Diagn Interv Radiol 2020; 26:11–21

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors

A. Emre Kavur
Naciye Sinem Gezer
Mustafa Barış
Yusuf Şahin
Savaş Özkan
Bora Baydar
Ulaş Yüksel
Çağlar Kılıkçıer
Şahin Olut
Gözde Bozdağı Akar
Gözde Ünal
Oğuz Dicle
M. Alper Selver

PURPOSE
We aimed to compare the accuracy and repeatability of emerging machine learning-based (i.e., deep learning) automatic segmentation algorithms with those of well-established interactive semi-automatic methods for determining liver volume in living liver transplant donors at computed tomography (CT) imaging.

METHODS
A total of 12 methods (6 semi-automatic, 6 full-automatic) were evaluated. The semi-automatic segmentation algorithms were based on both traditional iterative models including watershed, fast marching, region growing, active contours and modern techniques including robust statistics segmenter and super-pixels. These methods entailed some sort of interaction mechanism such as placing initialization seeds on images or determining a parameter range. The automatic methods were based on deep learning and included three framework templates (DeepMedic, NiftyNet and U-Net), the first two of which were applied with default parameter sets and the last two involved adapted novel model designs. For 20 living donors (8 training and 12 test datasets), a group of imaging scientists and radiologists created ground truths by performing manual segmentations on contrast-enhanced CT images. Each segmentation was evaluated using five metrics (i.e., volume overlap and relative volume errors, average/root-mean-square/maximum symmetrical surface distances). The results were mapped to a scoring system and a final grade was calculated by taking their average. Accuracy and repeatability were evaluated using slice-by-slice comparisons and volumetric analysis. Diversity and complementarity were observed through heatmaps. Majority voting (MV) and simultaneous truth and performance level estimation (STAPLE) algorithms were utilized to obtain the fusion of the individual results.

RESULTS
The top four methods were automatic deep learning models, with scores of 79.63, 79.46, 77.15, and 74.50. Intra-user score was determined as 95.14. Overall, automatic deep learning segmentation outperformed interactive techniques on all metrics. The mean volume of liver of ground truth was 1409.93±271.28 mL, while it was calculated as 1342.21±231.24 mL using automatic and 1201.26±258.13 mL using interactive methods, showing higher accuracy and less variation with automatic methods. The qualitative analysis of segmentation results showed significant diversity and complementarity, enabling the idea of using ensembles to obtain superior results. The fusion score of automatic methods reached 83.87 with MV and 86.20 with STAPLE, which were only slightly less than fusion of all methods (MV, 86.70) and (STAPLE, 88.74).

CONCLUSION
Use of the new deep learning-based automatic segmentation algorithms substantially increases the accuracy and repeatability for segmentation and volumetric measurements of liver. Fusion of automatic methods based on ensemble approaches exhibits best results with almost no additional time cost due to potential parallel execution of multiple models.

From the Graduate School of Natural and Applied Sciences (A.E.K., U.Y.), Dokuz Eylül University, İzmir, Turkey; Departments of Radiology (N.S.G., M.B., O.D.) and Electrical and Electronics Engineering (M.A.S. ✉ *aselver@gmail.com*), Dokuz Eylül University School of Medicine, İzmir, Turkey; Department of Computer Engineering (Y.Ş., Ş.O., G.Ü.), İstanbul Technical University, İstanbul, Turkey; Department of Electrical and Electronics Engineering (S.Ö., B.B., G.B.A.), Middle East Technical University, Ankara, Turkey; Department of Computer Engineering (Ç.K.), Uludağ University, Bursa, Turkey.

Received 21 January 2019; revision requested 05 March 2019; last revision received 02 May 2019; accepted 10 June 2019.

Published online 1 November 2019.

DOI 10.5152/dir.2019.19025

Computer-aided determination of liver volume from healthy subjects is useful for various applications such as investigating factors influencing the liver size, preoperative volumetric assessment of donor liver and printing three-dimensional (3D) models (1–3). Segmentation is the key element of these analyses as it eliminates the information that does not belong to the liver from the images (4). Up until the recent developments in machine learning technology, the semi-automatic methods were accepted as the primary

You may cite this article as: Kavur AE, Gezer NS, Barış M, et al. Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors. Diagn Interv Radiol 2020; 26:11–21.

tools by providing the highest performance measurements and substantially reducing the time needed for segmentation tasks, particularly in the liver (5, 6).

The semi-automatic methods take advantage of various interaction mechanisms such as inserting seed points to initialize algorithms, manually measuring maximum diameters of the liver and distances in midclavicular line or using advanced interfaces providing visual or some other form of feedback for optimal parameter specification (7–10). However, these interactions are user dependent and additional analyses are required to show significant intra- and interobserver agreements. Unfortunately, such analyses are usually performed with a very limited number of operators (i.e., one to three) and cannot reflect a generalization over repeatability and consistency (11). Moreover, interaction procedures and times may become tedious for challenging cases; therefore, several studies aim to automatize the interaction tasks by utilizing additional image processing strategies (12, 13).

On the other hand, the recent developments in emerging deep learning technologies enable construction of systems that are shown to be able to achieve higher accuracy and repeatability in a fully automatic manner (14). Briefly, deep learning is a branch of machine learning that employs multi-layered neural networks having a much more complex architecture and internal feature extraction mechanism. In other words, deep learning has changed traditional feature extraction followed by classification pipeline to a simpler input-output strategy through utilization of deeper layers, which requires much more training data compared with machine learning. Deep learning and related models, especially convolutional neural networks (CNNs), are attracting growing number of researchers in all fields of medical image processing (15, 16). Besides successful applications on classification, detection, and quality control; the most addressed application area is reported to be the segmentation (14, 17–19). In many cases, deep learning models that are prepared for segmentation are shown to accelerate the progress of the ongoing research studies in terms of precision, sensitivity, and processing time (20–22).

In 2007, SLIVER07 challenge provided such a comparative study of a range of algorithms for liver segmentation under several intentionally included difficulties such as patient orientation variations or tumors and lesions (19). Its outcomes reported a snapshot of the methods that were popular for medical image analysis and since then, abdomen-related challenges mostly aim disease and tumor detection rather than organ segmentation. However, healthy liver segmentation has many challenges as well as important application areas. In the last decade, machine learning based automatic strategies, especially deep learning through CNNs, introduced significant novelties and improvements to medical image segmentation. In order to reflect these recent improvements to the field, a new challenge is organized and carried out for liver segmentation from computed tomography (CT) with participation of 14 teams.

In this article, the annotation framework, resulting data set, evaluation setup, details of participating methods together with their results and performance analysis are presented. The accuracy and repeatability of emerging machine learning based (i.e., deep learning) automatic segmentation algorithms are compared with those of well-established interactive semi-automatic methods for determining liver volume in living liver transplant donors at CT imaging.

Main points

- A comparative study was performed for segmentation of liver and volume determination in living liver transplant donors at computed tomography (CT) imaging.
- The use of emerging deep learning frameworks for automatic liver segmentation outperformed the well-established semi-automatic (interactive) methods.
- Segmentation performance can be further increased via ensemble methods and can save a substantial amount of time while improving repeatability.

Methods

Data analysis

This study was approved by the Institutional Review Board and informed consent was obtained from patients who participated in the study. The CT database consists of 20 contrast-enhanced abdominal data sets acquired from different patients using two different CT scanners, either 16-slice (Brilliance 16 Philips; Philips Medical Systems) or 64-slice (Brilliance 64 Philips; Philips Medical Systems) multidetector CT scanner. The pixel size (i.e., x-y spacing attribute of DICOM meta data) of series differs between 0.7 and 0.8 mm, while their slice thickness is 3 to 3.2 mm. Each patient data includes an average of 90 slices (minimum 77, maximum 105 slices) that contains images of a healthy liver. In total, 597 slices (30% of data) were provided for training and 1325 slices (70% of data) were used for tests. All images in a single CT series have similar Hounsfield unit (HU) range of adjacent organs while the same tissue across different data sets have varying HU ranges due to the injection of contrast media. Atypical liver shapes (i.e., unusual size, orientation or atypical contour of the liver) composed 15% of the database.

