Diagnostic and Interventional Imaging (2019) **100**, 211—217

ORIGINAL ARTICLE / *Computer developments*

Kidney cortex segmentation in 2D CT with U-Nets ensemble aggregation

V. Couteaux[a,b,*], S. Si-Mohamed[c,d], R. Renard-Penna[e], O. Nempont[a], T. Lefevre[a], A. Popoff[a], G. Pizaine[a], N. Villain[a], I. Bloch[b], J. Behr[f], M.-F. Bellin[g], C. Roy[h], O. Rouvière[i], S. Montagne[j], N. Lassau[k], L. Boussel[c,d]

[a] *Philips Research France, 33, rue de Verdun, 92150 Suresnes, France*
[b] *LTCI, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France*
[c] *CREATIS, CNRS UMR 5220, Inserm U1206, INSA-Lyon, Claude Bernard Lyon 1 University, 69100 Villeurbanne, France*
[d] *Department of Radiology, Hospices Civils de Lyon, 69002 Lyon, France*
[e] *Department of Radiology, Hôpital Tenon, AP—HP, GRC-UPMC n°5 Oncotype-URO, Sorbonne universités, 75020 Paris, France*
[f] *Department of Radiology, CHRU de Besançon, 25000 Besançon, France*
[g] *Department of Radiology, Hôpitaux Universitaires Paris Sud, 94270 Le Kremlin Bicêtre, France*
[h] *Department of Radiology, CHU de Strasbourg, Nouvel Hôpital Civil, 67000 Strasbourg, France*
[i] *Department of Uroradiology, Hospices Civils de Lyon, Faculté de Médecine Lyon Est, 69002 Lyon, France*
[j] *Department of Radiology, Hôpital Pitié Salpétrière, AP—HP, 75013 Paris, France*
[k] *Department of Radiology, Gustave Roussy, IR4M, UMR8081, CNRS, Université Paris-Sud, Université Paris-Saclay, 94805 Villejuif, France*

KEYWORDS
Renal cortex;
Image segmentation;
Artificial intelligence (AI);
Computed tomography (CT)

Abstract
Purpose: This work presents our contribution to one of the data challenges organized by the French Radiology Society during the *Journées Francophones de Radiologie*. This challenge consisted in segmenting the kidney cortex from coronal computed tomography (CT) images, cropped around the cortex.
Materials and methods: We chose to train an ensemble of fully-convolutional networks and to aggregate their prediction at test time to perform the segmentation. An image database was made available in 3 batches. A first training batch of 250 images with segmentation masks was provided by the challenge organizers one month before the conference. An additional training batch of 247 pairs was shared when the conference began. Participants were ranked using a Dice score.

* Corresponding author. Philips Research France, 33, rue de Verdun, 92150 Suresnes, France.
E-mail address: vincent.couteaux@telecom-paristech.fr (V. Couteaux).

https://doi.org/10.1016/j.diii.2019.03.001

Results: The segmentation results of our algorithm match the renal cortex with a good precision. Our strategy yielded a Dice score of 0.867, ranking us first in the data challenge.
Conclusion: The proposed solution provides robust and accurate automatic segmentations of the renal cortex in CT images although the precision of the provided reference segmentations seemed to set a low upper bound on the numerical performance. However, this process should be applied in 3D to quantify the renal cortex volume, which would require a marked labelling effort to train the networks.

Renal diseases are often associated with cortical morphological changes, such as volume reduction or notch defect. All these features are considered as surrogate markers of renal diseases and can be visible on imaging examinations, such as ultrasound, magnetic resonance imaging (MRI), or computed tomography (CT) [1,2]. Despite a well-established qualitative assessment of the renal cortex with these modalities, a quantitative approach helps improve the diagnostic work-up of renal diseases [3]. However, to date quantitative assessment of renal cortex is hampered by complex and time-consuming analyses such as semi-automated segmentations based on a pixel value threshold algorithm, region growing, appearance models combined with graph cuts or random forests [4—8]. The recent development of convolutional neural networks (CNN), as well as the access to very large imaging databases, could help overcome these limitations. Very promising results have recently been obtained in several applications such as the segmentation of cardiac chambers, and the brain [9,10]. However, the appropriate artificial intelligence (AI) tools for kidney analysis still need to be developed.

