Article Type: Research Article

Title: Fully automatic, multi-organ segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFs), convolutional neural networks (CNNs) and a multi-atlas (MA) approach.

Authors: Ioannis Lavdas PhD [1], Ben Glocker PhD [2], Konstantinos Kamnitsas PhD [2], Daniel Rueckert PhD [2], Henrietta Mair MBBS, MA [3], Amandeep Sandhu MEng, MBBS [4], Stuart A. Taylor MRCP, FRCR [3] ,Eric O. Aboagye PhD [1], Andrea G. Rockall MRCP, FRCR [5]

[1] Imperial College Comprehensive Cancer Imaging Centre (C.C.I.C.), Hammersmith Campus, Du Cane Road, W12 0NN, Commonwealth Building Main Office, Ground Floor, [2] Biomedical Image Analysis Group, Department of Computing, Huxley Building, 180 Queen's Gate, Imperial College London, SW7 2AZ, [3] Department of Imaging, University College London Hospitals NHS Foundation Trust, Euston Road, NW1 2BU,

[4] Department of Radiology Hammersmith Hospital, Imperial College Healthcare NHS Trust, DuCane Road, W12 0NN, [5] Department of Radiology, The Royal Marsden NHS Foundation Trust, Fulham Road, London, SW3 6JJ

Corresponding Author: Ioannis Lavdas, Comprehensive Cancer Imaging Centre (C.C.I.C.), Hammersmith Campus, Du Cane Road, W12 0NN, Commonwealth

Building Main Office, Ground Floor, Tel.: +44 020 83838598, Fax: +44 020 8383 1783, Email: ilavdas@imperial.ac.uk

Dr Ioannis Lavdas, Imperial College London

ABSTRACT

Purpose: As part of a programme to implement automatic lesion detection methods for whole body magnetic resonance imaging (MRI) in oncology, we have developed, evaluated and compared three algorithms for fully automatic, multi-organ segmentation in healthy volunteers. **Methods:** The first algorithm is based on classification forests (CFs), the second is based on 3D convolutional neural networks (CNNs) and the third algorithm is based on a multi-atlas (MA) approach. We examined data from 51 healthy volunteers, scanned prospectively with a standardised, multi-parametric whole body MRI protocol at 1.5T. The study was approved by the local ethics committee and written consent was obtained from the participants. MRI data were used as input data to the algorithms, while training was based on manual annotation of the anatomies of interest by clinical MRI experts. Five-fold cross-validation experiments were run on 34 artefact-free subjects. We report three overlap and three surface distance metrics to evaluate the agreement between the automatic and manual segmentations, namely the Dice similarity coefficient (DSC), recall (RE), precision (PR), average surface distance (ASD), root mean square surface distance (RMSSD) and Hausdorff distance (HD). Analysis of variances was used to compare pooled label metrics between the three algorithms and the DSC on a ‘per-organ’ basis. A Mann-Whitney U test was used to compare the pooled metrics between CFs and CNNs and the DSC on a ‘per-organ’ basis, when using different imaging combinations as input for training. **Results:** All three algorithms resulted in robust segmenters that were effectively trained using a relatively small number of data sets, an important consideration in the clinical setting. Mean overlap metrics for all the segmented structures were: CFs: DSC=0.70±0.18, RE=0.73±0.18, PR=0.71±0.14, CNNs: DSC=0.81±0.13, RE=0.83±0.14, PR=0.82±0.10, MA: DSC=0.71±0.22, RE=0.70±0.34, PR=0.77±0.15. Mean surface distance metrics for all the segmented structures were: CFs: ASD=13.5±11.3 mm, RMSSD=34.6±37.6 mm and HD=185.7±194.0 mm, CNNs; ASD=5.48±4.84 mm, RMSSD=17.0±13.3 mm and HD=199.0±101.2 mm, MA: ASD=4.22±2.42 mm, RMSSD=6.13±2.55 mm and HD=38.9±28.9 mm. The pooled performance of CFs improved when all imaging combinations (T2w+T1w+DWI) were used as input, while the performance of CNNs deteriorated, but in neither case, significantly. CNNs with T2w images as input, performed significantly better than CFs with all imaging combinations as input for all anatomical labels, except for the bladder. **Conclusions:** Three state-of-the-art algorithms were developed and used to automatically segment major organs and bones in whole body MRI; good agreement to manual segmentations performed by clinical MRI experts was observed. CNNs perform favourably, when using T2w volumes as input. Using multi-modal MRI data as input to CNNs did not improve the segmentation performance.

Keywords: whole body MRI, fully automatic segmentation, classification forests, convolutional neural networks, multi-atlas segmentation

INTRODUCTION

Recent technological advances in magnetic resonance imaging (MRI) technology, specifically the use of continuously moving table technology, more powerful and faster gradients, phased array coils and parallel acquisition techniques, have allowed whole body MRI to be performed clinically with uncompromised image quality and within reasonable time. The addition of diffusion-weighted imaging (DWI) to whole body protocols [1] means that whole body MRI is now becoming increasingly popular not only for cancer diagnosis and staging, but also for treatment response assessment [2, 3], without the burden of ionising radiation.

One of the most important challenges when reading whole body MRI scans, however, is the increased volume of resulting imaging data, especially when multi-parametric acquisitions are used. As a result, the reading process can become rather time-consuming, with increased risk of misinterpretations. Furthermore, whole body DWI for staging cancer patients suffers from some limitations with respect to its diagnostic performance [4], when compared to other whole body imaging techniques, for example Positron Emission Tomography (PET). Whole body DWI is particularly prone to false-positives, resulting from tissues with normally occurring restricted diffusivity [5].

It would therefore be very beneficial in terms of reading speed and diagnostic performance to develop and evaluate fully automatic methods that identify and segment malignant lesions in whole body MRI scans, whilst recognising normal organs and benign lesions. Such automatic segmentation methods could also find applications in whole body imaging when, for example, adipose or muscle tissue volume evaluation is required [6, 7]. In cancer staging and treatment response monitoring automatic segmentation methods could assist in, for example, automatic tumour detection and volumetric whole body lesion burden assessment [8].

A plethora of segmentation methods has been described in the literature for the main medical imaging modalities used in the clinic (for example MRI or Computed Tomography-CT). Here, we provide a brief overview of the fully automatic segmentation techniques that refer to whole body MRI and/or relate to the machine learning methods we employ in this work. Automatic segmentation methods, which use algorithms other than the ones described here, have been previously described in whole body MRI for the quantification of adipose and muscle tissue [6, 7]. Algorithms based on classification forests (CFs) and variants, have been previously used for the localisation of spinal anatomy [9] or specific/multi-organ segmentation [10, 11] in CT scans and also for automatic detection and segmentation of high grade gliomas [12]. One study has used regression forests to perform multi-organ segmentation in whole body DIXON imaging [13]. A multi-atlas (MA) approach, analogous to the one employed in this work, has been used for segmentation in cardiac MRI [14], while variants have been used in CT imaging [15, 16]. To our knowledge, the use and performance comparison of CFs, convolutional neural networks (CNNs) or MA approaches to perform multi-organ segmentation in whole body MRI, has not been described before.