The CT data were manually segmented by a group of image specialists and radiologists in order to create ground truth masks. The ground truths were further annotated by another expert radiologist and the final masks are generated upon consensus. The training batch including anonymized DICOM images and ground truth masks are distributed to the registered competitors in order to prepare their algorithms before the challenge. Only anonymized DICOM images are included in the test bench and the ground truth maps are never shared with the participants. Instead, the participants submitted their results (i.e., binary image series) and only the evaluation results (i.e., grades) of their algorithms are provided to them.

Image segmentation methods

Among 14 participating teams, 12 segmentation results were submitted: 6 of these results have used automatic approaches while 6 of them utilized interactive methods. The following subsections describe the participating methods. Moreover, a concise comparison of the methods is presented in the Table 1.

Semi-automatic (interactive) image segmentation methods

In this category, six well-established semi-automatic image segmentation methods have participated to the challenge. The parameters of all these methods were adjusted by educated guess from biomedical, electronics and/or computers engineers with radiologic and digital image processing background. Two 2D methods, namely MATLAB based Active Contours (2D-AC) and Super-pixels (2D-SP), and four Slicer software integrated 3D methods, namely Watershed (3D-WS), Fast Marching (3D-FM),

Table 1. Comparison of the methods in terms of their advantages and disadvantages

	Segmentation method	Advantages		Disadvantages
AUTOMATIC (DEEP LEARNING) TECHNIQUES	DEU_DeepMedic	• Easy to adapt the system for different modalities and tissues • Easier installation procedure than other CNN based systems	• Ability to learn highly discriminative features automatically during the training process • Higher performance compared to conventional machine learning strategies	CNN-based automatic segmentation methods have similar disadvantages: • It is hard to determine an adequate architecture for different semantic segmentation problems • Experience is required to find optimal parameters for convergence • They need huge amounts of data for training • They need too much computational power (i.e., powerful graphics card, higher memory) • There are too many parameters needed for optimization
	DEU_Nifty-Net	• Maximum simplicity to understand • Supports multi-modal input		
	ITU_U-Net METU_U-Net X_U-Net Y_U-Net	• Increases resolution of the output images iteratively by fusing discriminative features from hidden layer		
SEMI-AUTOMATIC (INTERACTIVE) METHODS	3D-RG	• Easy to implement • Segments very quickly on smooth tissues		• Needs too many user interactions • Hard to determine proper thresholds • Fails when border gradients are low
	2D-AC	• Robust against small artifacts and noise		• The boundary conditions must be tuned very well to achieve high performance
	3D-FM	• Easy to implement • Segments very quickly on smooth tissues		• Needs too many user interactions • Sometimes iteration overflows from target object at lower gradient borders
	RSS	• Performs well on challenging areas		• Needs too much time for iterations • Might need tedious user interactions
	3D-WS	• Shows higher performance on borders having lower gradient values		• Needs too much computational power • Target organ area must be selected manually at each iteration
	2D-SP	• Works on meaningful regions not on the individual pixels itself		• Might need too many optimization steps for convergence

CNN, convolutional neural network

Region Growing (3D-RG), and Robust Statistics Segmenter (RSS), were utilized (23). All these methods are well-established and their performances as well as pitfalls are known. Thus, they created a great baseline for comparisons with automatic methods.

Automatic approaches

All teams in automatic category have participated using a deep learning-based strategy which clearly reflects the current trend in radiologic image analysis. Despite its outstanding performance, training a deep learning CNN from scratch is difficult due to two requirements: 1) a large amount of labelled data and 2) a significant amount of expertise to ensure convergence. In order to address these, 1) challenges were organized to provide the necessary medical data, and 2) template models were prepared and released by experts. These models use different approaches designed by various architectures for good localization and enriched use of context at the same time. They aim modular structural designs for sharing networks and pre-trained models, using which it is possible to get started directly with established built-in tools, adapt existing networks to the imaging data, and quickly build new solutions to the other particular image analysis problems.

All the proposed methods, which are given in detail below, are based on well-known models, DeepMedic, U-Net, and Nifty-Net (24–26). Unfortunately, setting the parameters of these models or making even slight revisions in the architecture to obtain higher performance are still far away from being trivial. On the other hand, it was recently shown by an extensive study that selection of these parameters can have drastic effects on performance (18).

DEU_DeepMedic: DeepMedic consists of a 3D CNN coupled with a 3D fully connected conditional random field. The generic nature of our system allows its straightforward application for different lesion segmentation tasks without major adaptations. DeepMedic was originally developed for brain and its lesions' segmentation. In the context of this challenge, DeepMedic was adopted to liver segmentation problem. However, it was run with its default parameters as given.

DEU_Nifty-Net: Nifty-Net is another CNN platform designed for medical image analysis researches and has also been used with its default parameters as given (26). Nifty-Net is an open source system and it uses TensorFlow framework.

ITU_U-Net: The model is designed as a variation of U-Net architecture, which is built upon a fully convolutional network extended by large number of features at the up sampling that allows the network to propagate context information better to higher resolution layers. ITU_U-Net architecture starts with a 2D convolution. After that convolution, the result is down-sampled in five down-sampling blocks having

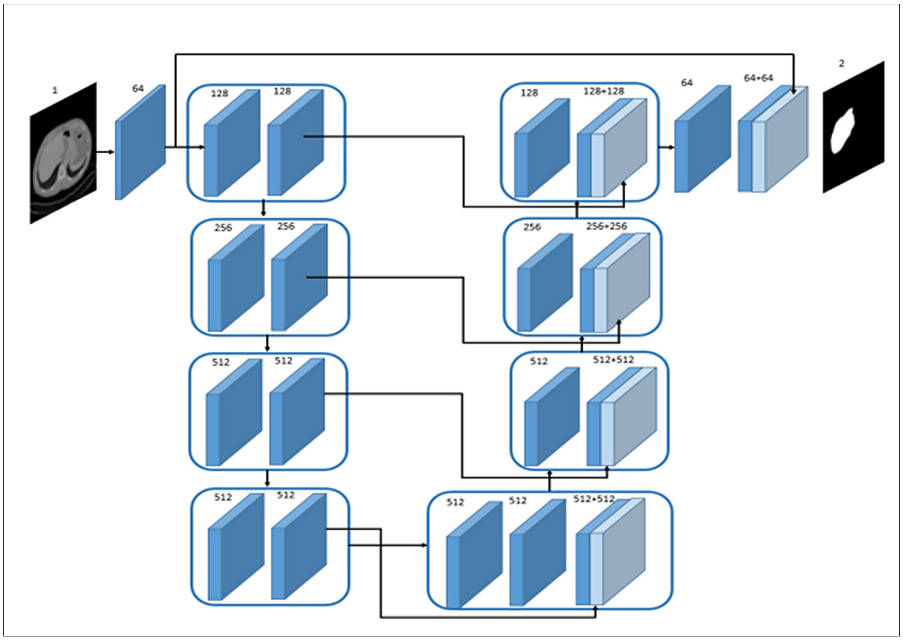

Figure 1. ITU_U-Net architecture and design.

a leaky Rectified Linear Unit (ReLU), convolution and batch normalization (BN), respectively. Then, at the smallest size, convolution transpose is applied before the BN. In the up-sampling layers, the feature maps are again processed through a ReLU, a convolution transpose and BN. Moreover, corresponding output from down-sampling layers are merged to the output. The architecture is given in Fig. 1, where each block represents an output of a parametrized layer (convolution, BN, deconvolution) and numbers represent channel counts. For both convolution and convolution transpose operations, 4 by 4 kernels with 2 strides is used. The model is trained for 300 epochs using SGD optimizer with 0.9 momentum and 0.0002 learning rate. Data sets from SLIVER challenge are added to the training set. Based on the analysis of images belonging to the training set, intensities outside (-1500,1500) HU range are removed, DICOM meta window level and width information and histogram equalization are used to emphasize the intensity range for the liver as pre-processing steps. Histogram equalization and a simple morphologic post-processing are also explored. The latter aimed to eliminate small outliers in the resulting image. However, they were not used in the final version of the method.