Fully-convolutional networks have drastically improved the state-of-the-art in image segmentation [11]. U-Nets are currently a standard approach for two-dimensional (2D) or three-dimensional (3D) medical image segmentation problems [12—18].

The *Journées Francophones de Radiologie* was held in Paris in October 2018. For the first time this year, the French Society of Radiology organized an AI competition. Teams of industrial researchers, students, and radiologists were invited to take part in five data challenges. In this paper, we present our approach to address the kidney cortex segmentation challenge aiming at segmenting the renal cortex on 2D coronal CT images.

Method

Kidney cortex segmentation challenge

An image database was made available in 3 batches. A first training batch of 250 images with segmentation masks was provided by the challenge organizers one month before the conference. An additional training batch of 247 pairs was shared when the conference began. Two days later, the teams were ranked on a test batch of 299 images.

CT images in the coronal plane, cropped and resized around the kidney (192×192 pixels with a pixel size of 1×1 mm and intensity in Hounsfield units [HU]) were provided (Fig. 1). The reference segmentation was provided as a binary mask for each image of the training set. Due to the usual difficulties of manual segmentation, in particular for irregularly shaped objects such as the renal cortex, several reference segmentations were debatable or even erroneous. We observed that a proportion of the pixels at the edge of the cortex were either left out when they should not have been, or mislabeled as cortex while clearly outside (Fig. 1c). Moreover, blood vessels inside the kidney were occasionally included in the reference segmentation, but this was inconsistent throughout the dataset (Fig. 1d). In fact, it can be hard to distinguish actual renal columns from some blood vessels. We clipped the image intensity values between −150 HU and 200 HU and rescaled them between 0 and 1. This range has been chosen manually to contain all the renal cortex dynamic and limit the influence of high values in the image, corresponding to bones, and very low values, corresponding to air.

To address the specific difficulties of this challenge, such as the imprecision of the reference segmentations, we adopted several popular strategies such as artificial data augmentation, meta parameter optimization, pre-training and post-processing with connected components analysis [19—22]. We also used ensemble aggregation, a standard machine learning technique frequently applied to deep learning [12,22,23].

Network architecture

We chose a U-Net architecture with 5 levels of depth, residual blocks, and rectified linear units (ReLU) activation functions, and added convolutions on the skip connections (Fig. 2) [18,24—26]. We set the meta-parameters using a Bayesian optimization approach [19,20]. We used artificial data augmentation during training to limit overfitting, by randomly applying translations, rotations, zooms, noise, brightness and contrast shifts to the input samples. The training was performed until convergence and lasted

