fig. 2.

Abdomen/BraTS. Results of the segmentation of abdominal organs are displayed in Tables 5 and 6. For all organs, the hybrid pipeline was more accurate than the standard, producing more true positives and fewer false positives or negatives than the standard one, except for the spleen, in whose segmentation the hybrid pipeline resulted in a higher median number of false negatives. Similarly, the hybrid pipeline also outperformed the standard one in the segmentation of brain tumours (Tables 3, 4).

Further experiments. Results from further experiments can be found in the appendix. Of particular note was the experiment which trained networks using the hybrid approach with the three databases, but with reduced data. Mean DICE values for the thrombus/stent and lumen/whole aneurysm were 0.877/0.927/0.903 when 20 supervised volumes were in the training data (instead of 30 in the original experiments), outperforming the standard method with full data. Similar trends were seen with the BraTS dataset with only 60 volumes (instead of 120) giving mean DICE values of 0.761 and in the Abdomen dataset trained with 10 (instead of 20) giving mean DICE values of 0.867/0.807/0.841/0.823/0.922 for left kidney/right kidney/spleen/aorta/liver segmentation.

Discussion.

Our results suggest that the combination of a mathematical (variational) approach with a deep learning algorithm may improve the accuracy of automated segmentation of organs and tumours on CT.

For general purposes, variational and deep learning segmentation have been studied widely, though limited work has been conducted into hybrid approaches[7–9]. Recently, deep learning has become the preferred approach to segmentation and can produce outstanding results on a wide variety of applications[21–23]. Wang et al.[24,25] used a similar approach to ours, albeit fully based on deep learning. In their proposed method, an initial network

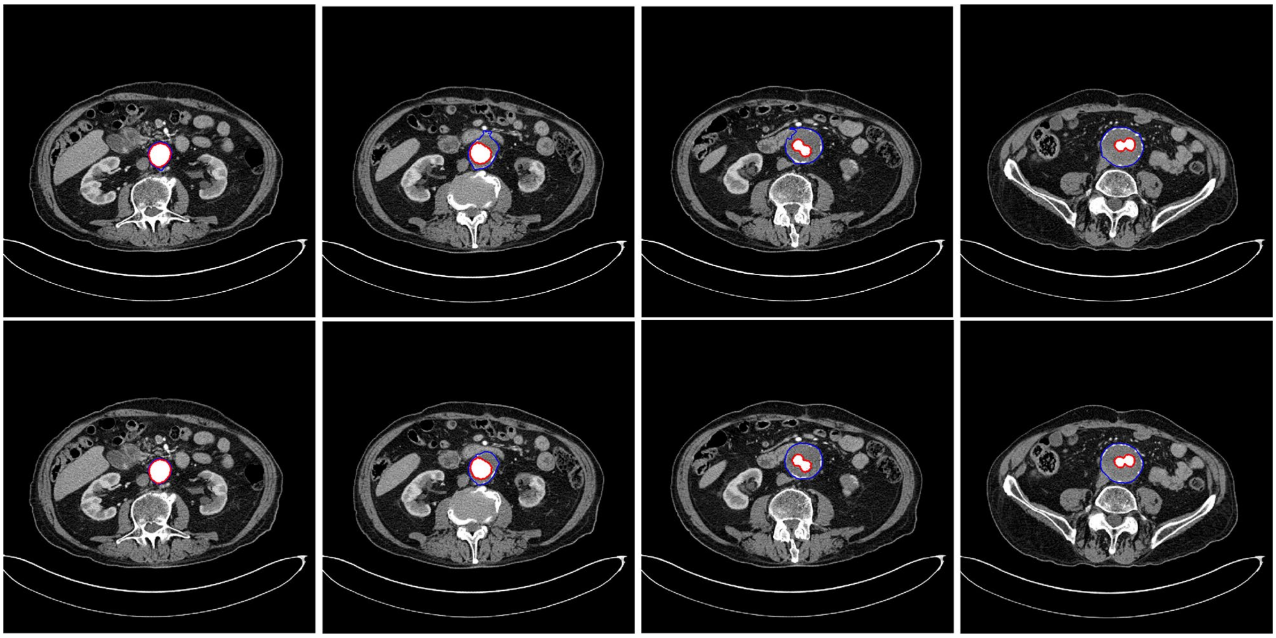

Figure 1. An example of segmentation by the two pipelines, with axial cross-sections of an aneurysm at four different levels. The standard pipeline is shown on the top row, with the hybrid pipeline on the bottom row. Here the standard pipeline is unable to accurately follow the contour of the aneurysm. The hybrid pipeline appears accurate.