METU_U-Net: This model is also designed as a variation of U-Net for taking advantage of skipping connections that provide better error backpropagation and avoiding loss of information at deeper layers (Fig. 2). In addition, a Conditional Adversarial Network (CAN) is introduced in the proposed model. Contrary to ITU_U-Net, BN is performed before convolution (27). In this way, vanishing gradients are prevented and selectivity is increased. Moreover, parametric ReLU is employed to preserve the negative values using a trainable leakage parameter. In order to improve the performance around the edges, a CAN, which generates similar images to the training data set based on the provided conditions, is employed during training (not as a post-process operation). This introduces a new loss of function to the system which regularizes the parameters for sharper edge responses. Although the proposed system is a 2D network, we have utilized 3D information by concatenating the neighbor slices of the target input slice. Only normalization of each CT image is performed for pre-processing and 3D connected component analysis is utilized for post-processing.

X_U-Net and Y_U-Net: The remaining two teams have also used U-Net with different configurations. Their systems had the same U-Net architecture, but their results were different because of the differences between the architecture and the tuning parameters.

Evaluation strategy of segmentation results

Selecting the proper evaluation metric(s) has critical importance for an informative comparison. In the literature, there are many proposed metrics that compare the similarity of two 3D objects. On the other hand, none of them is sufficient to perform a fair evaluation individually (28, 29). In order to overcome this problem, a weighted average of five different performance metrics was determined and considered as the final grade of the segmentation. These five different performance metrics are:

1. Volumetric overlap (VO)
2. Relative volume difference (RVD)
3. Average symmetric surface distance (ASSD)
4. Root mean square symmetric surface distance (RMSD)
5. Maximum symmetric surface distance (MSSD)

All outputs of these metrics were converted to scores with help of some expert selected thresholds.

Volumetric overlap (VO)

Volumetric overlap is the number of voxels in the intersection of segmentation and reference, divided by the number of voxels in the union of segmented volume and reference volume.

$$VO = \frac{V_{seg} \cap V_{ref}}{V_{seg} \cup V_{ref}} \; x\ 100$$

Here $V_{seg} \cap V_{ref}$ and $V_{seg} \cup V_{ref}$ symbolise number of voxels in the intersection and union of the segmented and reference (ground truth) object. Its value is equal to 100 for a perfect segmentation and 0 as the lowest possible value when there is no overlap at all between segmentation and reference. The threshold is determined as 50%. If a VO of result has lower than 50%, the grade will be 0. If it is higher than 50%, the grade will remain as calculated. The grade conversion is shown in Fig. 3a.

Relative volume difference (RVD)

Relative volume difference (RVD) is the total volume difference between the segmentation and reference which is divided by the total volume of the reference object. The absolute value is taken, and the result is multiplied by 100.

$$RVD = \left| \frac{V_{seg} - V_{ref}}{V_{ref}} \right| x100$$

This value is 0 for a perfect segmentation and larger than zero otherwise. Note that the 0% can also be obtained for a segmentation, which is not identical with reference but has same number of voxels. That is why only RVD is not sufficient for a fair evaluation and many different error calculation metrics were preferred. In evaluation of the results, the RVD values higher than 10 get a grade of 0. The RVD values between 10 and 0 are mapped between 50 and 100 as shown at Fig. 3a. Since the mapping calculation from actual value to percent value has an inverse proportion, lower RVD represents higher performance.

Average symmetric surface distance (ASSD)

Symmetric surface distance metrics provides an alternative comparison method with a different approach. Let the distance of a voxel from a set of voxels belonging it be defined as:

$$d(x,A) = \min_{y \in A}\big(d(x,y)\big)$$

where $d\ (x,\ y)$ is the Euclidean distance vector between the voxels incorporating the real spatial resolution of the image. To calculate symmetric surface distances, the border voxels of segmented and reference objects are determined. For each voxel at the border of first object, the closest border voxel in the second object is calculated. All these distances are stored for all border voxels from both reference and segmentation. The process is illustrated in Fig. 3e.

In this work the distances were chosen at pixel level. The average of all the symmetric surface distances gives the average symmetric surface distance:

$$ASSD = \frac{1}{|V_{seg}| + |V_{ref}|} x \left(\sum\nolimits_{x \in V_{seg}} d(x, V_{ref}) + \sum\nolimits_{y \in V_{ref}} d(y, V_{seg})\right)$$

This value is 0 for a perfect segmentation. There is no upper limit. In our evaluation, the ASSD values greater than 10 get a grade of 0. The values between 10 and 0 are mapped between 50 and 100. There is inverse proportion between ASSD and grades (Fig. 3b).

Root-mean-square symmetric surface distance (RMSSSD)

This metric is like ASSD, but it calculates the squared distances between the two sets of border voxels. After averaging the squared values, the root is extracted and gives the symmetric RMS surface distance.

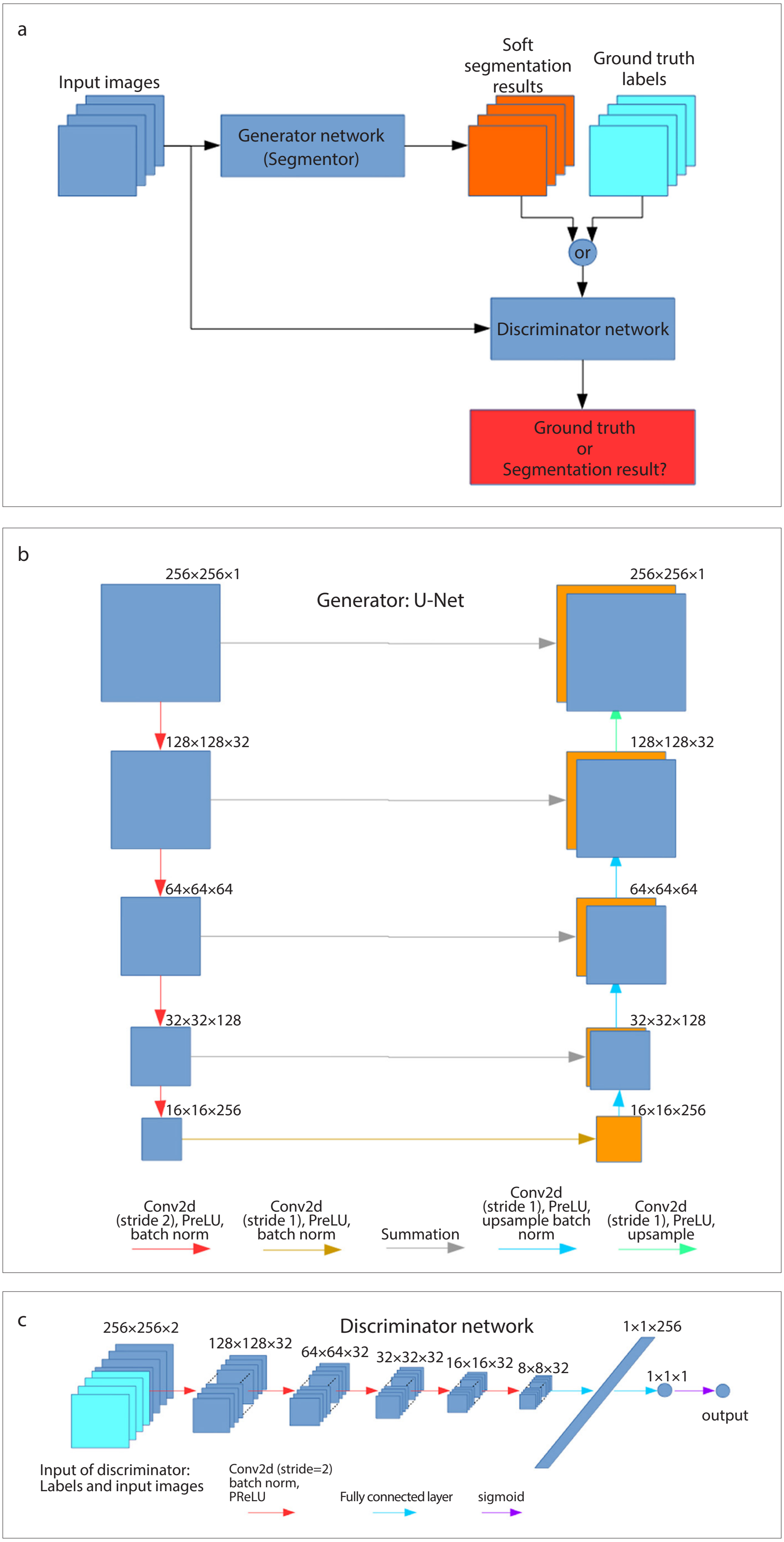
a
Input images
Generator network (Segmentor)
Soft segmentation results
Ground truth labels
or
Discriminator network
Ground truth or Segmentation result?
b
Generator: U-Net
256×256×1
256×256×1
128×128×32
128×128×32
64×64×64
64×64×64
32×32×128
32×32×128
16×16×256
16×16×256
Conv2d (stride 2), PreLU, batch norm
Conv2d (stride 1), PreLU, batch norm
Summation
Conv2d (stride 1), PreLU, upsample batch norm
Conv2d (stride 1), PreLU, upsample
c
Discriminator network
256×256×2
128×128×32
64×64×32
32×32×32
16×16×32
8×8×32
1×1×256
1×1×1
output
Input of discriminator: Labels and input images
Conv2d (stride=2) batch norm, PReLU
Fully connected layer
sigmoid