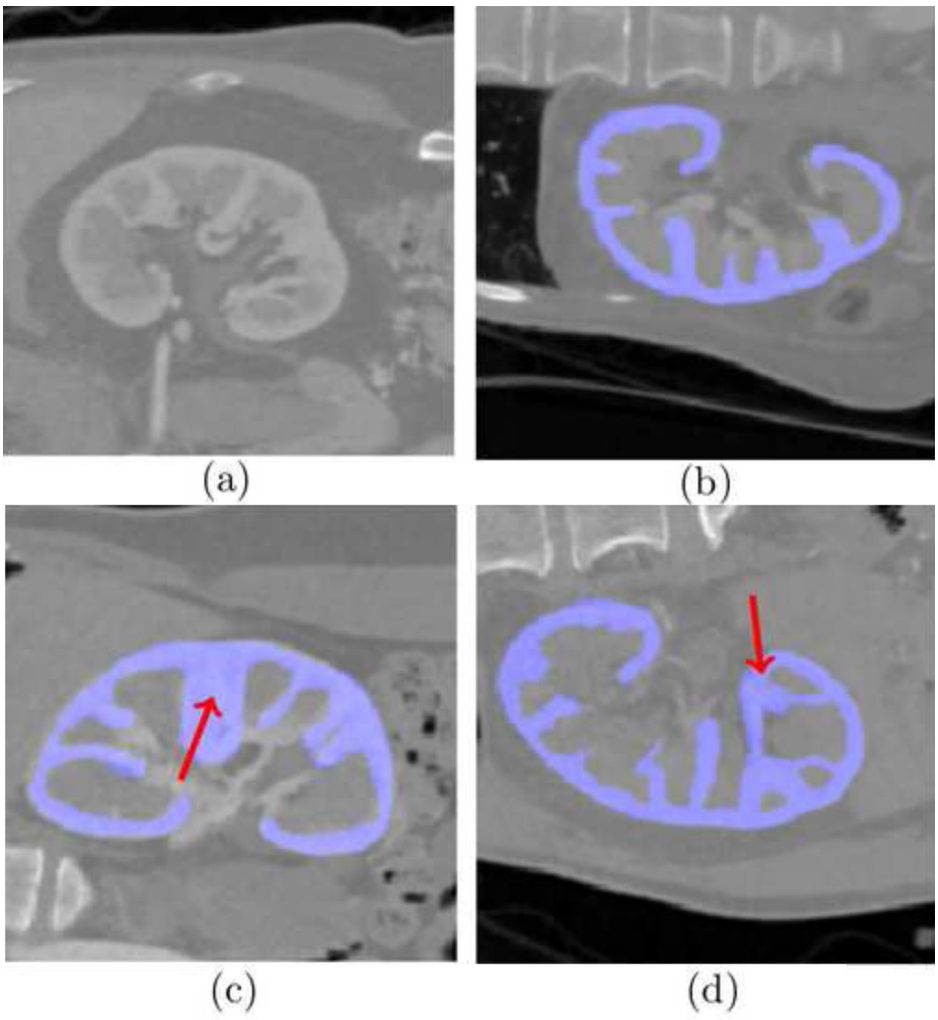

Figure 1. CT images of the kidney from the training set provided by the data challenge organizers. The reference segmentation is overlapped in blue; a: image only; b: correct segmentation; c: inaccurate segmentation and renal column clusters (arrow); d: blood vessels included in the segmentation (arrow).

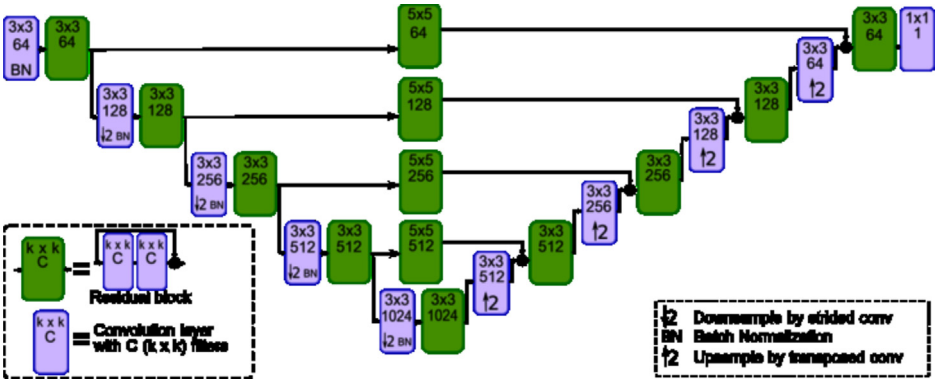

Figure 2. Selected network architecture to achieve the segmentation task. Green boxes are residual blocks, blue boxes are simple convolutional layers with ReLU activation. Batch normalization is applied after convolution and before activation.

between one and two hours. We used Adam optimizer with a learning rate of 1.10^{-4} on batches of 10 images.