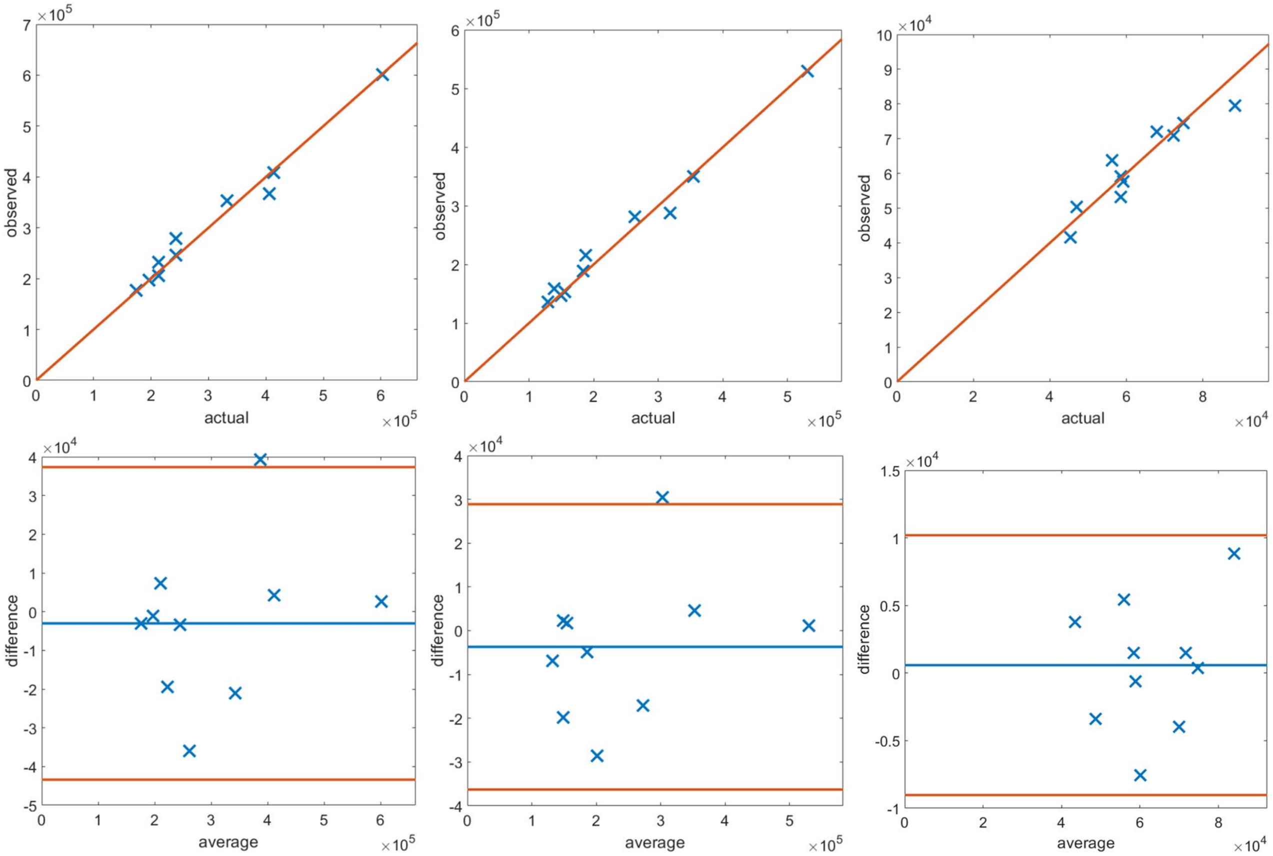

Figure 2. Correlation (top row) and agreement (bottom row) between the proposed hybrid approach and the ground truth in voxel detection for the aortic data. Units are the number of voxels.