Figure 2. a–c. METU adversarial network: **(a)**, architecture; **(b)**, generator; **(c)**, discriminator network.

$$RMSSSD = \sqrt{\frac{1}{|V_{seg}| + |V_{ref}|}} x$$

$$\sqrt{\sum_{x \in V_{seg}} d^2(x, V_{ref}) + \sum_{y \in V_{ref}} d^2(y, V_{seg})}$$

This value is 0 for a perfect segmentation. There is no upper limit. In our evaluation, the RMSSSD values greater than 15 get a grade of 0. The values between 15 and 0 are mapped between 50 and 100. Again, there is inverse proportion between RMSSSD and grades (Fig. 3c).

Maximum symmetric surface distance (MSSD)

This metric is like the previous two, but only the maximum of all voxel distances is taken into account instead of the average.

$$MSSD = \max_{x \in V_{seg}} \left(\min_{y \in V_{ref}} (d(x, y)) \right)$$

Maximum symmetric surface distance is one of the most critical error metrics because it represents the maximum allowed error margin in surgical operations. Its value is 0 for a perfect segmentation. There is no upper limit. In our evaluation, the MSSD values greater than 50 get a grade of 0. The values between 50 and 0 are mapped between 50 and 100 (Fig. 3d).

In the results section, the participating methods are compared using the grades calculated by the abovementioned metrics. Scores over all data sets are given in Fig. 4 and average values are given in Table 2. Moreover, the volumes calculated with algorithms are compared to the ground truth (Fig. 5). Then, the complementarity and diversity of the automatic and interactive methods are quantitatively analyzed by heatmap illustrations (Fig. 6). Based on the findings, two ensemble methods, namely majority voting (MV) and simultaneous truth and performance level estimation (STAPLE) are employed to fuse the results of automatic methods and the results are compared to individual method performances in Table 2.

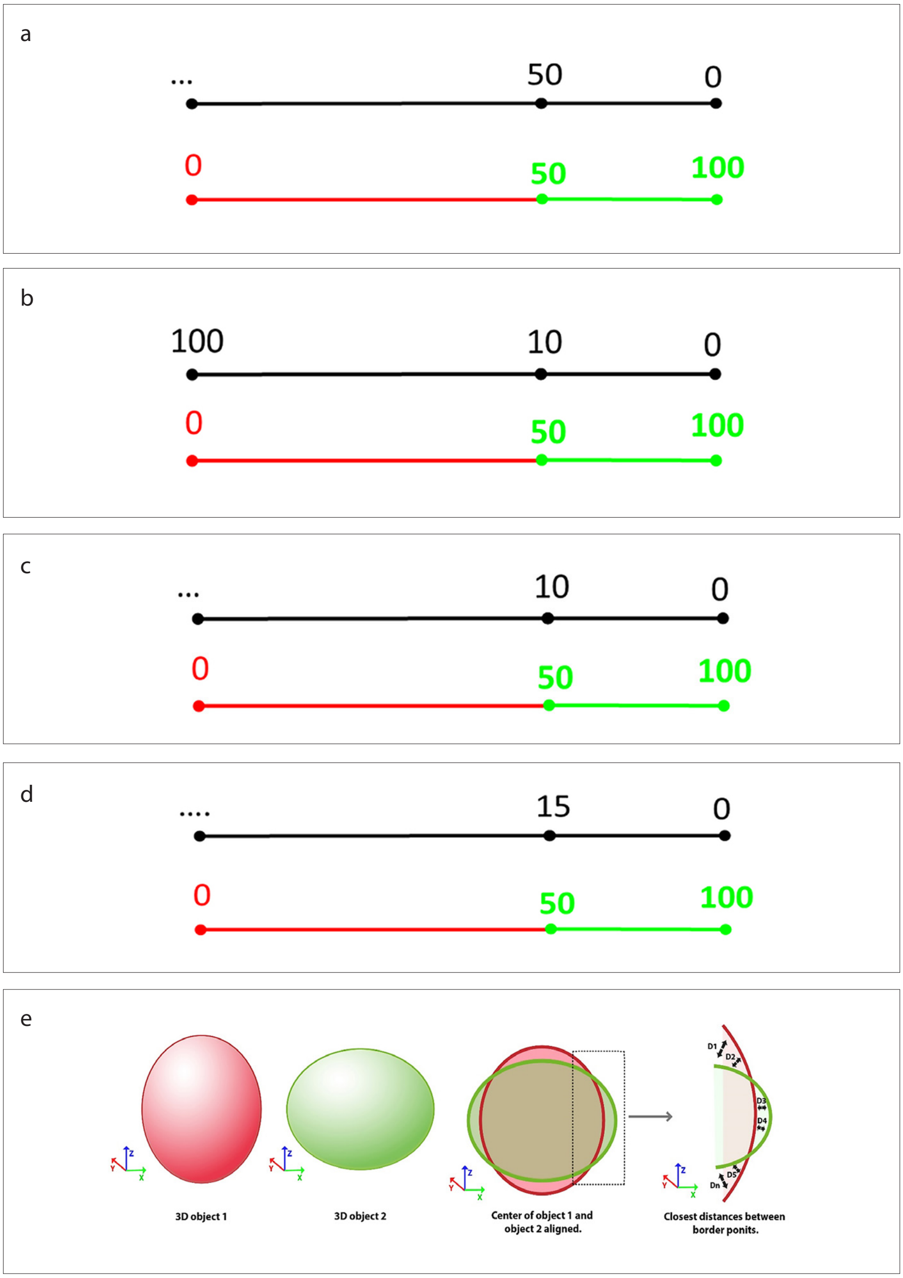

Figure 3. a–e. Illustration of computing grades from metrics of: **(a)**, volumetric overlap and RVD; **(b)**, ASSD; **(c)**, RMSD; **(d)**, MSSD; and **(e)**, symmetric surface distances. Green part of the score line represents the range above the given threshold, while the red part represents the results have zero score (i.e., below the threshold).