The purpose of this study was to develop and evaluate three robust algorithms for automatic, multi-organ segmentation in whole body MRI from healthy volunteers, using three state-of-the-art machine leaning approaches. This is a necessary preparatory step towards developing automatic lesion detection methods for whole body MRI in oncology.

MATERIALS AND METHODS

Healthy Volunteers and Imaging Protocol

The study was approved by the local ethics committee, and written consent was obtained from the participants before imaging. Fifty-one healthy volunteers (24 male-mean age=37, range=23-67 years and 27 female-mean age=39, range=23-68 years) were scanned with whole body MRI from February 2012 to May 2014 [17].

Whole body MRI was performed in a moving-table 1.5T system (Siemens Avanto with Syngo MR B17, Erlangen, Germany), using the body coil for transmission and the neck/body phased array coils as receive coils. Four different imaging stations were used to achieve full body coverage, from the top of the neck to mid-thighs. Axial slices were acquired during free-breathing for DWI (*b*=0, 150, 400, 750 and 1000 s/mm^2), while breath-holds were employed for the three first stations for anatomical imaging. DWI slice-matched T1w with DIXON and T2w imaging was also performed. Apparent Diffusion Coefficient (ADC) maps were generated online using a monoexponential fit to the equation: $S=S_0 \cdot e^{-b \cdot ADC}$. The full imaging protocol is shown in Table 1.

Table 1. Whole body imaging protocol used for the healthy volunteers.

Sequence type	SS SE EPI [a]	VIBE [b] with DIXON	HASTE [c]
FOV (mm)	450×366	450×351	450×366
Matrix size	128×128 interpolated	320×202	256×256
No of slices/ thickness/distance (mm)	50/5/0%	56/5/20%	50/5/0%
***TR* (ms)**	9000	7.54	767
***TE* (ms)**	72	2.38/4.76	92
Bandwidth (Hz/pixel)	2056	300	399
Flip Angle	90	10	130
N_A	4	1	2
Fat suppression	STIR [d] (*TI*=180 ms)	N/A	N/A
***b*-values (s/mm^2)**	0,150,400,750,1000	N/A	N/A
Parallel Acquisition	GRAPPA [e] 2	GRAPPA 2	GRAPPA 2
No stations	4, free-breathing	4, (3 with breath-holds)	4, (3 with breath-holds)
T_A (min)/station	8.17	0.15	1.18

[a] SS SE EPI=single-shot spin echo echo planar imaging, [b] VIBE=3D volumetric interpolated breath-hold examination, [c] HASTE=half-Fourier acquisition single-shot turbo spin-echo, [d] STIR=short inversion time inversion recovery, [e] GRAPPA=generalised autocalibrating partially parallel acquisition

Classification Forests (CFs) algorithm

CFs are powerful, multi-label classifiers that facilitate simultaneous segmentation of multiple organs. They have very good generalisation properties, meaning that the algorithm can be effectively trained using a relatively small amount of annotated example data, a particularly important advantage in the clinical setting.

CFs is a supervised, discriminative learning technique, which is based on random forests (RFs); an ensemble of weak classifiers called decision trees [18]. Each decision tree is constructed in a way that it produces a partitioning of the training data, e.g., image points that carry organ label information, in a way that training data with same labels are grouped together. This is achieved by building the trees from the root node down to the leaf nodes. Internal nodes, so called split nodes, separate the incoming data into two sets. Leaf nodes then correspond to small clusters of training data from which label statistics are computed and are used for predictions at testing time. Data splitting in the trees is based on an objective function, which maximises the information gain over empirical label distributions. The goal is to select discriminative features at split nodes that are best for partitioning the data. Different trees are built by injecting randomness for both feature selection and training data subsampling. This ensures decorrelation between trees and has proven to yield good generalisation properties. During testing, image points from a new image are ‘pushed’ through each tree until a leaf node is reached. The label statistics over training data that are stored in the leaf nodes are aggregated over different trees by simple averaging, and a final decision on the most likely label is made based on this aggregation. Intuitively, image points will fall into leaf nodes that contain similar image points from the training data with respect to the features that are evaluated along the path from root to leaf node.

An attractive property of CFs is their ability to automatically select the right image features for a given task, from a potentially very large and high-dimensional pool of possible features [19]. This is important because pre-selecting or hand-crafting image features beforehand can be very difficult, as it is not known in advance which features are discriminative for the task at hand. In CFs the user only has to provide weak guidance on the ranges of parameters that are used to randomly generate potential features. In this work, we make use of the popular offset box-features, which have been shown to provide effective means of capturing local and contextual information [12]. Box-features are very efficient to compute, which is beneficial for training and testing. In box-features, intensity averages are calculated within randomly sized and displaced 3D boxes. Two types of features are computed; single-box and two-box features. Single-box features simply correspond to the average intensity of all voxels from a particular MRI sequence that fall into a 3D box. Two-box features return the difference between the averages computed for each of the two boxes and generalise intensity gradient features. Here, each box can be taken from a different MRI sequence and thus yield cross-sequence information.

Tuning parameters for our algorithm have been set accordingly to knowledge from previous applications, such as vertebra localization in whole-body CT scans [9]. We have used CFs extensively for related tasks for which cross-validation has been used to optimise hyperparameters such as tree depth [9, 12]. In this work, we used 50 trees with a maximum tree depth of 30. The stopping criterion for growing trees is if either the objective function (information gain) cannot be further improved or the number of training samples in a leaf fall below a threshold of four samples. We found that neither increasing the number of trees nor the tree depth increases the segmentation accuracy of the CFs.

Convolutional Neural Networks (CNNs) algorithm

CNNs are feed-forward artificial neural networks, which have recently emerged as powerful machine learning methods for image analysis tasks, such as segmentation. CNNs are capable of learning complex, non-linear data associations between the input images and segmentation labels through layers of feature extractors. Each layer performs multiple convolutional filter operations on the data coming in from the previous layer and outputs feature responses, which are then processed by the next layer. The last layer in the network combines all the outputs to make a prediction about the most likely class label for each voxel in an image. The parameters of the convolutions and weights for combining feature responses are learned during the training stage, using an algorithm called back-propagation. The layered architecture enables CNNs to learn complex features automatically without any need for guidance from the user. The features correspond to a sequence of filter kernels learned in consecutive layers of the neural network. A final feature that is used for classification thus, can correspond to a non-linear combination of individual features that are extracted hierarchically. This is also called features-of-features, as filter kernels in deeper layers are applied to the feature responses of earlier layers. This is different to CFs, where the user has to define a pool of potential features beforehand from which the most discriminative ones are then selected during CF training. However, CNNs come with an increased computational cost during training, and they have multiple meta-parameters that need to be highly tuned to achieve optimal performance, a process which can be challenging for less experienced users. In addition, defining the right network architecture is a challenge on its own and a field of active research.

