Deep quantitative liver segmentation and vessel exclusion to assist in liver assessment

Benjamin Irving[1], Chloe Hutton[1], Andrea Dennis[1], Sid Vikal[1], Marija Mavar[1], Matt Kelly[1], and Sir J. Michael Brady[12]

[1] Perspectum Diagnostics, Oxford, UK,
[2] Department of Oncology, University of Oxford, UK,
ben.irving@perspectum-diagnostics.com

Abstract. Liver disease, especially Non-Alcoholic Fatty Liver Disease has reached high levels, and there is a need for non-invasive tests based on quantitative MRI to replace biopsy in order to better assess liver health. An automated quantitative liver segmentation approach is required to automate these tests and in this work we propose a fully convolutional framework with a novel objective function for quantitative liver segmentation. The method has (to date) been tested on quantitative T1 maps generated from the UK Biobank study. We obtained extremely encouraging results on an unseen test set with a Dice score of 0.95, and Sensitivity 0.98 and Specificity 0.99.

Keywords: segmentation, MRI, liver, convolutional neural networks, deep learning

1 Introduction

Liver disease has already reached high levels worldwide [12]. In some developed countries, up to one third of all adults have some form of liver disease, increasingly Non-Alcoholic Fatty Liver Disease (NAFLD). Up to 12 percent of people with NAFLD go on to develop the more severe Non-Alcoholic Steatohepatitis (NASH) [2]. The current reference standard for the diagnosis and grading of liver disease is biopsy, but this is limited by its invasiveness, frequent complications, and sampling: a liver biopsy typically represents just 1/50,000th of the whole liver volume and is unable to characterise heterogeneous liver tissue. Furthermore, there can be considerable variability in histological interpretation of liver biopsy samples [3]. These factors highlight the need for a non-invasive method to assess quantitatively a greater volume of the liver.

Quantitative MRI has been shown to be one of the key modalities for non-invasive assessment of NAFLD [10]. It enables quantitative and repeatable assessment of the whole liver region. In practice, quantitative measurements are typically performed manually by a trained human operator who places regions of interest within the liver image, avoiding vessels and image artefacts. Such manual analysis inevitably leads to inter-rater variability and possibly bias, not least when the liver tissue is particularly heterogeneous. This highlights the need

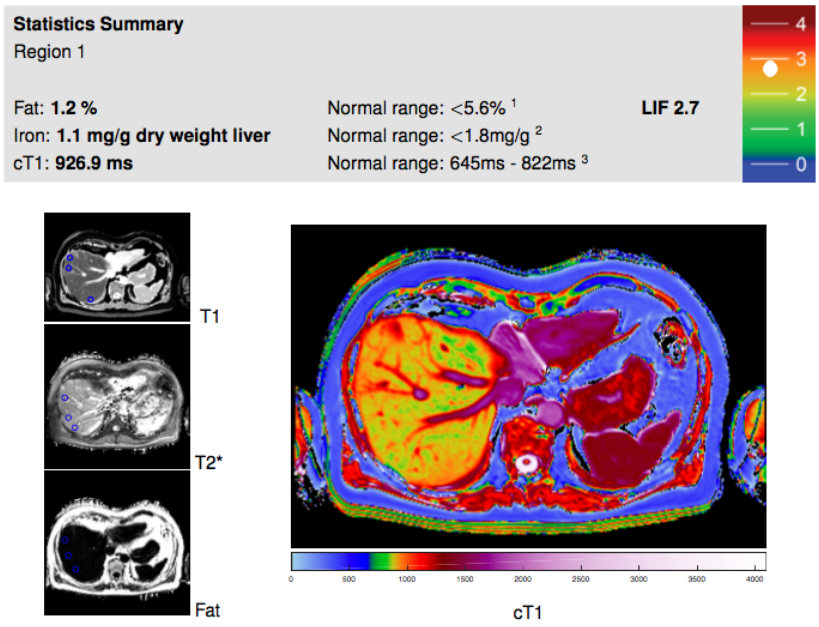

Fig. 1. Perspectum Diagnostics's LIF liver score calculation and corrected T1 based on multi-parametric quantitative analysis

to automate this part of the analysis, and a critical step is automated liver segmentation.