Weight initialization and pre-training

Considering the low amount of data available for training following the popular practice initiated in [21], we considered that pre-training the network on a large and publicly-available dataset would be advantageous. We therefore pre-trained our U-Nets to segment persons, the common objects in context (COCO) dataset [26]. We compared training experiments using randomly initialized weights or pre-training (Fig. 3). Although the final score was similar, the training converges faster using a pre-trained network,

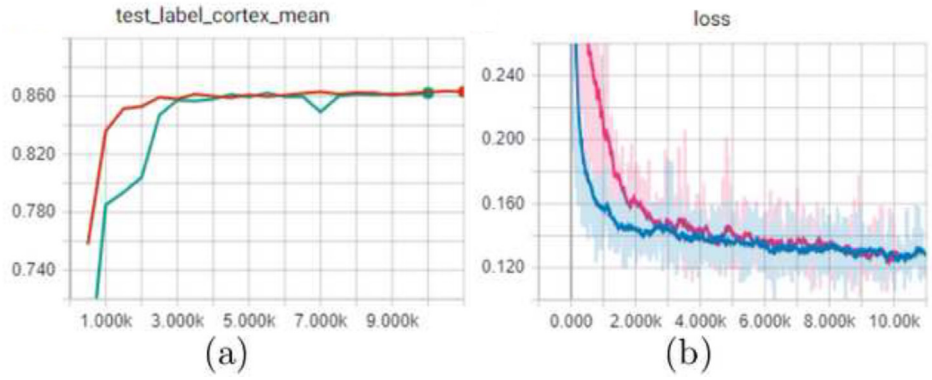

Figure 3. Impact of pre-training on the training procedure: a: evolution of the Dice score on the validation set during training (red is pre-trained, green is not); b: evolution of the binary cross-entropy on the training set (blue is pre-trained, pink is not). The *x*-axis represents the number of training steps.

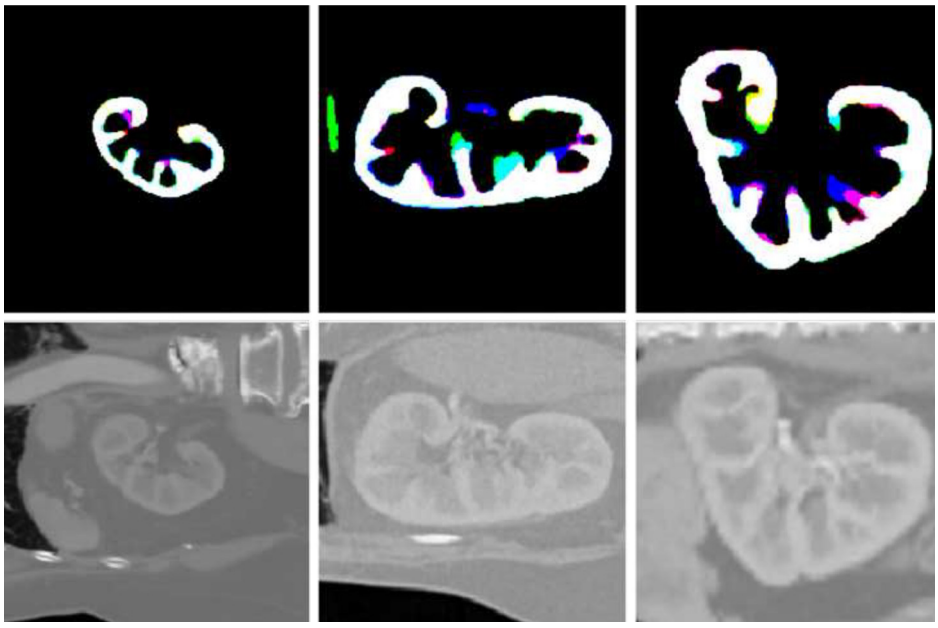

Figure 4. Top line: segmentation achieved by three networks trained on three different folds of the training database (each output is displayed on a different color channel, so that white represents a consensus for positively-labeled regions. We observe inconsistencies on the inner parts of the renal columns, and to a lesser extent on the outermost edge of the renal cortex). Bottom line: corresponding input CT images.

and was more stable overall. Therefore, we used pre-trained networks.