Here, we make use of a recently published CNN approach that we developed originally for the task of brain lesion segmentation in multi-parametric MRI [20]. The approach, called DeepMedic, uses a dual pathway CNN that processes an image at different levels of resolution simultaneously. This has the advantage that features are based on both local and contextual information, something that can be particularly appealing in the case of whole body multi-organ segmentation. For example, the left and right kidneys might look very similar locally and share similar features at small scale, but the contextual features that cover larger regions of the images, allow the discrimination between the left and right body parts.

The CNN configuration used here follows largely the default configuration that has been previously used for brain lesion segmentation [20]. To accommodate for larger context in the case of organ segmentation, the receptive field for the low-resolution pathway has been increased by using an image downsampling factor of 3. We use a dual pathway (two resolutions), 11-layer deep CNN, where the last two layers correspond to fully connected layers, which combine the features extracted on the two resolution pathways. We employ 50-70 feature maps (that is different kernels) for each layer. The network architecture is fully convolutional and there are no max-pooling layers, which we find to increase segmentation accuracy. The CNN architecture is a balance between model capacity, training efficiency, and memory demands. Further details about DeepMedic are provided in [20]. An open source implementation is available at URL [21].

Multi-atlas (MA) algorithm

Our third algorithm is based on a MA label propagation approach [14]. Multi-atlas segmentation uses a set of atlases (images with corresponding segmentations) that represent the inter-subject variability of the anatomy to be segmented. Each atlas is registered to the new image to be segmented using a deformable image registration. The MA approach accounts for anatomical shape variability and is more robust than single atlas propagation methods in that at any errors associated with propagation, are averaged out when combining multiple atlases. The approach employed here makes use of efficient 3D-3D intensity-based image registration [22] with free-form deformations as the transformation model and correlation coefficient as the similarity measure. Majority voting is used to derive the final tissue label at each voxel.

The source code for all the algorithms described in this work is publicly available, and we can provide configuration files upon request.

Training of CFs and CNNs is a demanding process computationally and in our case took up to 12 hours for CFs and 30 hours for CNNs for a single fold with 27 images, when using a quad-core Intel Xeon 3.5 GHz workstation with 32 GB RAM and an NVIDIA Titan X graphics processor unit (GPU). Our CFs implementation uses all available central processor units (CPUs), while the CNN implementation runs mostly on the GPU. Training only needs to be performed once. Testing of new data points to obtain the full segmentation of an image is a particularly efficient process and takes about a minute for CFs and CNNs. Note, that the MA algorithm does not require any training, but has considerably longer running time during testing which scales linearly with the number of atlases. To segment a single image using 27 atlases takes about 15 minutes on CPU. Table 2 is comparing the strengths and weaknesses of the three algorithms.

Table 2. Strengths and weaknesses of the three developed algorithms, with respect to each other.

	CFs	**CNNs**	**MA**
Strengths	• Straightforward training • Relatively short training time • Easy to implement • Runs on standard CPU [a]	• Automatic feature learning • Capable of learning complex data associations • Spatially smooth predictions	• No training required • Straightforward to add new atlases • Very intuitive as based on image alignment • Preserves anatomical structure
Weaknesses	• Limited feature complexity • Noisy predictions	• Complexity of training configuration • Increased training time • Requires high-end GPUs [b] • Difficult to implement	• Increased testing time • Not so good generalisation • Misses fine details in structural variation

[a] CPU: Central processing unit, [b] GPU: Graphics processing unit

DICOM data from individual imaging stations were stitched into single NIfTI volumes (https://nifti.nimh.nih.gov/). The MRI data were used as input data to the algorithms, while training was based on manual annotation of the anatomies of interest on the T2-weighted volumes, first segmented by two radiology trainees (HM and AS) and an MR physicist (IL, 5 years of experience in whole body MRI). An MRI expert (AR, 17 years of experience in MRI) then checked the segmentations, which were adjusted, if needed, and agreed in consensus. When multi-modal MRI data were used as input to CFs and CNNs (for example, T2w+T1w+DWI data-where T1w refers to T1w in- and opposed-phase images from the DIXON acquisitions and DWI refers to *b*=1000 s/mm^2 images and ADC maps) an extra, registration, step was added to the data preparation pipeline. During this step, T1w and DWI volumes were affinely registered to the T2w volumes. A schematic overview of the data preparation process, including the registration step, is given in Figure 1.

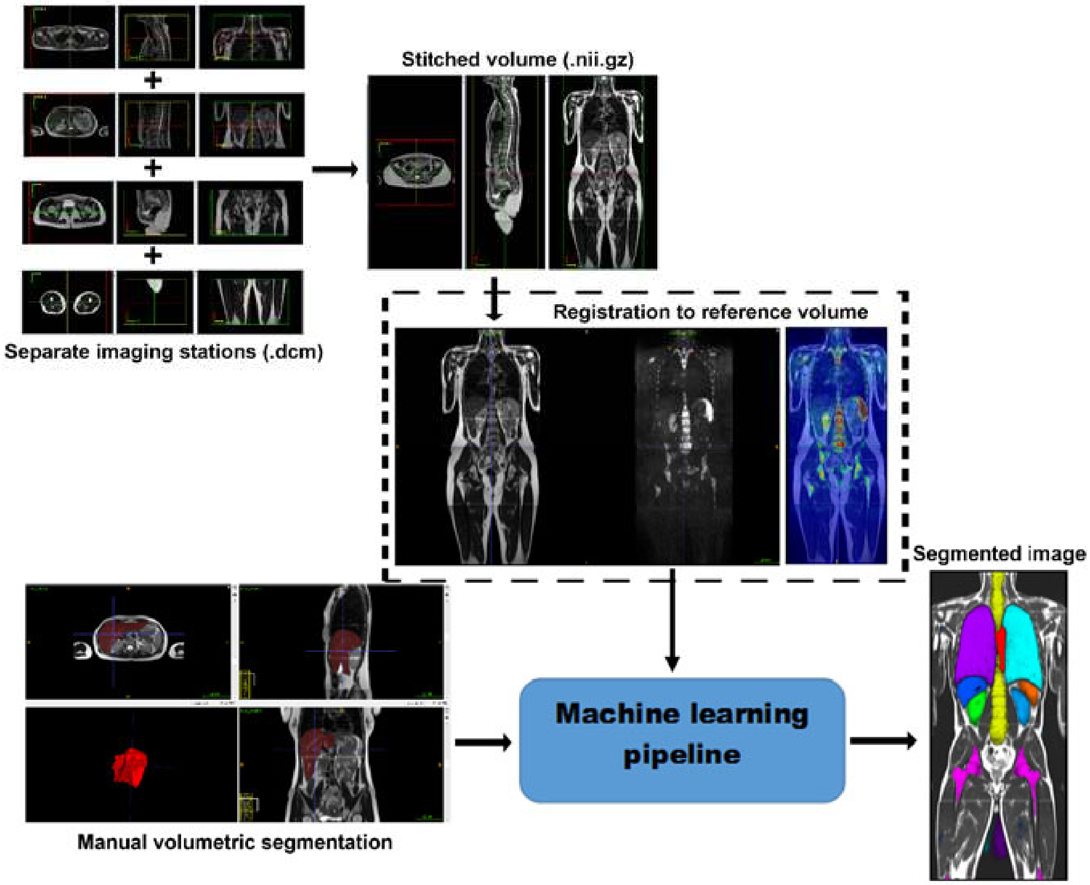
Stitched volume (.nii.gz)
Separate imaging stations (.dcm)
Registration to reference volume
Segmented image
Machine learning pipeline
Manual volumetric segmentation