Perspectum Diagnostics (www.perspectum-diagnostics.com) provides a cloud-based service that enables quantification of liver health based on cT1 (quantitative T1 corrected for iron), T2* (to measure iron burden), and Proton Density Fat Fraction (PDFF), whose fusion as LiverMultiscan is effective for detecting NAFLD [1]. The cT1 is measured in milliseconds, typically in the range 500-1500ms, but is often mapped onto a scale of 0-4, called the Liver Inflammation and Fibrosis (LIF) Score [10] (see Figure 1). This enables a hepatologist to relate the LIF score to histology grades such as the Ishak score. The LIF score is based on the distribution of cT1s in patients whose disease status is confirmed by biopsy. Perspectum is analysing many thousands of cases per annum, including those from the ongoing UK Biobank study (http://www.ukbiobank.ac.uk) which currently has over 10,000 cases [13]. Studies of this scale further emphasise the need for automated and objective, quantitative analyses of data in order to provide population based biomarkers for use in prospective studies of liver disease. Semi-automatic liver segmentation methods have already been built into Perspectum's analysis workflow; however, this paper reports a fully automatic segmentation method to reduce user interaction and inter-rater variability.

Liver segmentation from MRI is challenging, not just because of the intrinsic variability of (diseased) liver tissue, but because of the variability in acquisitions and protocols, as well as motion artefacts arising from potentially longer acquisition times compared to CT. However, quantitative imaging sequences enable calculation of the underlying tissue parameters such as T1, T2* and PDFF, which are robust to variation in acquisition (T1 and T2* are still related to field strength). These techniques require the acquisition of a series of images to construct the quantitative maps and so currently just one (or a limited number of) 2D slices are acquired to minimise acquisition time and, therefore, liver motion effects. This makes the segmentation more challenging because we cannot use volumetric shape information as a prior and regions of the liver may appear disconnected in the axial slice. In addition, for quantitative assessment of liver parenchyma, it is also important to exclude ducts and larger blood vessels from the segmentation – making this a unique challenge. In NAFLD, there is also considerable variation in liver health due to fibrosis, steatosis, and any segmentation method must be robust to such variations.

Deep convolutional networks have shown considerable potential in image analysis for detection and segmentation [11, 8]. In this paper, we demonstrate that a deep network is effective for the delineation of the liver and exclusion of vessels in quantitative images, even where there is considerable variation of liver health. This work makes a number of novel contributions: we are not aware of any studies that have used quantitative MRI scans to perform automated liver segmentation even though this provides a reliable quantification of the underlying tissue, and deep learning in quantitative MRI also appears to be novel.

Our method builds on fully convolutional neural network research and applies it to quantitative liver segmentation. We make a number of improvements including modification of the loss function, to make it more appropriate to segmentation in biological images. In Section 2 we discuss previous methods for liver segmentation. We introduce our approach in Section 3. Our method has been tested on a cohort from the Biobank trial (Section 4) and results from an independent test set are shown in Section 5.

2 Background

Few methods have been developed to segment the liver region from MRI because of the challenges in scanner and sequence variability; rather, most reported methods have been developed for CT [6]. Cheng et al. propose a MRI liver segmentation method that uses a 2D liver shape model as a prior [4]. However, it assumes that the liver appears as a single connected region in the acquisition slice, which is not always the case for certain slices through the liver. We also wish to exclude liver vessels as part of the segmentation and so a shape model is not appropriate. Masoumi et al. [9] combine the watershed transform with a neural network to optimise the segmentation; the network is used to iteratively optimise the parameters of the watershed transform. This approach is limited

by the definition of the watershed transform and would not be effective for cases with a large variation in liver health and pathology.

Instead, we aim to take advantage of modern approaches to segmentation – in particular, fully convolutional neural networks such as U-Nets [11, 8]. Such methods use stacked convolutional, ReLu, and pooling layers to automatically learn features (low level features such as edges in the first layers to high level features such as textures and objects in the later layers). Fully convolutional networks do not have any fully connected layers and have been shown to be effective for pixelwise labelling of an image [8]. In the U-Net formulation, the first half of the network combines convolutional and max pooling layers to learn higher order representations within the image, while the second part is based on upsampling and convolution, and translates the representation back into a pixelwise labelling. Merge layers combine low level features into the final segmentation [11].