Post-processing and ensemble aggregation

We noticed that networks trained on different folds of the training database behave differently, especially on ambiguous pixels (Fig. 4). To improve the robustness and reduce the variability, we used ensemble aggregation.

We trained five networks on random folds of the training dataset, and two others on the complete training dataset. For each image at test time, we thus obtained seven segmentation masks taking pixel values in the interval ''[0,1]''. In each mask we only kept the largest connected component in order to remove obvious false positives (see, for instance Fig. 4, top middle: a blob is falsely labeled positively by one of the networks). Finally, we aggregated the results by taking the median value for each pixel, as it has shown to produce better results than the mean, by reducing the influence of extreme or outlier values.

Results

Participants were ranked using a Dice score: $S = \frac{2|P \cap T|}{|P|+|T|}$, where P is the predicted mask and T is the reference mask. We obtained a score of 0.867 on the test dataset and won the challenge by a narrow margin. The slight improvement obtained by the ensemble aggregation enabled us to win this challenge, as the second ranked team scored higher than our best network.

The segmentation results of our algorithm match the renal cortex with a good precision (Fig. 5). However, some of the flaws of the provided reference segmentations remain,

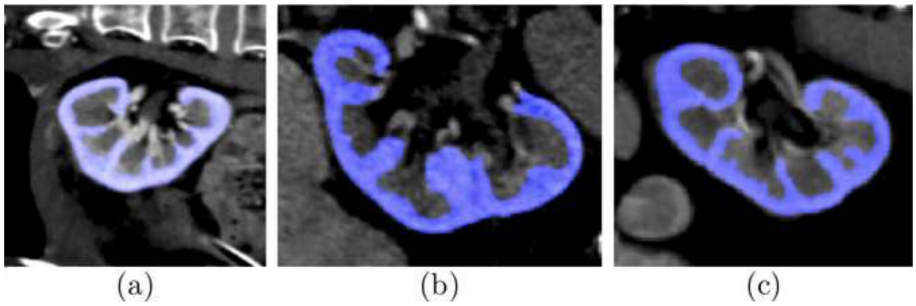

Figure 5. Illustration of automatic segmentation results obtained with the proposed approach (overlapped in blue on the input CT image); a: correct segmentation; b: cluster of renal columns; c: overextended segmentation.

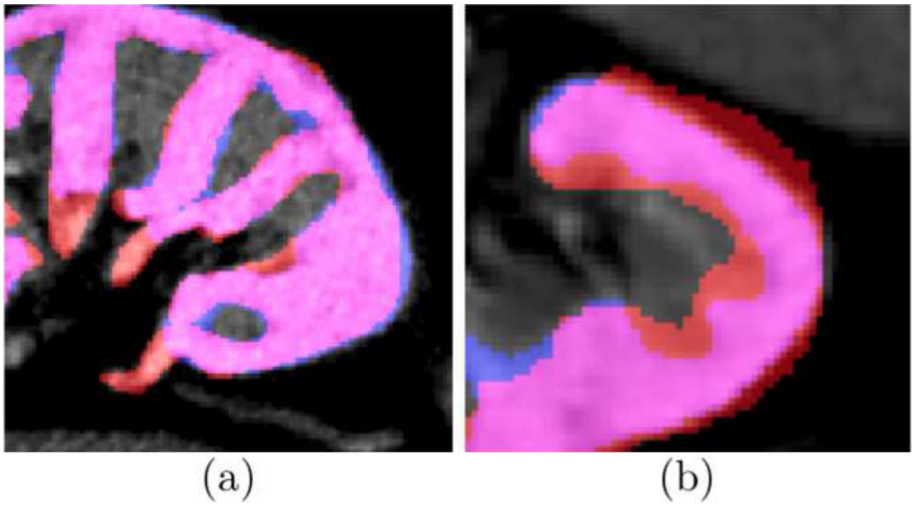

Figure 6. Illustration of test cases where the automatic segmentation results (blue) seem more accurate than the provided reference segmentation (red). Intersection in pink; a: vessels included in the reference mask but not in automatic segmentation result; b: reference segmentation obviously too wide.