Thrombus:	Standard	Hybrid
As measured	180,651 (138,491–590,159)	202,160 (135,817–529,732)
TP	160,358 (120,456–493,880)	174,384 (119,911–503,438)
FN	11,708 (3631–167,718)	13974 (4484–56,037)
FP	21,378 (7768–96,279)	19994 (9306–46,248)
Stent and lumen		
As measured	60,190 (45,333–74,340)	61,499 (41,540–79,582)
TP	56,863 (42,893–72,441)	56,187 (40,472–74,887)
FN	2074 (220–39,972)	2726 (1504–13,532)
FP	2868 (891–8112)	2680 (1045–9293)
Whole aneurysm		
As measured	247,781 (183,816–662,998)	262,863 (177,255–600,584)
TP	220,060 (167,745–566,548)	233,612 (167,327–575,127)
FN	9329 (2283–206,380)	13,762 (4582–66,201)
FP	20,447 (7345–96,450)	17,912 (7500–50,589)

Table 1. Standard and Hybrid pipeline accuracy in aneurysm segmentation. Units displayed are volume (mm^3), given as median (range).

Thrombus	VM	Standard	Hybrid
DICE	0.734 ± 0.088	0.873 ± 0.094	0.909 ± 0.054
JACCARD	0.587 ± 0.107	0.784 ± 0.132	0.837 ± 0.087
VS	0.866 ± 0.117	0.927 ± 0.095	0.972 ± 0.026
HD	41.0 ± 11.4	95.7 ± 81.5	60.4 ± 67.1
TPR	0.660 ± 0.104	0.893 ± 0.150	0.922 ± 0.048
TNR	0.999 ± 0.001	0.999 ± 0.0001	0.999 ± 0.0001
Time (s)	345 ± 31.5	11.6 ± 0.258	11.6 ± 0.258
Stent and lumen			
DICE	0.923 ± 0.025	0.928 ± 0.082	0.937 ± 0.028
JACCARD	0.857 ± 0.044	0.875 ± 0.123	0.884 ± 0.045
VS	0.982 ± 0.012	0.955 ± 0.086	0.970 ± 0.022
HD	56.6 ± 68.2	60.7 ± 69.9	49.0 ± 60.4
TPR	0.942 ± 0.027	0.928 ± 0.134	0.935 ± 0.044
TNR	0.999 ± 0.0001	0.999 ± 0.0001	0.999 ± 0.0001
Time (s)	332 ± 38.7	11.6 ± 0.258	11.6 ± 0.258
Whole aneurysm			
DICE	0.803 ± 0.058	0.902 ± 0.093	0.933 ± 0.041
JACCARD	0.674 ± 0.079	0.831 ± 0.133	0.877 ± 0.070
VS	0.903 ± 0.081	0.935 ± 0.092	0.976 ± 0.024
HD	55.8 ± 52.4	109 ± 80.0	52.4 ± 55.5
TPR	0.740 ± 0.075	0.918 ± 0.151	0.942 ± 0.043
TNR	0.999 ± 0.001	0.999 ± 0.0001	0.999 ± 0.0001
Time (s)	678 ± 58.5	11.6 ± 0.258	11.6 ± 0.258

Table 2. Performance on segmentation metrics of the variational method (VM) only, and standard and hybrid pipeline in aneurysm segmentation. Units given as mean ± standard deviation.

produces an initial segmentation; if this is not desirable a second network is used to refine the result. Our proposed method is similar, but, in place of a network, the first step is a variational algorithm, which we recently developed specifically to improve segmentation in the presence of low contrast[10].

Use of the variational method is innovative, and it allows us greater control over the initial segmentation when compared with the work by Wang et al.[24,25]. Further, a major weakness of deep learning methods is their heavy reliance on large, labelled datasets. As we can rely on the variational method to provide a reasonably good initial segmentation, we can also use unlabelled data in the training stage, thus expanding our training dataset. The output segmentation should therefore be more resilient to variation.

BraTS:	Standard	Hybrid
As measured	133,209 (2775–915,633)	95,233 (7253–315,035)
TP	82,413 (2775–215,828)	81,508 (7249–214,419)
FN	5480 (15–80,514)	5047 (594–57,515)
FP	14,077 (0–783,250)	8625 (5–231,803)

Table 3. Standard and Hybrid pipeline accuracy in brain tumour segmentation. Units displayed are volume (mm^3), given as median (range).

BraTS:	VM	Standard	Hybrid
DICE	0.746 ± 0.148	0.596 ± 0.308	0.832 ± 0.130
JACCARD	0.614 ± 0.159	0.490 ± 0.314	0.730 ± 0.170
VS	0.792 ± 0.156	0.633 ± 0.316	0.863 ± 0.132
HD	28.3 ± 18.0	49.4 ± 29.5	36.8 ± 20.5
TPR	0.637 ± 0.165	0.800 ± 0.262	0.884 ± 0.141
TNR	0.999 ± 0.001	0.998 ± 0.001	0.999 ± 0.001
Time (s)	46.6 ± 6.76	3.71 ± 0.102	3.73 ± 0.123

Table 4. Performance on segmentation metrics of the variational method (VM) only, and standard and hybrid pipeline in brain tumour segmentation. Units given as mean ± standard deviation.