3 Method

A fully convolutional deep neural network was developed for this analysis based on U-Nets [11]. The input is a 2D T1 quantitative map whose dimensions are 288×384. The network architecture is shown in Figure 2 and uses 15 stacked convolutional layers with 3×3 kernels. Between every second layer of the first 6 there are 2×2 pooling layers to produce a high level representation and 2×2 upsampling layers in the second half to convert that representation into a pixelwise segmentation. Convolutional weights were initialised with a Glorat uniform initialiser and biases were set to 0 . The upsampling followed by a merge layer is used to combine the low level features with higher level features into the final segmentation (upsampling is used to create images of the same dimension for merging). The network is somewhat shallower than U-Nets because of the currently limited training data. We also modify the network to avoid representionational bottlenecks by extending the number of filters in the layer before applying maxpooling rather than the layer after maxpooling as shown in Figure 2. This minimises the loss of the representation during the pooling stage. The number of filters per layer are shown in the figure. The final layer translates the features into a fuzzy image segmentation. A threshold of 0.5 is used to convert the segmentation into a binary labelling.

Data augmentation To date, we have worked with a relatively small dataset and there is an inevitable risk of overfitting to the training set. For this reason, data augmentation is required. We applied random affine transforms to each batch during training in a range of 4 degrees rotation, 10% translation, and 10% scaling. This transformation is applied to each case at training time, so at every epoch the same case will have a different transformation.

***SensSpec* Objective function** A novel objective function was used to train the method. This score aims to jointly optimise sensitivity and specificity of the

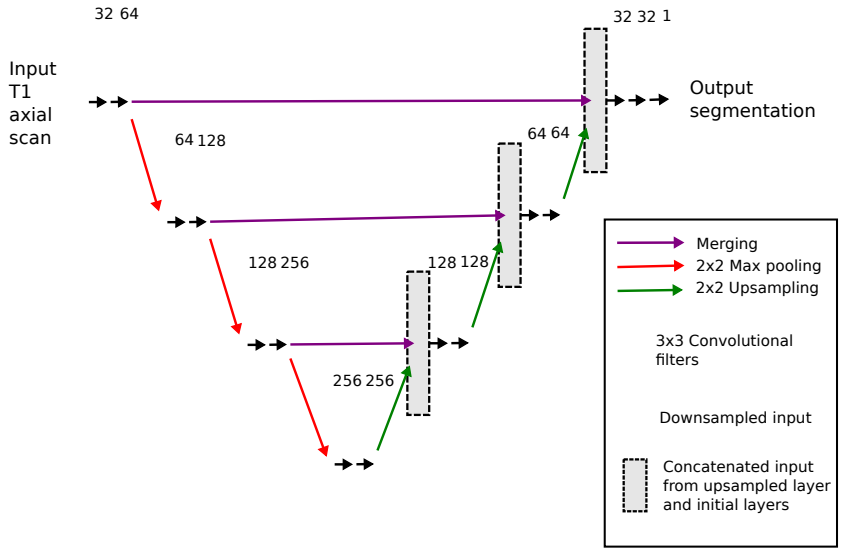

Fig. 2. Fully convolutional segmentation framework for T1 liver segmentation

detection compared to the ground truth, which we found to produce a stable optimisation.

We chose this objective function because in medical images the labelled region often occupies only a small proportion of the entire image – unlike scene labelling problems used in computer vision. We found that this makes commonly used objective functions such as cross-entropy and mean squared error less effective because there was a bias towards the background label on problems such as this where the training set is unbalanced. The SensSpec objective that we propose is as follows:

$$L(\theta, \hat{\theta}) = -\ln(Se\ Sp) \tag{1}$$

where

$$Se = \sum \theta_i \hat{\theta}_i / \sum \theta_i,\ \forall \theta_i = 1, \text{ and } \ Sp = \sum (1-\theta_i)(1-\hat{\theta}_i) / \sum (1-\theta_i),\ \forall \theta_i = 0 \tag{2}$$

where Se and Sp are Sensitivity and Specificity respectively. θ is a vector of ground truth image pixel labels and $\hat{\theta}$ are the fuzzy predicted labels. The $\forall$ operator means that only ground truth foreground values are used to calculate Sensitivity and only ground truth background values are used to calculate specificity.