such as the large clusters of renal columns, or when parts of the cortex are too widely segmented and join each other. Nonetheless, our algorithm seems to be less imprecise than the provided annotation, especially at the boundary of the cortex (Fig. 6).

Discussion

The state-of-the-art in image segmentation has improved greatly during the past five years, thanks to the progress accomplished in Deep Learning, to the point that some segmentation problems, which would have been considered a challenge ten years ago, now seem easy [27,28]. This is the case of renal cortex segmentation, where one can quickly achieve good results by training a UNet with any recent architecture found in the literature [18]. To the best of our knowledge, all the contestants chose a deep learning approach and the gap between participants was less than 0.03 Dice points.

The precision of the reference segmentations provided for this challenge seemed to set a low upper bound on the performance, as corroborated by the narrow gap between the first and second place (< 0.003 Dice points), and the gap between all the candidates (< 0.03 Dice points). As a consequence, the performance gain achieved by each of our algorithm details (image intensity scaling, data augmentation, pre-training, meta-parameter optimization, connected components analysis and ensemble aggregation) was difficult to quantify and barely significant if at all when considered alone, but enabled us, when added together, to improve the overall performance and win the challenge.

In conclusion, although 3D segmentation is useful clinically, the choice of 2D makes sense for a data challenge as it simplifies data collection, annotation, and storage [13,15—17]. Future research is needed to address the problem of renal cortex segmentation in 3D volumes.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

Funding

This work received funding from Association Nationale de la Recherche et de la Technologie (Contract 2018/2439)

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Credit author statement

Vincent Couteaux: conceptualization and design; data curation; writing-original draft preparation; review & editing.

Salim Si-Mohamed: conceptualization and design; data curation; supervision; resources; writing — original draft preparation; review & editing.

Raphaele Renard-Penna: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Olivier Nempont: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Thierry Lefevre: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Alexandre Popoff: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Guillaume Pizaine: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Nicolas Villain: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Isabelle Bloch: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Julien Behr: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Marie-France Bellin: conceptualization and design; resources; data curation.

Catherine Roy: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Olivier Rouviere: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Sarah Montagne: conceptualization and design; resources; data curation.

Nathalie Lassau: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Anne Cotten: conceptualization and design; data curation; resources; review & editing.

Loïc Boussel: conceptualization and design; supervision; writing — original draft preparation; review & editing.

Disclosure of interest

The authors declare that they have no competing interest.

References

[1] van den Dool SW, Wasser MN, de Fijter JW, Hoekstra J, van der Geest RJ. Functional renal volume: quantitative analysis at gadolinium-enhanced MR angiography–feasibility study in healthy potential kidney donors. Radiology 2005;236:189—95.

[2] Gandy SJ, Armoogum K, Nicholas RS, McLeay TB, Houston JG. A clinical MRI investigation of the relationship between kidney volume measurements and renal function in patients with renovascular disease. Br J Radiol 2007;80:12—20.

[3] Grantham JJ, Torres VE, Chapman AB, Guay-Woodford LM, Bae KT, King Jr BF, et al. Volume progression in polycystic kidney disease. N Engl J Med 2006;354:2122—30.

[4] Chen X, Summers RM, Cho M, Bagci U, Yao J. An automatic method for renal cortex segmentation on CT images: evaluation on kidney donors. Acad Radiol 2012;19:562—70.