Results

The results were evaluated quantitatively by the grade values described in previous section and qualitatively by visual illustrations. The mean values of the grades of all the segmentation methods on all test data (14 patients) are illustrated in Fig. 4 and the numerical details of these results are presented in Table 2. Automatic and interactive methods are indicated with different colors to analyze their results in detail. Moreover, the results of automatic methods, 3D and 2D interactive methods are given in Fig. 4b–4d. It can be clearly observed from the top rows of the Table 2 and Fig. 4b that automatic methods using deep learning-based solution perform better segmentation than the interactive methods. An important drawback of these methods is observed as in some datasets, they might completely fail and generate very low scores. The first four algorithms on the chart are deep learning-based automatic methods and the first two are the proposed novel U-Net designs, which are followed by well-established interactive strategies. These results point out the enhanced performance of automatic methods due to the recent developments in deep learning technology. Fig. 4c shows that the scores of 3D interactive methods

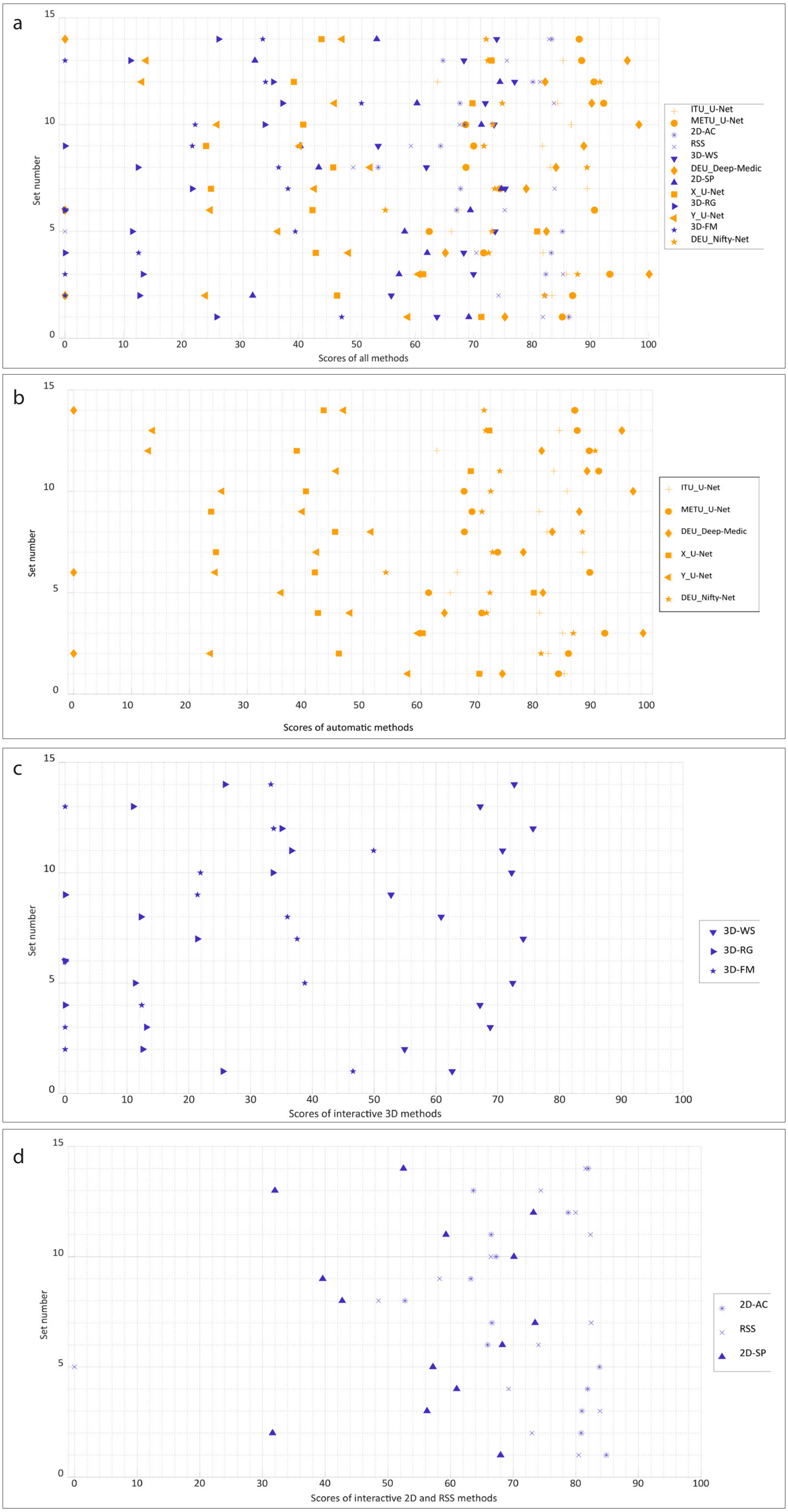

Figure 4. a–d. Scores of: **(a)**, all segmentation methods; **(b)**, automatic methods; **(c)**, interactive 3D methods; **(d)**, interactive 2D and RSS methods on test dataset. Each method has a unique marker. Automatic methods are represented by orange color, while interactive methods are blue.

are between 0 and 75 points, which are less than the deep-learning-based automatic methods. It can be observed from Fig. 4d that RSS and 2D interactive methods have better segmentation performance compared with the 3D interactive methods (maximum 85 points). However, they still left behind the performance of the automatic methods. Considering the results of the previous challenges on liver segmentation, performance differences between the two approaches were always in favor of interactive methods, but this has been changed by the deep learning models.

In addition to these scores, measured volumes of the livers were also analyzed for automatic and interactive methods in Fig. 5a, 5b. The mean volume of liver from ground truth was found to be 1409.93±271.28 mL. The mean volume from automatic methods was calculated as 1342.21±231.24 mL while it was 1201.26±258.13 mL for interactive methods showing higher accuracy and less variation on behalf of automatic methods. This information and regression analysis in Fig. 5a also support the fact that emerging deep learning methods are more reliable and stable than interactive methods for liver segmentation. In Fig. 5b, Bland-Altman plot of the automatic and interactive methods shows the agreement between the utilized methods.

The diversity and complementarity of the segmentation results were analyzed for qualitative evaluation of the outcomes. In order to do that, the binary results of the segmentation algorithms were summed cumulatively, and the values were mapped to virtual color scheme to obtain heatmaps. These maps are generated separately for 6 interactive and 6 automatic methods in order to observe their characteristic differences. The heatmaps are visualized according to two different color spectrums. The first one is generated to examine the true positive (TP) performance of the segmentation algorithms. A TP pixel that is found by all algorithms, would get a value of 6 and be represented by blue. A TP pixel that is found only by a single algorithm, would get a value of 1 and be represented by green. All TP values in between correspond to intermediate number of findings and assigned to a color inside the spectrum (shown on the right side of the Fig. 6). A TP pixel that could not be found by any algorithm is represented by purple to create contrast and draw attention. The second color spectrum aims

Table 2. Mean and standard deviation of all methods' results on test dataset (14 patients)