4 Data

UK Biobank is a groundbreaking trial for assessing risk factors in an apparently healthy population aged between 40-69 years. The trial is ongoing and aims to collect biomarkers, lifestyle factors, and medical images from the UK population. As part of this study, Perspectum Diagnostics performs Liver Inflammation and Fibrosis (LIF) analysis using quantitative MRI acquired in the study participants [13].

MR imaging was performed on a Siemens 1.5T MAGNETOM Aera at the dedicated Biobank Imaging Centre at Cheadle (UK). The shMOLLI acquisition protocol, which samples the T1 recovery curve using a single-shot steady state free precession acquisition, was used to acquire single slice T1 relaxation time maps in a transverse plane through the right lobe of the liver and spleen. Acquisition parameters were TR=4.94ms, TE=1.93 ms, flip angle = 35 degrees and voxel size = 1.15 x 1.15 x 8mm. Quantitative T1 parameter maps were then calculated by fitting T1 recovery curves to the acquired data.

For this initial study, we used 170 cases that had been segmented by an expert. Since the UK Biobank participant population mainly represents a disease-free population, this study included 100 participants with LIF ≥ 2 and 70 with LIF < 2. This is because LIF ≥ 2 has been shown to correlate closely with subsequent adverse liver events [10]. The dataset ensures a "sufficient" number of unhealthy cases in the cohort. The cohort was then randomly split into 80% training and 20% test (unseen by the algorithm). A second cohort of 100 Biobank cases (also unseen by the algorithm) was acquired after the initial submission as a general test with no specific requirement for pathology.

As noted above, "ground truth" liver segmentations had been delineated by an expert using an in-house semi-automatic liver segmentation tool for all cases. The tool used level sets based on user-defined landmark points to segment the liver and remove vessels. The user can then manually refine the segmentation if required. Finally, a third cohort of 166 cases was semi-automatically segmented by two readers, and was used to assess the DSC inter-rater agreement. This third cohort partly overlaps the first two. It is not exactly the same because we selected data that had already been annotated by two readers for other purposes.

5 Experimentation and Results

The fully convolutional network was implemented in Python using Keras [5] with Tensorflow as the backend. The method was trained on the 136 training set for 1500 epochs in batches of 20 cases (with augmentation). The Adam optimiser [7] with a learning rate of 5E-5 was used with our proposed SensSpec objective function. Training time took 290 minutes using an Ubuntu system with an Nvidia Titan X GPU. Once trained, the mean time was 2.81s per case on a Macbook with a dual core i5 CPU (The TitanX machine is only used during training).

Dice Similarity Coefficient (DSC) was used to evaluate the similarity between the semi-automatic ground truth and the automatic segmentation defined as

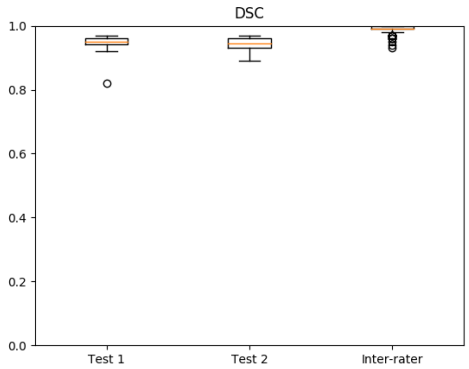

Fig. 3. A boxplot showing the DSC of the original test set (Test 1 with n=34), an extended test set (Test 2 with n=100) and the DSC inter-rater agreement using a semi-automatic method.

follows:

$$\frac{2|\theta \cap \hat{\theta}|}{|\theta| + |\hat{\theta}|} \tag{3}$$

where θ and $\hat{\theta}$ are the automated segmentation and ground truth. A DSC of 1 is a perfect match between the regions and a DSC of 0 means that there was no overlap. On the unseen test sets, the method achieved a DSC 0.95 and 0.94, respectively for the two test sets. The inter-rater variability using the semi-automatic method was 0.99. A boxplot of the DSC scores for the entire test set is shown in Figure 3. Mean sensitivity and specificity for the test set were 0.98 and 0.99, respectively. Figures 4 and 5 show the ground truth and automated segmentation for example cases in the test sets.