Left kidney:	Standard	Hybrid
As measured	155,878 (70,294–222,822)	158,221 (107,374–220,349)
TP	131,585 (66,150–191,654)	123,870 (101,112–198,947)
FN	11,081 (965–52,481)	7518 (736–55,575)
FP	19,458 (4144–54,788)	15,512 (6262–46,373)
Right kidney		
As measured	197,531 (14,369–236,772)	164,799 (79,980–222,058)
TP	132,750 (0–184,937)	132,802 (69,714–188,599)
FN	17,956 (1781–124,803)	14,531 (2098–59,380)
FP	41,861 (4039–99,438)	22,718 (4145–45,210)
Spleen		
As measured	185,886 (100,546–352,992)	256,998 (94,334-1,001,976)
TP	171,358 (87,343–315,736)	177,268 (82,381–397,355)
FN	7786 (1176–159,828)	9748 (5203–98,193)
FP	20,012 (3848–43,048)	14,703 (5526–876,251)
Aorta		
As measured	81,256 (38,618–200,230)	86,442 (33,559–216,247)
TP	70,117 (12,504–148,501)	64,838 (28,381–197,198)
FN	17,419 (4046–121,927)	9859 (1057–73,310)
FP	15,885 (2880–51,729)	7722 (2407–37,472)
Liver		
As measured	1,697,921 (1,329,048–2,333,988)	1,654,217 (1,288,193–2,237,004)
TP	1,601,916 (1,024,422–2,182,113)	1,602,463 (1,025,609–2,150,051)
FN	78,860 (28,886–202,664)	65,631 (27,699–281,215)
FP	121,130 (39,374–314,559)	86,373 (40,837–262,584)

Table 5. Standard and hybrid pipeline accuracy in organ segmentation. Units displayed are volume (mm^3), given as median (range).

Left kidney:	**VM**	**Standard**	**Hybrid**
DICE	0.872 ± 0.021	0.856 ± 0.072	0.894 ± 0.077
JACCARD	0.773 ± 0.033	0.753 ± 0.104	0.815 ± 0.114
VS	0.886 ± 0.033	0.904 ± 0.074	0.954 ± 0.041
HD	5.89 ± 6.31	9.70 ± 3.04	7.47 ± 8.52
TPR	0.784 ± 0.042	0.867 ± 0.149	0.904 ± 0.114
TNR	0.999 ± 0.0001	0.999 ± 0.0001	0.999 ± 0.0001
Time (s)	26.7 ± 3.56	2.12 ± 0.160	2.13 ± 0.143
Right kidney			
DICE	0.802 ± 0.107	0.721 ± 0.265	0.841 ± 0.078
JACCARD	0.680 ± 0.132	0.608 ± 0.239	0.732 ± 0.109
VS	0.832 ± 0.067	0.818 ± 0.226	0.893 ± 0.073
HD	12.0 ± 11.5	23.8 ± 18.6	12.3 ± 6.58
TPR	0.725 ± 0.073	0.797 ± 0.293	0.839 ± 0.163
TNR	0.999 ± 0.004	0.999 ± 0.001	0.999 ± 0.0001
Time (s)	30.5 ± 4.09	2.06 ± 0.093	2.05 ± 0.069
Spleen			
DICE	0.890 ± 0.034	0.886 ± 0.065	0.853 ± 0.224
JACCARD	0.804 ± 0.051	0.801 ± 0.098	0.785 ± 0.239
VS	0.923 ± 0.031	0.922 ± 0.069	0.896 ± 0.236
HD	6.57 ± 5.08	14.2 ± 7.30	40.0 ± 52.4
TPR	0.828 ± 0.025	0.892 ± 0.130	0.922 ± 0.066
TNR	0.999 ± 0.001	0.999 ± 0.0001	0.999 ± 0.0001
Time (s)	29.7 ± 4.09	2.06 ± 0.010	2.08 ± 0.115
Aorta			
DICE	0.731 ± 0.113	0.721 ± 0.169	0.847 ± 0.124
JACCARD	0.585 ± 0.121	0.587 ± 0.198	0.750 ± 0.160
VS	0.769 ± 0.077	0.879 ± 0.102	0.936 ± 0.067
HD	6.71 ± 9.20	38.4 ± 24.7	13.7 ± 14.9
TPR	0.646 ± 0.060	0.696 ± 0.200	0.821 ± 0.162
TNR	0.999 ± 0.003	0.999 ± 0.001	0.999 ± 0.0001
Time	30.9 ± 4.71	2.04 ± 0.085	2.05 ± 0.092
Liver			
DICE	0.928 ± 0.027	0.930 ± 0.032	0.941 ± 0.028
JACCARD	0.867 ± 0.045	0.870 ± 0.055	0.890 ± 0.049
VS	0.966 ± 0.017	0.970 ± 0.034	0.975 ± 0.034
HD	13.8 ± 8.81	24.4 ± 15.2	11.8 ± 7.96
TPR	0.911 ± 0.015	0.947 ± 0.031	0.949 ± 0.037
TNR	0.996 ± 0.004	0.999 ± 0.0001	0.999 ± 0.0001
Time (s)	31.0 ± 3.29	2.04 ± 0.091	2.03 ± 0.092