	VO		RVD		ASSD		RSMSSD		MSSD		Final
Team	%	Grade	%	Grade	mm	Grade	mm	Grade	mm	Grade	score
ITU_U-Net (24)	87.9±2.1	87.9±2.1	4.3±1.8	73.0±4.4	1.7±0.3	66.4±5.4	2.8±0.6	65.4±3.8	26.6±4.7	45.4±14.4	79.6
METU_U-Net (26)	90.4±1.3	90.4±1.3	2.0±1.1	90.0±4.8	1.5±0.3	67.5±1.9	3.1±1.0	64.6±4.8	35.5±10.6	31.2±14.5	79.5
DEU_DeepMedic (23)	85.4±8.3	85.4±8.3	4.4±3.0	72.9±6.7	1.1±2.5	88.7±3.9	1.5±6.1	85.8±4.9	19.9±16.8	53.0±14.6	77.2
DEU_Nifty-Net (25)	78.5±2.9	78.5±2.9	18.7±4.2	42.1±8.2	0.4±1.3	94.5±3.6	1.1±2.7	88.7±4.0	10.3±12.1	91.9±17.3	74.5
2D-AC (27)	88.1±1.3	88.1±1.3	9.7±1.3	31.3±4.3	1.5±0.8	67.2±1.6	2.5±0.3	66.6±1.7	22.6±3.8	50.3±13.2	72.8
RSS (27)	82.5±11.9	82.5±11.9	6.0±2.4	55.2±7.3	1.9±0.4	58.4±3.3	3.8±0.7	55.2 ±2.3	27.6±4.9	35.7 ±12.4	68.2
3D-WS (27)	80.3±11.6	80.3±11.6	6.1±1.5	59.9±12.3	3.8±0.7	49.6±2.0	6.6±1.1	46.5±2.3	29.8±5.1	31.9±13.6	62.3
2D-SP	74.1±2.8	74.1±5.5	7.0±2.7	56.4±8.2	4.6±0.7	48.4 ±1.6	7.7±1.3	46.9±2.1	44.6±6.4	24.0 ±16.8	56.1
X_U-Net (24)	70.7±2.6	70.7±4.7	10.8±2.7	26.6 ±8.1	4.4±0.7	52.7±1.4	7.4±2.1	45.5±3.8	55.4±18.0	18.9±17.9	49.7
3D-RG (27)	64.8±10.4	64.8±11.3	14.4±3.3	7.8±8.5	4.3±1.1	48.5±4.6	6.8±2.5	46.7±5.4	87.6±46.1	18.9±39.4	45.3
Y_U-Net (24)	79.5±4.9	79.5±8.4	12.1±4.2	36.1±4.3	5.4±1.6	39.6±6.4	14.8±3.7	13.3±6.2	110.3±16.1	0.0±0.0	37.6
3D-FM (27)	46.8±15.8	46.8±13.3	19.6±7.4	0.0±0.0	5.3±1.8	16.2±7.9	7.0±2.4	28.1±4.5	31.4±12.3	11.5±17.6	23.7

The results include both real outputs of metrics and their calculated grades with ±95% confidence intervals.

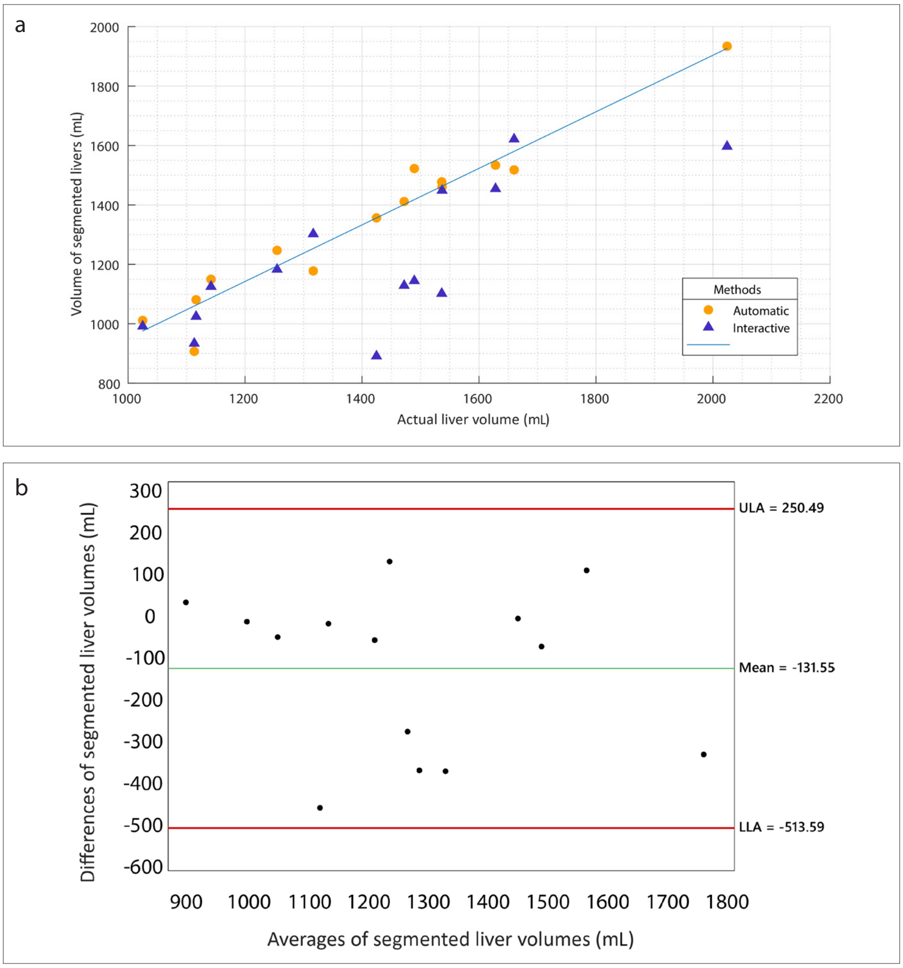

Figure 5. a, b. (a), Regression analysis and **(b)**, Bland Altman plot of liver volumes for both automatic and interactive methods.

to highlight the false positives (FPs). An FP pixel that is incorrectly found by a single algorithm, would get a value of 1 and be represented by orange color. An FP pixel that is incorrectly found by all algorithms, would get a value of 6 and be represented by red. Similar to TP case, the values in between are represented by corresponding colors in the spectrum.

According to Fig. 6, the inhomogeneous characteristics of the contrast-enhanced liver parenchyma and difficulties associated with its segmentation cause results to have varying characteristics. For example, a segmentation algorithm that is sensitive to intensity changes of voxels would probably work fine for the organ borders, but would miss the veins inside of the liver at the same time. It can be observed that there are significant differences between the heatmaps of interactive and automatic methods, which can be listed as follows:

1. Considering agreement rates on TPs, the interactive methods tend to make regional mistakes due to the spatial enlargement-based characteristics. They all seem to have almost no problem when the border of the liver is evident due to the attenuation difference of the contiguous anatomic structures such as fat tissue and gall bladder. However, it is harder to differentiate the outline when the liver is adjacent to isodense structures such as

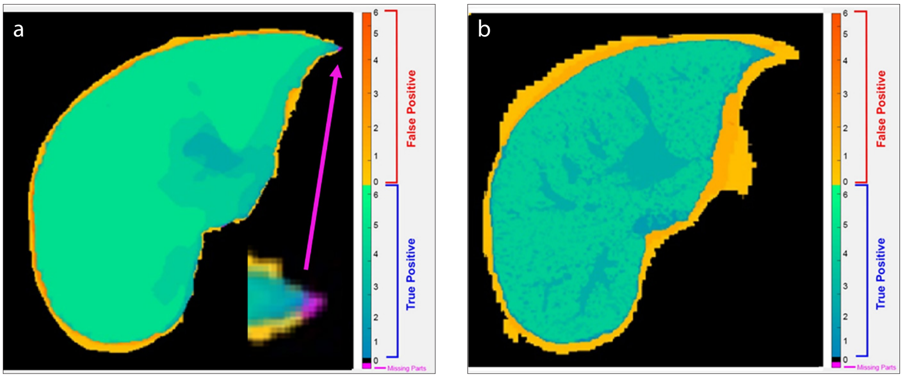

Figure 6. a, b. Colored heatmap example of **(a)**, interactive and **(b)**, automatic segmentation algorithms.