6 Discussion and Conclusion

This method provides a highly accurate and completely automated approach to segmenting the liver (and excluding vessels and ducts) from quantitative T1 images, and is the first step in complete automation of T1 liver tissue assessment from MRI. The method achieved a DSC of 0.95 and 0.94 for the two test sets compared to the semi-automatic ground truth. The inter-reader variability was 0.99. However, this was performed using the same in-house semi-automatic software so it is likely that there is a tendency towards similarity. Next, we plan to have the automatic and semi-automatic ground truth blindly assessed to determine which appears because we have had qualitative feedback that the automated method often appears to be better than the ground truth.

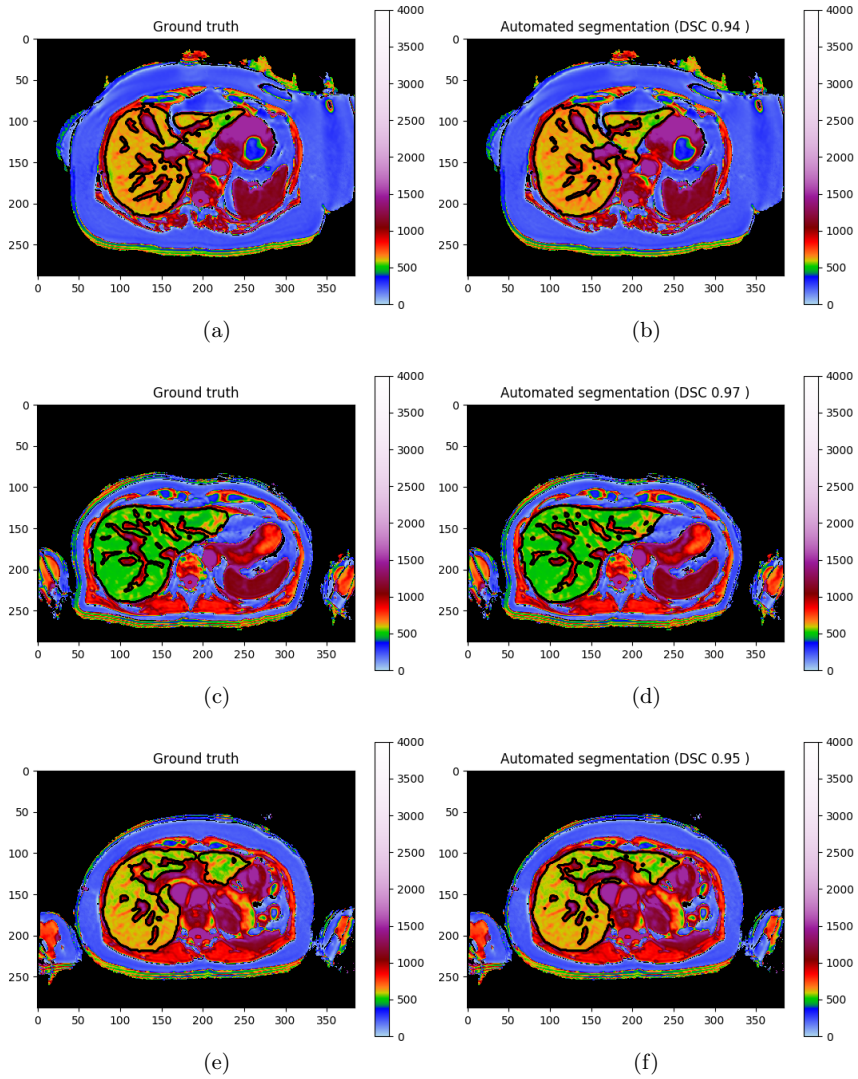

Fig. 4. Semi-automatic ground truth segmentation vs the proposed automated segmentation scheme on quantitative T1 maps. The colormap visualises normal and abnormal T1 ranges (in ms, green = normal, red = high)