Table 6. Performance on segmentation metrics of the Variational Method (VM) only, and Standard and Hybrid pipeline in organ segmentation. Units given as mean ± standard deviation.

Aortic segmentation. Automated segmentation of the abdominal aorta has been addressed before to some success. Early approaches include variational model-based methods such as level set methods (a mathematical way of representing a shape) by Loncaric et al.[26] and Subasic et al.[27,28], whereas Zohios et al.[29] used level sets and geometrical methods to segment the thrombus in the presence of calcifications. These models can be very time-consuming, and susceptible to imprecision where low contrast is present.

More recently Lalys et al.[30] proposed a fast 3D model based on the snakes model by Kass et al.[31], capable of segmenting both the lumen and thrombus but requiring some user input. While quoting a mean DICE score of 0.87 on post operative CTA scans, use of a shape based deformable model, using image registration to achieve segmentation, can suffer on unusual scans. Based on a similar idea, Lareyre et al.[32] proposed a fully automated pipeline for segmentation of AAAs, obtaining both lumen and thrombus incorporating the Chan-Vese model[2]. Evaluation of thrombus segmentation was performed on 525 selected slices from 40 CT scans, giving a mean: DICE of 0.88, Jaccard of 0.80 and Sensitivity of 0.91, which performs slightly worse than our pipeline's mean results of 0.91, 0.84 and 0.92 respectively.

Deep learning methods include that by Lopez et al.[33], who developed a fully automatic approach for segmenting the thrombus on CT scans of patients treated with Endovascular Aneurysm Repair (EVAR) using a new network architecture based on the proposed work by Xie et al.[34]. In a follow up work[35], the authors extended their work to segment 3D volumes, maintaining the fully automatic aspect. Notably, segmentation was performed both on preoperative and postoperative CT scans, with a mean DICE score of 0.89 and Jaccard of 0.81 on segmentation of the whole aneurysm on postoperative scans. Lu et al.[36] proposed a 3D pipeline using a V-Net architecture[37] combined with an ellipse fitting to estimate the maximum diameter of the aorta. Both contrast and non-contrast enhanced scans were used, quoting DICE scores of 0.89 and 0.90 on preoperative scans with and without contrast respectively. More recently the work by Caradu et al.[38] proposed a deep learning algorithm trained to segment preoperative infrarenal aortic aneurysm CT volumes effectively, with a mean DICE score of 0.95 on 100 scans. Adam et al.[39]introduced an automated method named Augmented Radiology for Vascular Aneurysm (ARVA), trained on a large dataset of 489 CT volumes (a combination of both preoperative and postoperative scans), dedicated to segmenting the entire aorta from the ascending portion to the iliac arteries, with a mean DICE score of 0.95 on preoperative scans and 0.93 on postoperative scans, thus comparable to ours, but achieved with fewer training data. This study nevertheless confirms that the use of an initial variational method can reduce the need for larger datasets.