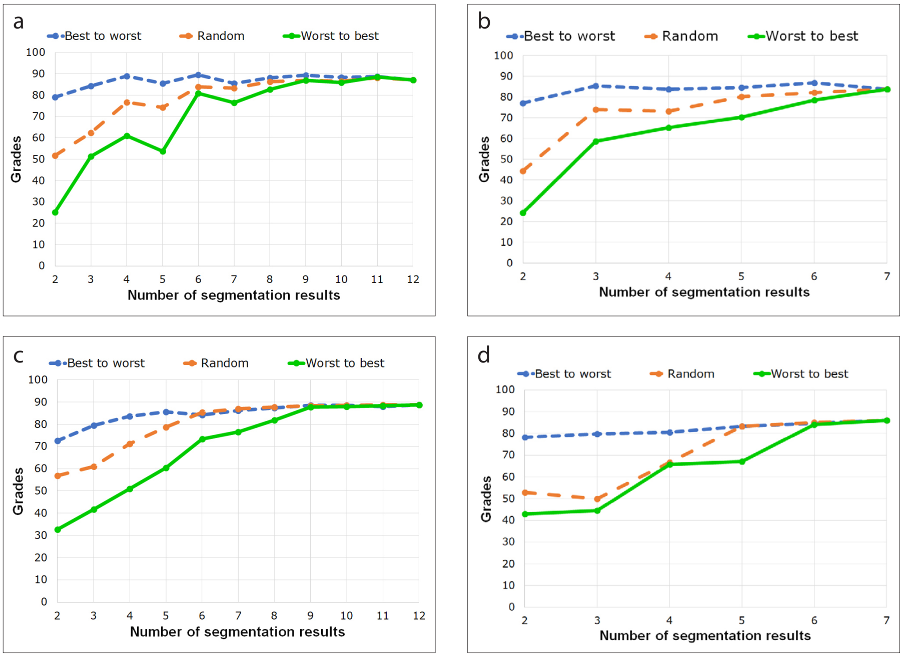

Figure 7. a–d. Performance of **(a)**, majority voting (MV) with all methods; **(b)**, MV with automatic methods; **(c)**, STAPLE with all methods; **(d)**, STAPLE with automatic methods with respect to number and quality of segmentation results.

gastric wall, diaphragm, and isodensely enhanced vena cava inferior. On the other hand, the automatic methods show much more distributed mistakes all over the liver region due to their classification-based characteristics.

2. Although there are almost no parts of the liver which cannot be detected by any of the employed six automatic methods (the percentage of false negative (FN) voxels is under 0.3%), there are small regions that cannot be detected by interactive ones (an example is shown in Fig. 6). This result is particularly important because FNs are much more unlikely to be recovered by post-processing operations after segmentation, while FPs can be reduced significantly.

3. Considering agreement rates on FPs (reddish color map, which was chosen to explore outcomes of the methods violating the ground truth) the interactive methods seem to make much less over segmentations. This is partly related to iterative parameter adjustment of the operator which prevents unexpected results. On the other hand, the FPs of automatic methods are distributed over a larger area. As expected, the agreements on FPs reduces as the distance from the liver region increases.

In the latest medical image processing applications and challenges, it is observed that the fusion of the outcomes of different methods through ensemble systems outperformed the utilization of each component method individually. Moreover, one of the main advantages of the automatic methods is that they can be run in parallel as they do not need any user input or interaction. Thus, it is possible to utilize all the six automatic methods at the same time and fuse their results in order to obtain a superior performance almost without any additional time. To analyze this possibility, two well-known fusion methods, MV and STAPLE, were adapted to segmentation results and the contributions of using ensembles were observed (30, 31).

MV is one of the most straightforward methods to achieve a segmentation fusion. It simply considers segmentation results of all algorithms pixel-by-pixel. If a pixel segmented as "true" by at least n/2 of n algorithms, the value of this pixel is determined as "true" where n is total number of segmentation results. Since MV is known to be sensitive to the number of segmentation results, the effect of n is analyzed by increasing its value (i.e., the number of contributing methods) one by one using three selection strategies *i*) random, *ii*) best-to-worst, and *iii*) worst-to-best (30).

For instance, considering random selection, MV starts by two randomly selected segmentation results and application of MV to whole data sets. At each iteration, another randomly selected segmentation result is added to MV system. The process is finalized after using all methods. In the second case, MV starts by using the two most successful results on Table 2. In the next iteration, the third most successful result is added, and this procedure is repeated until all results are used. The third case is just the opposite order of the second case. The results presented in Fig. 7a show that MV algorithm has a potential for increasing performance of fusion with respect to single segmentation results. The most successful segmentation had a score of 79.67, while MV reached the score of 86.70 when all methods were employed. Considering only the automatic methods, MV got 83.87, which is only 2.83 lower than using all methods (Fig. 7b).

The second fusion method, namely STAPLE algorithm (31), depends on expectation-maximization approach and it significantly differs from MV. The performances of the pre-segmentations are estimated based on comparison to an evolving estimate of the reference standard segmentation at each iteration of the expectation-maximization. The new performance parameters are used to update the reference standard segmentation. STAPLE estimates performance of the templates directly by estimating the ground truth. However, there is no direct association between the intensity similarity of the template and the target image, and the performance of the templates in the locally weighted fusion algorithms. In other words, STAPLE uses only segmentation results. There is no input of ground truth or the original DICOM images. Similar with MV trials, the STAPLE is also applied in three different ways (i.e., random, best-to-worst and worst-to-best).

According to Fig. 7c, STAPLE shows different performance curves under different conditions. It reaches maximum score at fusion of six segmentation results if the results are ordered best-to-worst. The success of STAPLE shows monotonically increasing curve when the segmentation results are ordered from worst-to-best. In this order, STAPLE needs at least two thirds of all segmentation results. On the other hand, the randomly sequenced results have a performance between these two cases as expected.

The final score of STAPLE reaches its maximum in any scenario and outperforms all individual segmentation methods and also MV with a score of 88.74. Considering only the automatic methods, STAPLE got 86.02, which is only 2.72 lower than using all methods (Fig. 7d). It can be easily said that STAPLE boosts the performance of final segmentation and it is a preferable fusion method. It has a unique approach for an ensemble problem. However, it needs too much time to converge a final estimation because of the requirements of many iterations.

Discussion

This study has evaluated the performance of the state-of-the-art deep learning models on automatic segmentation of the liver data. As such, it provided hints regarding the accuracy and robustness of various modern techniques that can be used to direct future research. It was also clearly shown that the participating teams at the automatic segmentation category outperformed the well-established semi-automatic approaches. Thus, the automatic methods have reached the high reliability level of the best semi-automatic methods and now provide a better alternative as they are much faster, operator independent, and executable in parallel.

Also, the challenge served as a great tool by offering expert annotations, retrospective analysis and descriptions of challenging cases and dense sampling of sparse conditions (e.g., atypical liver shapes). Moreover, it allowed application of different approaches to a common dataset and evaluation of the results of each method using the same metrics. This not only provided a comparative analysis to determine the state of the art, but also helped determine diversity and complementarity of different approaches. In the context of the outcomes of the competition, the most successful four automatic approaches are analyzed in detail. The models that took the third and fourth places used "DeepMedic" and "NiftyNet" frameworks, respectively, with slight parameter changes for adaptation to liver segmentation problem and can be considered to produce baseline results for deep learning approaches. The other two were novel deep learning models, ITU-U and METU-U, providing genuine modifications on another framework, namely U-Net. These two models are introduced for the first time in this study.

One of the most important advantages of the automatic approaches is that they can be utilized in parallel and their outcomes can be combined to obtain a superior performance through ensemble systems. The EMMA model, which won the BRATS challenge in 2018, is a great example of such fusion systems (32). Thus, in this study, the outcomes of the participating algorithms were also used as the inputs of two well-established fusion approaches: 1) MV and 2) STAPLE. The results of both algorithms have outperformed all component methods. Performance of MV is known to vary depending on number and order of successful results. As such, MV boosts the performance of segmentation if the results of many methods are included. On the other hand, it is also possible to obtain a superior result with MV if there are only just a few outstanding results. These outcomes are observed to be similar with another MV trial in the literature (33). STAPLE has provided the highest scores by performing slightly better than MV, but it required a significant time to converge. According to our analyses, the advantage of the MV is its speed and high reduction of false positive voxels outside of the liver. Our observations show that the results of MV have diverse characteristics outside of the liver and therefore, it is easy to eliminate false positives. On the other hand, STAPLE can estimate the true outcome better due to its inner machine learning based strategy allowing higher accuracy levels. Thus, it can be confidently concluded that the ensemble systems allow further improvements on automatic methods with almost no time loss if the algorithms can be executed in parallel.