Figure 1. Diagrammatic flowchart of the data preparation process for the machine learning pipelines. Data from different imaging stations are stitched to single volumes and then intra-patient registration is performed (when using multi-modal MRI data as input to the algorithms). Manual segmentation and annotation of the anatomies of interest is also performed to generate training data for the machine learning algorithms.

Quantitative Analysis and Evaluation

We run five-fold cross-validation experiments on 34 artefact-free data sets to assess the agreement of segmentations between the ones from the developed algorithms and the ones from the clinical experts. All data sets were inspected by an expert radiologist (AR) before being selected for validation. Data sets with severe motion artefacts or DWI data sets with severe distortion artefacts, and therefore severe misalignment, were excluded from validation.

We report six metrics (three overlap and three surface distance based measures) to assess the agreement between automatic segmentation results from our algorithms and the manual segmentations performed by the clinical experts. The Dice similarity coefficient (DSC) quantifies the match between the two segmentations (1=complete overlap, 0=no overlap). Recall (RE) can be expressed in terms of sensitivity (1=no misses) and precision (PR) can be expressed in terms of specificity (1=no false positives). The average surface distance (ASD) is the average of all the distances from points on the boundary of the automatic segmentation to the boundary of the manual segmentation (0=perfect match), the root mean surface distance (RMSSD) is calculated in the same way as the ASD, except that the distances are now squared (0=perfect match). Finally, the Hausdorff distance (HD) or maximum surface distance, is the maximal distance from a point in the first segmentation to a nearest point in manual segmentation (0=perfect match) [23]. The three surface distance metrics are expressed in mm and are unbounded.

We measured the above metrics for the right and left lungs (RLNG and LLNG), liver (LVR), gallbladder (GBLD), right and left kidneys (RKDN and LKDN), spleen (SPLN), pancreas (PNCR), bladder (BLD), spine (SPN) and pelvic bones, including the femurs (PLVS) for all three algorithms, when using T2w volumes as inputs. Then, we did the same when using all imaging combinations (T2w+T1w+DWI) as inputs to CFs and CNNs.

Statistical Analysis

One-way analysis of variances (ANOVA) was used to compare the mean metrics for all the examined structures between the three algorithms. Post hoc analysis (multiple comparisons) was performed with a Tukey test. In cases where the homogeneity of variances was violated, a Kruskal-Wallis test was used. A Mann-Whitney U test was used to compare the performance between CFs and CNNs when using T2w volumes as input to the algorithms and when using all imaging combinations (T2w+T1w+DWI). ANOVA and Mann-Whitney U tests were similarly used to compare the DSC of individual anatomical labels between the three algorithms and between CFs and CNNs when using different imaging inputs. Finally, a Mann-Whitney U test was used to compare the DSC between CFs with all imaging combinations (T2w+T1w+DWI) as input and CNNs with T2w images as input only, for each anatomical label. A significance level of 0.05 was used for all tests. Statistical analysis was performed in SPSS 21.0 for Windows (SPSS, Chicago, Ill).

RESULTS

A visual example of automatic segmentation results from the three algorithms in the coronal and axial plane is shown in Figure 2.