BraTS. The BraTS dataset is a widely used imaging dataset in the literature[40–42]. All scans include several MR sequences, including T1, post-contrast T1-weighted, T2-weighted and T2-FLAIR. Labels are also provided for the whole tumour, tumour core, and enhancing tumour regions. In order to simplify the experiments, we only segmented the T2-FLAIR sequence, which is commonly used in brain imaging. The BraTS dataset draws a fair amount of attention: for example, Jiang et al.[40] developed a two-stage model using a U-Net[4], utilising all sequences and using a post-processing thresholding technique. If the enhancing tumour region was less than a hand-tuned threshold, then the region was replaced with necrosis, which may cause significant improvement to the results . An average DICE of 0.888 for tumours was reported. Zhao et al.[41] made use of a pipeline involving a CNN and a number of expedients including different methods of sampling, patch-based training and teacher-student models, resulting in a reported mean DICE score of 0.883. Ali et al.[42] exploited multiple CNNs trained separately, with final predictions based on ensembling the probability maps from each CNN, with a mean DICE of 0.906 for the whole tumour. These works developed specific pipelines with a particular focus on brain tumour segmentation, making use of all available sequences. Our proposed pipeline, which was not brain-specific, did not produce quite as good results but was only based on T2 FLAIR sequences. It is possible that its accuracy would have been greater if more sequences had been used. It is also possible that its performance on MR-acquired images and/or brain images may not as good as that on CT-acquired images and/or aortic ones.

Abdominal organs. Gibson et al.[43] proposed a new network architecture for the purpose of multi-organ segmentation in abdominal CT scans. A mean DICE of 0.93, 0.95, 0.95 was reported for the left kidney, the spleen and liver respectively. Another method was developed by Cai et al.[44] who developed a novel shape learning network architecture, building an expected shape for the organ into the model. A mean DICE score of 0.96 and 0.94 was reported for the spleen and liver respectively. Weston et al.[45] implemented a deep learning method aimed at segmenting the complete abdomen and pelvis using a locally collected dataset. A variation on the 3D UNet was implemented, and a mean DICE score of 0.93, 0.93, 0.88, 0.95 was reported for the kidneys(combined), spleen, aorta and liver respectively. These methods were developed specifically for abdominal segmentation. Although our pipeline was not quite as accurate, the quantitative results demonstrate the advantage of the hybrid pipeline over a conventional approach.

A summary of the discussed methods from the literature and the proposed method can be found in Table 7.

The main limitation of our study is the limited testing of the hybrid pipeline. It is possible that its accuracy may not be reproducible when applied to other organs, to scans performed with different settings (for example: inferior resolution), to patients with less well-defined boundary conditions between body structures on CT or MR, or in presence of artefacts. Notably, all the above issues would also interfere with manual segmentation, which is still considered the gold standard. More extensive training/testing of the pipeline would be necessary to clarify its generalisability.

Conclusion

Our fully automatic segmentation pipeline combining elements from both mathematical modelling and artificial intelligence has been shown to be an accurate method for segmenting aortic 3D volumes and performed well also when applied to MR images of brain tumours and abdominal organs.

We plan to test and refine the pipeline further on similar but not identical clinical tasks, such as segmentation of preoperative CT scans of the aorta. This could be achieved by using the current pipeline as a baseline and potentially modifying it to suit preoperative scans. Our segmentation pipeline provides the groundwork for developing a method for serial comparison of images, potentially reducing operator time and bias.

Ultimately, further testing on larger datasets will be necessary before attempting clinical experimentation and translation.

	Region	Data	DICE	JACCARD
Aortic				
Lalys et al.[30]	Whole Aneurysm	92 preop CT scans	0.86 ± 0.06	0.82 ± 0.07
		15 postop CT scans	0.87 ± 0.03	0.83 ± 0.04
Lareyre et al.[32]	Lumen	40 preop CT scans	0.93 ± 0.04	0.87 ± 0.07
	Thrombus		0.88 ± 0.12	0.80 ± 0.15
Lopez et al.[35]	Thrombus	12 preop CT scans	0.84 ± 0.07	0.73 ± 0.10
	Thrombus	16 postop CT scans	0.89 ± 0.04	0.81 ± 0.07
Lu et al.[36]	Whole Aneurysm	57 preop CT scans	0.89 ± 0.05	–
Caradu et al.[38]	Whole Aneurysm	100 preop CT scans	0.95 ± 0.01	0.91 ± 0.02
Adam et al.[39]	Thoracic Aorta, Whole Aneurysm, Iliac Sections	22 postop CT scans	0.93	–
Proposed	Thrombus	10 postop CT scans	0.91 ± 0.05	0.84 ± 0.09
	Stent and Lumen		0.94 ± 0.03	0.88 ± 0.05
	Whole Aneurysm		0.93 ± 0.04	0.88 ± 0.07
BraTS				
Jiang et al.[40]	Whole tumour	125 MRI scans	0.89	–
Zhao et al.[41]			0.88	–
Ali et al.[42]			0.91	–
Proposed		40 MRI scans	0.83 ± 0.13	0.73 ± 0.17
Abdomen				
Gibson et al.[43]	Left Kidney	90 CT scans	0.93	–
	Spleen		0.95	–
	Liver		0.95	–
Cai et al.[44]	Spleen	10 CT scans	0.96 ± 0.01	–
	Liver	32 CT scans	0.94 ± 0.03	–
Weston et al.[45]	Left and right kidneys	18 CT scans	0.93	–
	Spleen		0.93	–
	Aorta		0.88	–
	Liver		0.95	–
Proposed	Left kidney	10 CT scans	0.89 ± 0.08	0.82 ± 0.11
	Right kidney		0.84 ± 0.08	0.73 ± 0.11
	Spleen		0.85 ± 0.22	0.79 ± 0.24
	Aorta		0.85 ± 0.12	0.75 ± 0.16
	Liver		0.94 ± 0.02	0.89 ± 0.05