This study has several limitations. First, slice thickness can be considered as a limiting factor on segmentation performance. Having 3 mm thickness gaps, the liver parts having oblique interfaces relative to the imaging plane with the surrounding organs (especially the heart and the stomach) are observed to be affected most from mis-segmentations, since they are much more sensitive to partial volume effects. On the other hand, borders that are more perpendicular to the imaging plane (such as caudate lobe) are less sensitive and segmented with better performance. These observations also agree with (29) and (33). Second limitation is the number of data sets and modalities included in the study (i.e., 20 donors, 8 for training, 12 for testing, acquired by 2 modalities). A larger number of donors would allow a better training especially for deep learning models, which have used additional techniques (such as data augmentation or mirroring) or other similar databases (29) to compensate this limitation. Increasing the number of modalities (probably within the framework of a multicenter study) would allow more diversity of the data sets, which is known to have positive effects on the performance of machine learning based techniques. Third, the evaluation and grading strategy also has some limitations. In this study, the volumetric measurements assume a unit density for the liver and the effects of blood vessels in the segmented liver, partial volume effects and slice thickness have been ignored. An interobserver analysis was performed to observe maximum possible score as well as variability and repeatability. The two manual segmentations performed by the same expert on the same data set at different times resulted in liver volumes of 1491 mL and 1496 mL. The volumetric overlap is found to be 97.21%, while RVD is 0.347%, ASSD is 0.611 (0.263 mm), RMSD is 1.04 (0.449 mm), and MSSD is 13.038 (5.632 mm). These measurements yielded a total grade of 95.14, which is higher than all algorithms but not close to perfect. Especially the MSSD, which has by far the worst score (i.e., 86) among the five measurement parameters, is significantly sensitive to image characteristics. Thus, this variability should be considered when evaluating the performance of the algorithms.

In conclusion, in this validation study performed with living donors for liver transplantation, the use of emerging deep learning frameworks for automatic liver segmentation outperformed the well-established semi-automatic (interactive) methods according to five different metrics. Moreover, deep learning models can work in parallel, if necessary computational power is available. Deep learning methods can further increase the segmentation performance via ensemble methods and can save a substantial amount of time while improving repeatability. Promising future research studies would include performing similar comparative analysis for segmentation of liver from magnetic resonance imaging (34) and extending CT segmentation to vascular analysis (35).

Financial disclosure

This study is supported by TUBITAK under grant number 116E133.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Kromrey ML, Ittermann T, Plodeck V, et al. Reference values of liver volume in Caucasian population and factors influencing liver size. Eur J Radiol 2018; 106:32–37. [CrossRef]
2. Lodewick TM, Arnoldussen CW, Lahaye MJ, et al. Fast and accurate liver volumetry prior to hepatectomy. HPB (Oxford) 2016; 18:764–772. [CrossRef]
3. Witowski JS, Pędziwiatr M, Major P, Budzyński A. Cost-effective, personalized, 3D-printed liver model for preoperative planning before laparoscopic liver hemihepatectomy for colorectal cancer metastases. Int J Comput Assist Radiol Surg 2017; 12:2047–2054. [CrossRef]
4. Garvey B, Türkbey B, Truong H, Bernardo M, Periaswamy S, Choyke PL. Clinical value of prostate segmentation and volume determination on MRI in benign prostatic hyperplasia. Diagn Interv Radiol 2014; 20:229. [CrossRef]
5. Molinari F, Pirronti T, Sverzellati N, et al. Intra- and interoperator variability of lobar pulmonary volumes and emphysema scores in patients with chronic obstructive pulmonary disease and emphysema: comparison of manual and semi-automated segmentation techniques. Diagn Interv Radiol 2013; 19:279. [CrossRef]
6. Hermoye L, Laamari-Azjal I, Cao Z, et al. Liver segmentation in living liver transplant donors: comparison of semiautomatic and manual methods. Radiology 2015; 234:171–178. [CrossRef]
7. Fischer F, Selver MA, Hillen W, Guzelis C. Integrating segmentation methods from different tools into a visualization program using an object-based plug-in interface. IEEE Trans Inf Technol Biomed 2010; 14:923–934. [CrossRef]
8. Roloff, AM, Heiss P, Schneider TP, et al. Accuracy of simple approaches to assessing liver volume in radiological imaging. Abdom Radiol 2016; 41:1293–1299. [CrossRef]
9. Goja S, Yadav SK, Yadav A, et al. Accuracy of preoperative CT liver volumetry in living donor hepatectomy and its clinical implications. Hepatobiliary Surg Nutr 2018; 7:167. [CrossRef]
10. Selver MA, Exploring brushlet based 3D textures in transfer function specification for direct volume rendering of abdominal organs. IEEE Trans Vis Comput Graph 2015; 21:174–187. [CrossRef]
11. Yoon JH, Lee JM, Jun JH, et al. Feasibility of three-dimensional virtual surgical planning in living liver donors. Abdom Imaging 2015; 40:510–520. [CrossRef]
12. Huynh HT, Karademir I, Oto A, Suzuki, K. Computerized liver volumetry on MRI by using 3D geodesic active contour segmentation. AJR Am J Roentgenol 2014; 202:152–159. [CrossRef]
13. Selver MA, Segmentation of abdominal organs from CT using a multi-level, hierarchical neural network strategy. Comput Methods Programs Biomed 2014; 113:830–852. [CrossRef]
14. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Analysis 2017; 42:60–88. [CrossRef]
15. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging 2016; 35:1299–1312. [CrossRef]
16. Hoo-Chang S, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016; 35:1285. [CrossRef]
17. Payan A, Montana G. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv 2015; 1502.02506.
18. de Vos BD, Wolterink JM, de Jong PA, Viergever MA, Išgum I. 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. Proc SPIE Medical Imaging 2016; 97841Y. [CrossRef]
19. Tarroni G, Oktay O, Bai W, et al. Learning-Based Quality Control for Cardiac MR Images. IEEE Trans Med Imaging 2018; 37:11.
20. Xue Y, Xu T, Zhang H, et al. SegAN: Adversarial network with multi-scale L_1 loss for medical image segmentation. Neuroinformatics 2018; 1–10. [CrossRef]
21. Chen L, Bentley P, Mori K, et al. DRINet for medical image segmentation. IEEE Trans Med Imaging 2018; 37:2453–2462. [CrossRef]
22. Oktay O, Ferrante E, Kamnitsas K, et al. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans Med Imaging 2018; 37:384–395. [CrossRef]
23. Pieper S, Halle, M, Kikinis R. 3D Slicer. In Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging: Nano to Macro, Arlington, VA, USA. ISBI 2004; 632–635.
24. Kamnitsas K, Ledig C, Newcombe VF, et al. Efficient multiscale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Analysis 2017; 36:61–78. [CrossRef]
25. 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. MICCAI 2015; 234–241. [CrossRef]
26. Gibson E, Li W, Sudre C, et al. NiftyNet: A deep learning platform for medical imaging. Comput Methods Programs Biomed 2018; 158:113–122. [CrossRef]
27. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning. ICML 2015; 37:448–456.
28. Heimann TB, van Ginneken, Styner MA. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging 2009; 28:1251–1265. [CrossRef]
29. Maier-Hein L, Eisenmann M, Reinke A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care? Nat Commun 2018; 9:5217. [CrossRef]
30. Artaechevarria X, Muñoz-Barrutia A, Solórzano CO. Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Trans Med Imaging 2009; 28:1266–1277. [CrossRef]
31. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004; 23:903–921. [CrossRef]
32. Kamnitsas K, Bai W, Ferrante E, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. arXiv 2017; 1711.01468.
33. Sabuncu MR, Thomas BT, Yeo K, et al. A generative model for image segmentation based on label fusion. IEEE Trans Med Imaging 2010; 29:1714–1729. [CrossRef]
34. Selvi E, Selver MA, Kavur AE, et al. Segmentation of abdominal organs from MR images using multi-level hierarchical classification. J Fac Archit Eng Gaz 2015; 30:533–546.
35. Selver MA, Kavur AE. Implementation and use of 3D pairwise geodesic distance fields for seeding abdominal aortic vessels. Int J Comput Assist Radiol Surg 2016; 11:803–816. [CrossRef]