[5] Halleck F, Diederichs G, Koehlitz T, Slowinski T, Engelken F, Liefeldt L, et al. Volume matters: CT-based renal cortex volume measurement in the evaluation of living kidney donors. Transpl Int 2013;26:1208—16.

[6] Jin C, Shi F, Xiang D, Jiang X, Zhang B, Wang X, et al. 3D fast automatic segmentation of kidney based on modified AAM and random forest. Trans Med Imaging 2016;35:1395—407.

[7] Pohle R, Toennies KD. A new approach for model-based adaptive region growing in medical image analysis. Computer Analysis of Images and Patterns Springer 2001;2124:238—46.

[8] Torimoto I, Takebayashi S, Sekikawa Z, Teranishi J, Uchida K, Inoue T. Renal perfusional cortex volume for arterial input function measured by semiautomatic segmentation technique using MDCT angiographic data with 0.5-mm collimation. AJR Am J Roentgenol 2015;204:98—104.

[9] Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 2017;30:449—59.

[10] Avendi MR, Kheradvar A, Jafarkhani H. Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach. Magn Reson Med 2017;78:2439—48.

[11] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640—51.

[12] Chen Y, Shi B, Wang Z, Zhang P, Smith CD, Liu. J.Hippocampus segmentation through multi-view ensemble ConvNets. 2017. p. 192—6.

[13] P.F. Christ, F. Ettlinger, F. Grün, M.E.A. Elshaera, J. Lipkova, S. Schlecht, et al. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. https://arxiv.org/abs/1702.05970 [Accessed on March 20, 2019].

[14] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse

annotation. Medical Image Computing and Computer-Assisted Intervention. MICCAI, 9901. Cham: Springer; 2016 [Lecture Notes in Computer Science].
[15] Dong H, Yang G, Liu F, Mo Y, Guo Y. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In: Valdés Hernández M, González-Castro V, editors. Medical Image Understanding and Analysis. MIUA. Communications in Computer and Information Science, 723. Cham: Springer; 2017.
[16] Erden B, Gamboa N, Wood S. 3D convolutional neural network for brain tumor segmentation. Computer Science. Stanford University; 2017 http://cs231n.stanford.edu/reports/2017/pdfs/526.pdf.
[17] F. Milletari, N. Navab, SA. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 3D Vision. IEEE 2016:565-71 [Accessed on March 20, 2019].
[18] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention; 2015. p. 234—41.
[19] Bertrand H, Ardon R, Perrot M, Bloch I. Hyperparameter optimization of deep neural networks: combining hyperband with bayesian model selection. France: CAP; 2017.
[20] Bertrand H, Perrot M, Ardon R, Bloch I. Classification of MRI data using deep learning and Gaussian process-based model selection. Biomedical Imaging. IEEE; 2017. p. 745—8.
[21] Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. Computer Vision and Pattern Recognition. IEEE; 2014. p. 1717—24.
[22] Rokach L. Ensemble-based classifiers. Artificial Intelligence Review 2009;33:1—39.
[23] Marmanis D, Wegner JD, Galliani S, Schindler K, Datcu M, Stilla U. Semantic segmentation of aerial images with an ensemble of CNNs, ISPRS Annals of the Photogrammetry. Remote Sens Spatial Info Sci 2016;III:473—80.
[24] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Computer Vision and Pattern Recognition. IEEE; 2016. p. 770—8.
[25] Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters improve semantic segmentation by global convolutional network. Computer Vision and Pattern Recognition. IEEE; 2017. p. 1743—51.
[26] Lin TY, Maire M, Belongie SJ, Bourdev LD, Girshick RB, Hays J, et al. Microsoft COCO: common objects in context. European Conference on Computer Vision; 2014. p. 740—55.
[27] Garcia-Garcia A, Orts S, Oprea S, Villena-Martinez V, Rodríguez JG. A review on deep learning techniques applied to semantic segmentation. Computer Vision and Pattern Recognition. Cornell University; 2017.
[28] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436—44.