Table 7. Recently published segmentation experiments on cross-sectional images of solid organs or lesions.

Data availability

The BraTS dataset is publically available from https://www.med.upenn.edu/cbica/brats2020/data.html. The Abdomen datset is publically available from https://www.synapse.org/#Synapse:syn3193805/wiki/217789. The aortic data that support the findings of this study are available from Royal Liverpool and Broadgreen University Hospitals but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Received: 7 March 2022; Accepted: 5 August 2022
Published online: 20 August 2022

References

1. Mumford, D. B. & Shah, J. Optimal approximations by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.* **25**, 25 (1989).
2. Chan, T. F. & Vese, L. A. Active contours without edges. *IEEE Trans. Image Process.* **10**, 266–277 (2001).
3. Cai, X., Chan, R. & Zeng, T. A two-stage image segmentation method using a convex variant of the mumford-shah model and thresholding. *SIAM J. Imag. Sci.* **6**, 368–390 (2013).
4. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241 (Springer, 2015).
5. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
6. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818 (2018).
7. Tang, M., Valipour, S., Zhang, Z., Cobzas, D. & Jagersand, M. A deep level set method for image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 126–134 (Springer, 2017).

8. Chen, X. et al. Learning active contour models for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 11632–11640 (2019).
9. Burrows, L., Chen, K. & Torella, F. On new convolutional neural network based algorithms for selective segmentation of images. In *Annual Conference on Medical Image Understanding and Analysis*, 93–104 (Springer, 2020).
10. Burrows, L., Guo, W., Chen, K. & Torella, F. Reproducible kernel hilbert space based global and local image segmentation. *Inverse Probl. Imaging* **15**, 1 (2021).
11. Shaikh, U. *et al.* Changes in aortic volumes following endovascular sealing of abdominal aortic aneurysms with the nellix endoprosthesis. *J. Endovasc. Ther.* **22**, 881–885 (2015).
12. Yafawi, A. *et al.* Aneurysm growth after endovascular sealing of abdominal aortic aneurysms (evas) with the nellix endoprosthesis. *Eur. J. Vasc. Endovasc. Surg.* **60**, 671–676 (2020).
13. Yafawi, A. *et al.* Stent frame movement following endovascular aneurysm sealing in the abdominal aorta. *J. Endovasc. Ther.* **26**, 54–61 (2019).
14. Yushkevich, P. A. *et al.* User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **31**, 1116–1128 (2006).
15. Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2014).
16. Bakas, S. *et al.* Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* **4**, 1–13 (2017).
17. Bakas, S. *et al.* Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv:1811.02629 (arXiv preprint) (2018).
18. Landman, B. *et al.* Segmentation outside the cranial vault challenge. https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn3193805&type=ENTITY. https://doi.org/10.7303/SYN3193805 (2015).
19. Theljani, A. & Chen, K. An unsupervised deep learning method for diffeomorphic mono-and multi-modal image registration. In *Annual Conference on Medical Image Understanding and Analysis*, 317–326 (Springer, 2019).
20. Bland, J. M. & Altman, D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **327**, 307–310 (1986).
21. Soltaninejad, M. *et al.* Supervised learning based multimodal mri brain tumour segmentation using texture features from supervoxels. *Comput. Methods Programs Biomed.* **157**, 69–84 (2018).
22. Chen, C. *et al.* Deep learning for cardiac image segmentation: A review. *Front. Cardiovasc. Med.* **25**, 25 (2020).
23. Hesamian, M. H., Jia, W., He, X. & Kennedy, P. Deep learning techniques for medical image segmentation: Achievements and challenges. *J. Digit. Imaging* **32**, 582–596 (2019).
24. Wang, G. *et al.* Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans. Med. Imaging* **37**, 1562–1573 (2018).