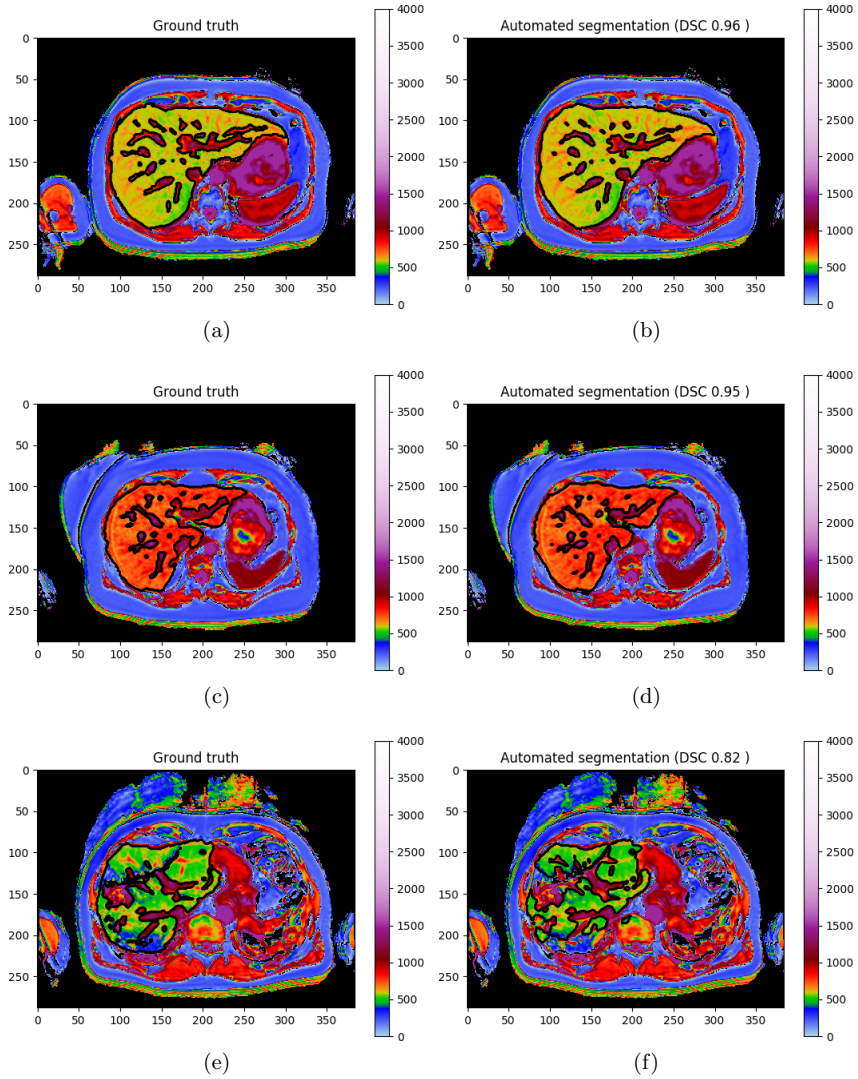

Fig. 5. Semi-automatic ground truth segmentation vs the proposed automated segmentation scheme on quantitative T1 maps (in ms)

Once trained, the segmentation method is fast with a mean time of 2.81s just on the CPU, and therefore a GPU is only required for training the network. This is useful for deployment in production systems.

The poorest performing case (outlier in Test 1 of Figure 3) had a DSC of 0.82 and is shown on the bottom row of Figure 5. This is a particularly complex case due to the presence of large vessels and ducts. There also appears to be breathing artefacts in this case. It could be argued that the automated segmentation method is more sensitive to liver parenchyma and excludes more vessels than the ground truth.

To understand if the method is overfitting we also calculated the DSC for the training set. The mean DSC for the training data was 0.96 which suggests that data augmentation has been effective in making the training robust and avoiding overfitting. We still expect that using a larger training set will make the method more robust.