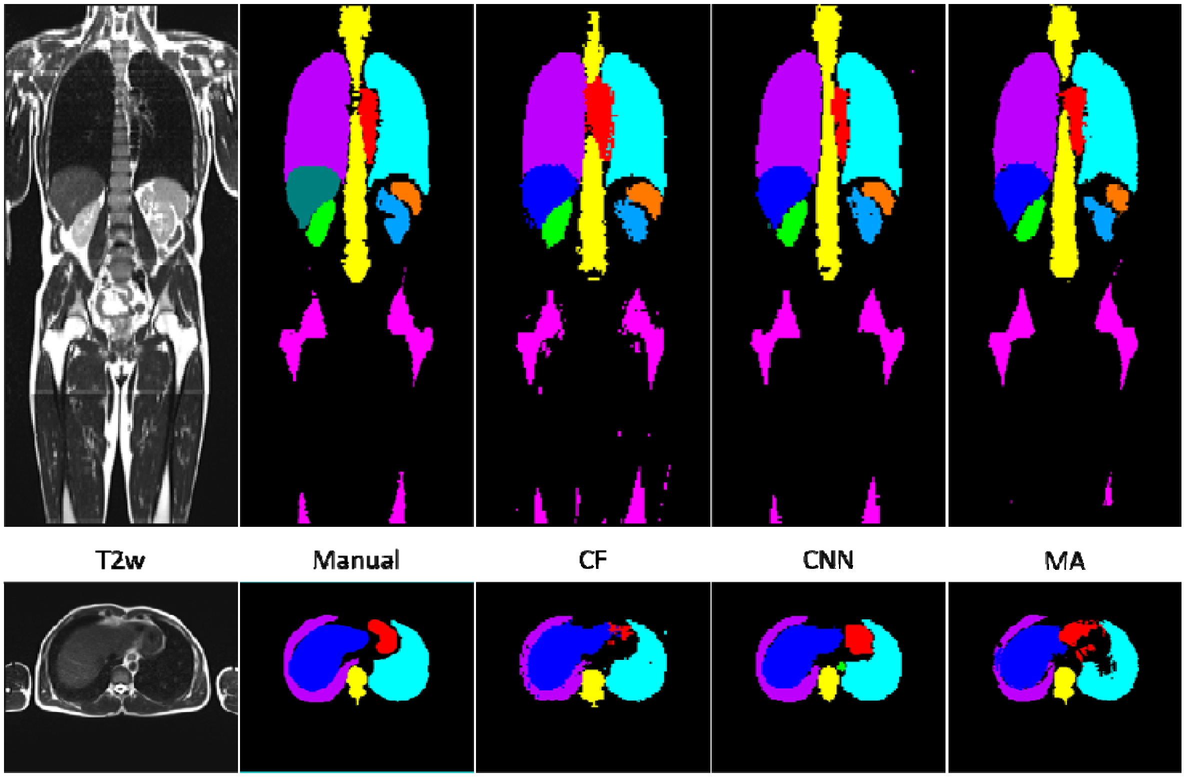

Figure 2. T2w representative coronal (top row) and axial slices (bottom row), manual and automatic segmentations of major organs (lungs, heart, kidneys, liver and spleen) and bones (spine and femurs) from the three algorithms.

A bar chart that provides a pictorial representation of the mean metrics (DSC, RE, PR, ASD, RMSSD and HD) for the segmented organs when using T2w volumes as input to all three algorithms, is shown in Figure 3.

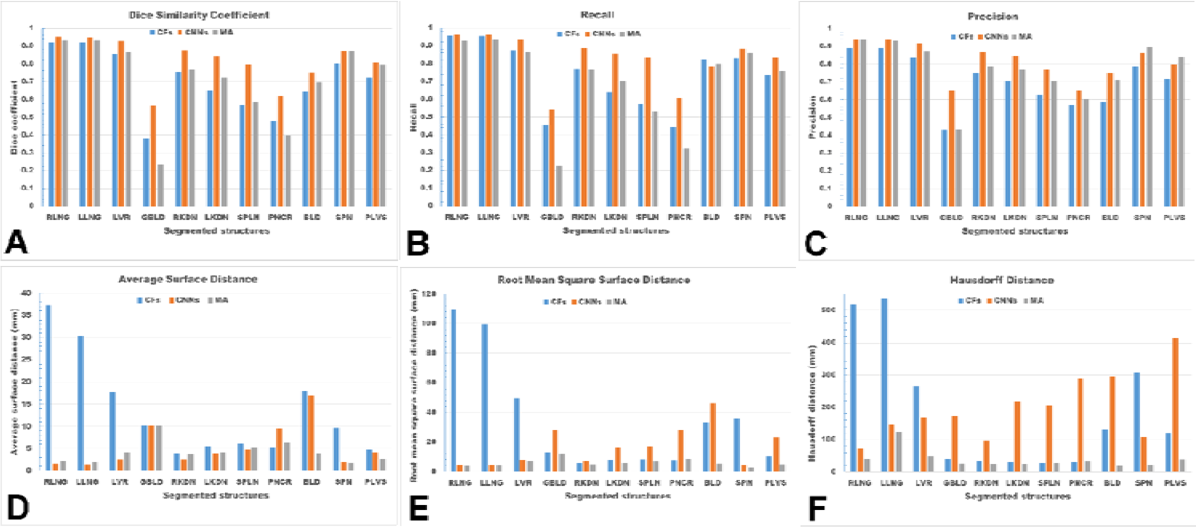

Figure 3. Bar chart showing the mean measured metrics, DSC (A), RE (B), PR (C), ASD (D), RMSSD (E) and HD (F) for the segmented organs (RLNG and LLNG: right and left lungs, LVR: liver, GBLD: gallbladder, RKDN and LKDN: right and left kidneys, SPLN: spleen, PNCR: pancreas and BLD: bladder) and bones (SPN: spine and PLVS: pelvis) for the three algorithms (CFs), (CNNs) and (MA).

It is noteworthy that an ‘at a glance’ qualitative assessment reveals that CNNs outperform CFs and the MA algorithm in DSC, RE and PR, while the MA algorithm seems to perform best in terms of surface distance metrics, namely ASD, RMSSD and HD.

Table 3 shows the pooled mean metrics ± standard deviation from all the segmented structures for the three algorithms. It also shows the *P* values from the analysis of variances when comparing the metrics between the three algorithms.

Table 3. Pooled mean metrics ± standard deviation from all the segmented structures from the three algorithms (CFs, CNNs and MA). In addition, *P* values from

the analysis of variances when comparing the metrics between the three algorithms (ANOVA for DSC, RE and PR and Kruskal-Wallis for ASD, RMSSD and HD). Significant values are shown in bold.

	DSC	**RE**	**PR**	**ASD (mm)**	**RMSSD (mm)**	**HD (mm)**
CFs	0.70±0.18	0.73±0.18	0.71±0.14	13.5±11.3	34.6±37.6	185.7±194.0
CNNs	0.81±0.13	0.83±0.14	0.82±0.10	5.48±4.84	17.0±13.3	199.0±101.2
MA	0.71±0.22	0.70±0.24	0.77±0.15	4.22±2.42	6.13±2.55	38.9±28.9
P	0.271	0.294	0.185	**0.005**	**0.004**	**0.001**

It is seen that CNNs provide the highest mean DSC (0.81±0.13), RE (0.83±0.14) and PR (0.82±0.10) than CFs and the MA algorithm, but not statistically significant (*P*=0.271, 0.294 and 0.185 respectively). On the contrary, the MA algorithm returns the lowest ASD (4.22±2.42 mm), RMSSD (6.13±2.55 mm) and HD (38.9±28.9 mm), when compared to CFs and CNNs, which is statistically significant (*P*=0.005, 0.004 and 0.001 respectively).

Table 4 reports the DSC, the most commonly used metric to assess agreement between manual and automatic segmentations, for individual anatomical structures (labels) when the three algorithms (CFs, CNNs and MA) are using the T2w images as inputs only. It also shows the *P* values from the analysis of variances, when comparing the DSC between the three algorithms for each anatomical label.

Table 4. DSC ± standard deviation for each anatomical label, segmented by the three algorithms (CFs, CNNs and MA), when using T2w images as input only. In addition, *P* values from the analysis of variances when comparing the DSC between the three algorithms (ANOVA for DSC, RE and PR and Kruskal-Wallis for ASD, RMSSD and HD). Significant values are shown in bold.