25. Wang, G. *et al.* Deepigeos: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1559–1572 (2018).
26. Loncaric, S., Subasic, M. & Sorantin, E. 3-d deformable model for aortic aneurysm segmentation from ct images. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 00CH37143)*, vol. 1, 398–401 (IEEE, 2000).
27. Subasic, M., Loncaric, S. & Sorantin, E. Region-based deformable model for aortic wall segmentation. In *3rd International Symposium on Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the*, vol. 2, 731–735 (IEEE, 2003).
28. Subašić, M., Lončarić, S. & Sorantin, E. Model-based quantitative aaa image analysis using a priori knowledge. *Comput. Methods Programs Biomed.* **80**, 103–114 (2005).
29. Zohios, C., Kossioris, G. & Papaharilaou, Y. Geometrical methods for level set based abdominal aortic aneurysm thrombus and outer wall 2d image segmentation. *Comput. Methods Programs Biomed.* **107**, 202–217 (2012).
30. Lalys, F., Yan, V., Kaladji, A., Lucas, A. & Esneault, S. Generic thrombus segmentation from pre-and post-operative cta. *Int. J. Comput. Assist. Radiol. Surg.* **12**, 1501–1510 (2017).
31. Kass, M., Witkin, A. & Terzopoulos, D. Snakes: Active contour models. *Int. J. Comput. Vis.* **1**, 321–331 (1988).
32. Lareyre, F. *et al.* A fully automated pipeline for mining abdominal aortic aneurysm using image segmentation. *Sci. Rep.* **9**, 1–14 (2019).
33. López-Linares, K. *et al.* Fully automatic detection and segmentation of abdominal aortic thrombus in post-operative cta images using deep convolutional neural networks. *Med. Image Anal.* **46**, 202–214 (2018).
34. Xie, S. & Tu, Z. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 1395–1403 (2015).
35. López-Linares, K., García, I., García-Familiar, A., Macía, I. & Ballester, M. A. G. 3d convolutional neural network for abdominal aortic aneurysm segmentation. arXiv:1903.00879 (arXiv preprint) (2019).
36. Lu, J.-T. *et al.* Deepaaa: clinically applicable and generalizable detection of abdominal aortic aneurysm using deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 723–731 (Springer, 2019).
37. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571 (IEEE, 2016).
38. Caradu, C., Spampinato, B., Vrancianu, A. M., Bérard, X. & Ducasse, E. Fully automatic volume segmentation of infrarenal abdominal aortic aneurysm computed tomography images with deep learning approaches versus physician controlled manual segmentation. *J. Vasc. Surg.* **74**, 246–256 (2021).
39. Adam, C. *et al.* Pre-surgical and post-surgical aortic aneurysm maximum diameter measurement: Full automation by artificial intelligence. *Eur. J. Vasc. Endovasc. Surg.* **62**, 869–877 (2021).
40. Jiang, Z., Ding, C., Liu, M. & Tao, D. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *International MICCAI Brainlesion Workshop*, 231–241 (Springer, 2019).
41. Zhao, Y.-X., Zhang, Y.-M. & Liu, C.-L. Bag of tricks for 3d mri brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, 210–220 (Springer, 2019).
42. Ali, M., Gilani, S. O., Waris, A., Zafar, K. & Jamil, M. Brain tumour image segmentation using deep networks. *IEEE Access* **8**, 153589–153598 (2020).
43. Gibson, E. *et al.* Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Trans. Med. Imaging* **37**, 1822–1834 (2018).
44. Cai, J. *et al.* End-to-end adversarial shape learning for abdomen organ deep segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 124–132 (Springer, 2019).
45. Weston, A. D. *et al.* Complete abdomen and pelvis segmentation using u-net variant architecture. *Med. Phys.* **47**, 5609–5618 (2020).

Author contributions

F.T., L.B., W.G. and K.C. contributed to the methodology. L.B. conducted the experiments. M.H. provided the manual segmentation for the aortic data. F.T. analysed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18173-0.

Correspondence and requests for materials should be addressed to L.B. or K.C.