The loss function that we propose provides a stable optimisation and is robust to training against unbalanced dataset where the region of interest is much smaller than the background. The loss function could be adapted to maximise a different classification operating point of the Receiver Operating Characteristics, if desired. For comparison, we attempted to use the mean squared error but, due to the unbalanced labelling, this loss function quickly fell into a local minima where the entire image was classified as non-liver.

UK Biobank has over 10,000 images, and is scheduled to grow to 100,000, which highlights the need for an automated approach. In this initial study we used 136 cases for training, an initial test set of 34 and later an additional test set of 100 because of the time taken to create a ground truth. We have shown that with data augmentation the fully convolutional network is capable of learning from relatively small datasets. In future we intend to train the approach on a much larger cohort once we have performed manual labelling. This would also allow testing of a deeper implementation of the network.

Acknowledgements

This research has been conducted using the UK Biobank Resource under application 9914.

References

1. Banerjee, R., Pavlides, M., Tunnicliffe, E.M., Piechnik, S.K., Sarania, N., Philips, R., Collier, J.D., Booth, J.C., Schneider, J.E., Wang, L.M., Delaney, D.W., Fleming, K.A., Robson, M.D., Barnes, E., Neubauer, S.: Multiparametric magnetic resonance for the non-invasive diagnosis of liver disease. Journal of Hepatology 60(1), 69 – 77 (2014)
2. Blachier, M., Leleu, H., Peck-Radosavljevic, M., Valla, D.C., Roudot-Thoraval, F.: The burden of liver disease in europe: A review of available epidemiological data. Journal of Hepatology 58(3), 593 – 608 (2013)

3. Castera, L., Pinzani, M.: Non-invasive assessment of liver fibrosis: are we ready? The Lancet 375(9724), 1419 (2010)
4. Cheng, K., Gu, L., Wu, J., Li, W., Xu, J.: A Novel Level Set Based Shape Prior Method for Liver Segmentation from MRI Images, pp. 150–159. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
5. Chollet, F.: Keras. https://github.com/fchollet/keras (2015)
6. Heimann, T., van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., Binnig, G., Bischof, H., Bornik, A., Cashman, P.M.M., Chi, Y., Cordova, A., Dawant, B.M., Fidrich, M., Furst, J.D., Furukawa, D., Grenacher, L., Hornegger, J., KainmÃIJller, D., Kitney, R.I., Kobatake, H., Lamecker, H., Lange, T., Lee, J., Lennon, B., Li, R., Li, S., Meinzer, H.P., Nemeth, G., Raicu, D.S., Rau, A.M., van Rikxoort, E.M., Rousson, M., Rusko, L., Saddi, K.A., Schmidt, G., Seghers, D., Shimizu, A., Slagmolen, P., Sorantin, E., Soza, G., Susomboon, R., Waite, J.M., Wimmer, A., Wolf, I.: Comparison and evaluation of methods for liver segmentation from ct datasets. IEEE Transactions on Medical Imaging 28(8), 1251–1265 (Aug 2009)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR abs/1412.6980 (2014), http://arxiv.org/abs/1412.6980
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
9. Masoumi, H., Behrad, A., Pourmina, M.A., Roosta, A.: Automatic liver segmentation in mri images using an iterative watershed algorithm and artificial neural network. Biomedical Signal Processing and Control 7(5), 429 – 437 (2012)
10. Pavlides, M., Banerjee, R., Sellwood, J., Kelly, C.J., Robson, M.D., Booth, J.C., Collier, J., Neubauer, S., Barnes, E.: Multiparametric magnetic resonance imaging predicts clinical outcomes in patients with chronic liver disease. Journal of hepatology 64(2), 308–315 (2016)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597 (2015), http://arxiv.org/abs/1505.04597
12. Wang, F.S., Fan, J.G., Zhang, Z., Gao, B., Wang, H.Y.: The global burden of liver disease: The major impact of china. Hepatology 60(6), 2099–2108 (2014)
13. Wilman, H.R., Kelly, M., Garratt, S., Matthews, P.M., Milanesi, M., Herlihy, A., Gyngell, M., Neubauer, S., Bell, J.D., Banerjee, R., et al.: Characterisation of liver fat in the uk biobank cohort. PloS one 12(2), e0172921 (2017)