	DSC			
	CFs	**CNNs**	**MA**	***P***
RLNG	0.92±0.03	0.95±0.01	0.93±0.01	**<0.001**
LLNG	0.92±0.03	0.95±0.01	0.93±0.01	**<0.001**
LVR	0.85±0.03	0.93±0.01	0.86±0.04	**<0.001**
GBLD	0.38±0.26	0.56±0.19	0.24±0.26	**<0.001**
RKDN	0.75±0.09	0.87±0.03	0.77±0.07	**<0.001**
LKDN	0.65±0.19	0.84±0.11	0.72±0.13	**<0.001**
SPLN	0.57±0.18	0.79±0.11	0.58±0.14	**<0.001**
PNCR	0.47±0.13	0.62±0.09	0.40±0.14	**<0.001**
BLD	0.65±0.22	0.75±0.21	0.69±0.23	0.162
SPN	0.80±0.04	0.87±0.01	0.87±0.02	**<0.001**
PLVS	0.73±0.05	0.81±0.03	0.79±0.06	**<0.001**

It is worth noting that CNNs performed significantly better (*P*<0.001) than CFs and the MA algorithm in segmenting all the anatomies of interest, except for the bladder (*P*=0.162).

A bar chart that provides a pictorial representation of the mean metrics (DSC, RE, PR, ASD, RMSSD and HD) for the segmented organs when using T2w volumes and all imaging combinations (T2w+T1w+DWI) as input to CFs and CNNs, is shown in Figure 4.

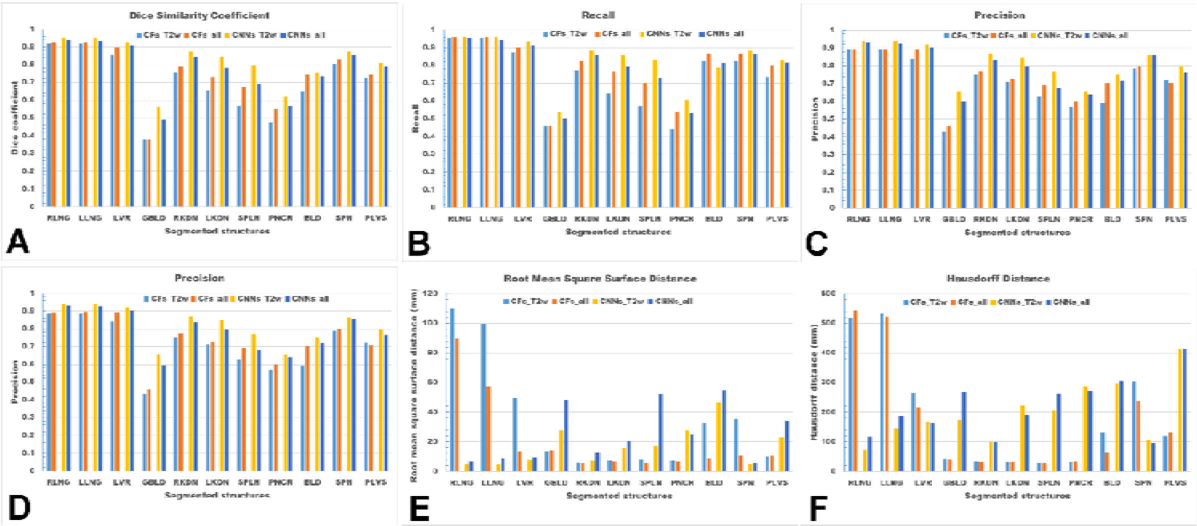

Figure 4. Bar chart comparing the mean measured metrics, DSC (A), RE (B), PR (C), ASD (D), RMSSD (E) and HD (F) for the segmented organs (RLNG and LLNG: right and left lungs, LVR: liver, GBLD: gallbladder, RKDN and LKDN: right and left kidneys, SPLN: spleen, PNCR: pancreas and BLD: bladder) and bones (SPN: spine and PLVS: pelvis), when using T2w volumes and all imaging combinations (T2w+T1w+DWI) as inputs to CFs and B CNNs.

It appears that the use of all imaging combinations (T2w+T1w+DWI) as input to CFs improves both the overlap (DSC, RE and PR) and surface distance (ASD, RMSSD and HD) metrics, but the opposite happens for CNNs, where the algorithm seems to perform better when using T2w volumes as input only.

Table 5 shows the pooled mean metrics ± standard deviation from all the segmented structures for CFs and CNNs, when using T2w only volumes and all imaging combinations (T2w+T1w+DWI) as inputs. It also shows the *P* values from the Mann Whitney U test when comparing the two input cases for CFs and CNNs.

Table 5. Pooled mean metrics ± standard deviation from all the segmented structures for CFs and CNNs, when using T2w only volumes and all imaging combinations (T2w+T1w+DWI) as inputs. In addition, *P* values from the Mann-Whitney U test when comparing the two input cases for CFs and CNNs.

	DSC	**RE**	**PR**	**ASD (mm)**	**RMSSD (mm)**	**HD (mm)**
CFs_T2w	0.70±0.17	0.73±0.18	0.71±0.14	13.5±11.2	34.6±37.6	185.7±194.0
CFs_all	0.74±0.16	0.78±0.16	0.74±0.13	7.89±7.55	20.9±27.1	170.7±194.0
P	0.491	0.412	0.533	**0.039**	0.309	0.974
CNNs_T2w	0.81±0.12	0.82±0.14	0.82±0.10	5.48±4.84	17.0±13.3	199.0±101.2
CNNs_all	0.77±0.14	0.79±0.15	0.79±0.11	9.23±8.04	25.2±19.1	215.9±98.6
P	0.412	0.450	0.450	0.178	0.224	0.224

It is confirmed that the performance of CFs is improved when all imaging combinations are used (T2w+T1w+DWI) as input, when compared to using T2w volumes only. This is reflected in all metrics (DSC=0.74±0.16 vs. 0.70±0.17, RE=0.78±0.16 vs. 0.73±0.18, PR=0.74±0.13 vs. 0.71±0.14, ASD=7.89±7.55 mm vs. 13.5±11.2 mm, RMSSD=20.9±27.1 mm vs. 34.6±37.6 mm and HD=170.7±194.0 mm vs. 185.7±194.0 mm). On the contrary, the performance of CNNs is better when using T2w volumes only as input, rather than using all imaging combinations (T2w+T1w+DWI). This is again reflected in all metrics (DSC=0.81±0.12 vs.

0.77±0.14, RE=0.82±0.14 vs. 0.79±0.15, PR=0.82±0.10 vs. 0.79±0.11, ASD=5.48±4.84 mm vs. 9.23±8.04 mm, RMSSD=17.0±13.3 mm vs. 25.2±19.1 mm and HD=199.0±101.2 mm vs. 215.9±98.6 mm). No significant differences were found in the performance of CFs and CNNs, when using different T2w only and all imaging combinations (T2w+T12w+DWI) as inputs.

Table 6 shows the DSC for all the anatomical labels, when CFs and CNNs are being used with T2w images only (CFs_T2w and CNNs_T2w) as inputs and when using all imaging combinations (T2w+T1w+DWI) as input to the two algorithms (CFs_all and CNNs_all). It also shows the *P* values from the Mann-Whitney U tests when comparing the DSC between CFs and CNNs used with different imaging inputs.

Table 6. DSC ± standard deviation from CFs and CNNs for all the anatomical labels, when using T2w only images (CFs_T2w and CNNs_T2w) and when using all imaging combinations (T2w+T1w+DFWI) as inputs (CFs_all and CNNs_all). In addition, *P* values from the Mann-Whitney tests. Significant values are shown in bold.

	DSC			DSC		
	CFs_T2w	**CFs_all**	***P***	**CNNs_T2w**	**CNNs_all**	***P***
RLNG	0.92±0.03	0.92±0.02	0.564	0.95±0.01	0.94±0.01	**0.001**
LLNG	0.92±0.03	0.92±0.02	0.500	0.95±0.01	0.93±0.03	**0.003**
LVR	0.85±0.03	0.90±0.02	**<0.001**	0.93±0.01	0.91±0.03	**<0.001**
GBLD	0.38±0.26	0.38±0.25	0.976	0.56±0.19	0.49±0.18	0.079
RKDN	0.75±0.09	0.79±0.06	0.093	0.87±0.03	0.84±0.05	**<0.001**
LKDN	0.65±0.19	0.73±0.13	**0.023**	0.84±0.11	0.78±0.13	**<0.001**

SPLN	0.57±0.18	0.67±0.15	**<0.001**	0.79±0.11	0.69±0.13	**<0.001**
PNCR	0.47±0.13	0.55±0.11	**0.017**	0.62±0.09	0.57±0.11	0.051
BLD	0.65±0.22	0.74±0.18	**0.046**	0.75±0.21	0.74±0.16	0.411
SPN	0.80±0.04	0.83±0.03	**<0.001**	0.87±0.01	0.85±0.05	**0.044**
PLVS	0.73±0.05	0.74±0.05	0.135	0.81±0.03	0.78±0.06	0.069

It is seen that the addition of extra imaging modalities (T1w+DWI) as input to CFs_T2w, significantly improves the segmentation performance (*P*<0.046) for many anatomical structures (LVR, LKDN, SPLN, PNCR, BLD and SPN). By contrast, the addition of T1w+DWI to CNNs_T2w, significantly deteriorates the DSC (*P*<0.044) for most the examined anatomies of interest (RLNG, LLNG, LVR, RKDN, LKDN, SPLN and SPN).

Finally, Table 7 shows and compares the DSC from all anatomical labels, when segmented by the two algorithms with the best DSC performance as reported above, namely CFs_all and CNNs_T2w. It also shows the *P* values from the Mann-Whitney U tests to compare the DSC between the two algorithms for all the examined structures.

Table 7. DSC ± standard deviation from all the examined structures for CFs_all and CNNs_T2w algorithms. Also, *P* values from the Mann-Whitey U tests to compare the DSC between the two algorithms for each segmented structure. Significant values are shown in bold.

	DSC		
	CFs_all	**CNNs_T2w**	***P***
RLNG	0.92±0.02	0.95±0.01	**<0.001**
LLNG	0.92±0.02	0.95±0.01	**<0.001**
LVR	0.90±0.02	0.93±0.01	**<0.001**
GBLD	0.38±0.25	0.56±0.19	**0.002**
RKDN	0.79±0.06	0.87±0.03	**<0.001**
LKDN	0.73±0.13	0.84±0.11	**<0.001**
SPLN	0.67±0.15	0.79±0.11	**<0.001**
PNCR	0.55±0.11	0.62±0.09	**0.008**
BLD	0.74±0.18	0.75±0.21	0.384
SPN	0.83±0.03	0.87±0.01	**<0.001**
PLVS	0.74±0.05	0.81±0.03	**<0.001**

It is striking that CNNs_T2w scored significantly better DSCs than CFs_all in all the examined organs (*P*<0.008), apart from the bladder (*P*=0.384). The segmentation performance was notably improved when using CNNs_T2w, even for organs with great variability in appearance, such as the gallbladder (0.38±0.25 for CNNs_T2w vs. 0.56±0.19 for CFs_all, *P*=0.002).

DISCUSSION

All the algorithms tested in this study, permitted automatic, multi-organ segmentation in whole body MRI of healthy volunteers with very good agreement to the segmentations, performed manually by clinical experts. Accurate, multi-organ, automatic segmentation in whole body MRI is the first step in training machine-learning algorithms to recognise normality. This will lead the way to developing automatic identification and segmentation algorithms for lesions, such as primary or metastatic tumours, with increased sensitivity and specificity. These algorithms could ultimately facilitate the process of reading whole body scans in cancer patients by reducing the reading time and, possibly, improving the diagnostic accuracy of whole body MRI. These algorithms may also assist in quantifying the extent of normal tissues such as muscle or fat.

Our analysis showed that CNNs outperformed CFs and the MA algorithm when T2w volumes were used as input to the algorithms and when using pooled overlap evaluation metrics (DSC, RE and PR) to assess the accuracy of segmentation. When the performance of the algorithms was assessed with pooled surface distance metrics (ASD, RMSSD and HD), it was the MA algorithm, that performed best. Single misinterpreted voxels in CFs and CNNs can greatly elevate ASD, RMSSD and HD; these metrics are particularly sensitive to outliers.

We then assessed the pooled metrics performance of CFs and CNNs when using all imaging combinations (T2w+T1w+DWI) as input, arguing that maximisation of training information to the algorithms might improve the performance of segmentation [12]. We found that the performance of CFs was improved, however not significantly, when using all imaging combinations as input for training. The opposite was observed for CNNs.

The findings for the pooled metrics analysis, described above, were corroborated by a 'per-organ' quantitative analysis of the commonly used DSC, to assess the performance of our segmentation algorithms. This analysis confirmed that for all individual anatomical structures (except for the bladder), the algorithm that returned the greatest DSC was CNNs with T2w images only used as input.

Because our structural scans were acquired using breath-holds and the DWI ones with free breathing, we found that there was significant displacement between soft tissues in anatomical areas adjacent to the diaphragm between these types of scans. As the employed affine registration method [24] cannot fully compensate for non-linear motions caused by breathing, we assume that misregistration could be the reason why the performance of CNNs, despite performing better than the other two algorithms when using T2w volumes as input only, was degraded when using all imaging combinations as input for training. A more robust, non-linear registration method could improve the accuracy of CNNs and further improve the performance of CFs, so we are currently looking into methods to address this issue. Alternatively, we could have generated training data by manually segmenting the structures of interest on each sequence type separately, but this would be a rather strenuous and time-consuming approach.

The performance of our methods cannot be directly compared to similar methods in the literature because there is no previous work describing automatic, simultaneous segmentation of healthy organs and bones in multi-parametric whole body MRI. We believe, however, that our methods may compare favourably to other machine learning methods for detection and segmentation in medical imaging in that our classifiers are inherently multi-label and have shown that can be effectively trained when using a relatively small number of data sets, something that is very important in the clinical setting. However, we would still need to address the performance limitations of our algorithms when segmenting organs with big variability in appearance (for example, the gallbladder or the pancreas).

CONCLUSIONS

In conclusion, we have developed and evaluated three state-of-the-art algorithms that automatically segment healthy organs and bones in whole body MRI with accuracy comparable to the one achieved manually by clinical experts. An algorithm based on CNNs and trained using T2w only images as input, performs favourably when compared to CFs or a MA algorithm, trained with either T2w only images or a combination of imaging inputs (T2w+T1w+DWI). Using multi-modal MRI data as input for training the developed algorithms did not improve the segmentation performance in this work, but it is anticipated to improve the segmentation performance if more effective whole body registration between the various imaging modalities can be performed. This investigation is the first step towards developing robust algorithms for the automatic detection and segmentation of benign and malignant lesions in whole body MRI scans for staging of cancer patients.

ACKNOWLEDGEMENTS

This work is supported by the National Institute of Health Research (EME Project: 13/122/01).

CONFLICTS OF INTEREST

None declared.

REFERENCES

1. Takahara T, Imai Y, Yamashita T, et al. Diffusion weighted whole body imaging with background body signal suppression (DWIBS): technical improvement using free breathing, STIR and high resolution 3D display. *Radiation Medicine*. 2004; 22(4): 275-282.
2. Koh DM and Collins DJ. Diffusion-Weighted MRI in the Body: Applications and Challenges in Oncology. *American Journal of Roentgenology*. 2007; 188(6): 1622-1635.
3. Schmidt GP, Reiser MF, and Baur-Melnyk A. Whole-body MRI for the staging and follow-up of patients with metastasis. *European Journal of Radiology*. 2009; 70(3): 393-400.
4. Wu L-M, Gu H-Y, Zheng J, et al. Diagnostic value of whole-body magnetic resonance imaging for bone metastases: a systematic review and meta-analysis. *Journal of Magnetic Resonance Imaging*. 2011; 34(1): 128-135.
5. Padhani AR, Koh D-M, and Collins DJ. Whole-Body Diffusion-weighted MR Imaging in Cancer: Current Status and Research Directions. *Radiology*. 2011; 261(3): 700-718.
6. Jin Y, Imielinska CZ, Laine AF, et al., Segmentation and Evaluation of Adipose Tissue from Whole Body MRI Scans, in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003: 6th International Conference, Montréal,*

Canada, November 15-18, 2003. Proceedings, R.E. Ellis and T.M. Peters, Editors. 2003, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 635-642.

7. Karlsson A, Rosander J, Romu T, et al. Automatic and quantitative assessment of regional muscle volume by multi-atlas segmentation using whole-body water–fat MRI. *Journal of Magnetic Resonance Imaging*. 2015; 41(6): 1558-1569.
8. Blackledge MD, Collins DJ, Tunariu N, et al. Assessment of Treatment Response by Total Tumor Volume and Global Apparent Diffusion Coefficient Using Diffusion-Weighted MRI in Patients with Metastatic Bone Disease: A Feasibility Study. *PLoS ONE*. 2014; 9(4): e91779.
9. Glocker B, Konukoglu E, and Haynor DR, Random Forests for Localization of Spinal Anatomy, in *Medical Recognition, Segmentation and Parsing*, S. Zhou, Editor. 2015, Academic Press, Elsevier: London. p. 94-109.
10. Glocker B, Pauly O, Konukoglu E, and Criminisi A, Joint Classification-Regression Forests for Spatially Structured Multi-object Segmentation, in *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, A. Fitzgibbon, et al., Editors. 2012, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 870-881.
11. Cuingnet R, Prevost R, Lesage D, et al., Automatic Detection and Segmentation of Kidneys in 3D CT Images Using Random Forests, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part III*, N. Ayache, et al., Editors. 2012, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 66-74.
12. Geremia E, Zikic D, Clatz O, et al., Classification Forests for Semantic Segmentation of Brain Lesions in Multi-channel MRI, in *Decision Forests for Computer Vision and Medical Image Analysis*, A. Criminisi and J. Shotton, Editors. 2013, Springer London: London. p. 245-260.
13. Pauly O, Glocker B, Criminisi A, et al., Fast Multiple Organ Detection and Localization in Whole-Body MR Dixon Sequences, in *Medical Image Computing and*

Computer-Assisted Intervention – MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part III, G. Fichtinger, A. Martel, and T. Peters, Editors. 2011, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 239-247.

14. Bai W, Shi W, O'Regan DP, et al. A Probabilistic Patch-Based Label Fusion Model for Multi-Atlas Segmentation With Registration Refinement: Application to Cardiac MR Images. *IEEE Transactions on Medical Imaging*. 2013; 32(7): 1302-1315.
15. Wolz R, Chu C, Misawa K, et al. Automated Abdominal Multi-Organ Segmentation With Subject-Specific Atlas Generation. *IEEE Transactions on Medical Imaging*. 2013; 32(9): 1723-1730.
16. Okada T, Yokota K, Hori M, et al., Construction of Hierarchical Multi-Organ Statistical Atlases and Their Application to Multi-Organ Segmentation from CT Images, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008: 11th International Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part I*, D. Metaxas, et al., Editors. 2008, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 502-509.
17. Lavdas I, Rockall AG, Castelli F, et al. Apparent Diffusion Coefficient of Normal Abdominal Organs and Bone Marrow From Whole-Body DWI at 1.5 T: The Effect of Sex and Age. *American Journal of Roentgenology*. 2015; 205(2): 242-250.
18. Breiman L. Random Forests. *Machine Learning*. 2001; 45(1): 5-32.
19. Criminisi A, Shotton J, and Konukoglu E. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. *Foundations and Trends® in Computer Graphics and Vision*. 2012; 7(2–3): 81-227.
20. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*. 2017; 36: 61-78.

21. Kamnitsas K. *DeepMedic source code 2016*. https://github.com/Kamnitsask/deepmedic. Accessed November, 2016.

22. Glocker B, Komodakis N, Tziritas G, Navab N, and Paragios N. Dense image registration through MRFs and efficient linear programming. *Medical Image Analysis*. 2008; 12(6): 731-741.

23. Heimann T, van Ginneken B, Styner MA, et al. Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. *Medical Imaging, IEEE Transactions on*. 2009; 28(8): 1251-1265.

24. Studholme C, Hill DLG, and Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*. 1999; 32(1): 71-86.