Proceedings of Machine Learning Research 102:447–456, 2019 MIDL 2019 – Full paper track

3D multirater RCNN for multimodal multiclass detection and characterisation of extremely small objects

Carole H. Sudre[1,2,3] CAROLE.SUDRE@KCL.AC.UK
Beatriz Gomez Anson[4] BGOMEZA@SANTPAU.CAT
Silvia Ingala[5] S.INGALA@VUMC.NL
Chris D. Lane[2] C.LANE@UCL.AC.UK
Daniel Jimenez[2] D.JIMENEZ@UCL.AC.UK
Lukas Haider[6] L.HAIDER@UCL.AC.UK
Thomas Varsavsky[1,3] THOMAS.VARSAVSKY@KCL.AC.UK
Lorna Smith[7] LORNA.SMITH@UCL.AC.UK
Sébastien Ourselin[1] SEBASTIEN.OURSELIN@KCL.AC.UK
Rolf H Jäger[8] R.JAGER@UCL.AC.UK
M. Jorge Cardoso[1,2,3] M.JORGE.CARDOSO@KCL.AC.UK
[1] *School of Biomedical Engineering and Imaging Sciences, King's College London, UK*
[2] *Dementia Research Centre, UCL Institute of Neurology, UK*
[3] *Department of Medical Physics and Biomedical Engineering, University College London, UK*
[4] *Santa Creu i Sant Pau Hospital, Universitat Autonòma Barcelona, Barcelona, Spain*
[5] *Vrije University Medical Centre Amsterdam, The Netherlands*
[6] *Queen Square Multiple Sclerosis Centre, UCL Institute of Neurology, London, UK*
[7] *Cardiometabolic Phenotyping Group, Institute of Cardiovascular Science, UCL, London, UK*
[8] *Brain Repair and Rehabilitation Group, Institute of Neurology, UCL, London, UK*

Abstract

Extremely small objects (ESO) have become observable on clinical routine magnetic resonance imaging acquisitions, thanks to a reduction in acquisition time at higher resolution. Despite their small size (usually $<$10 voxels per object for an image of more than 10^6 voxels), these markers reflect tissue damage and need to be accounted for to investigate the complete phenotype of complex pathological pathways. In addition to their very small size, variability in shape and appearance leads to high labelling variability across human raters, resulting in a very noisy gold standard. Such objects are notably present in the context of cerebral small vessel disease where enlarged perivascular spaces and lacunes, commonly observed in the ageing population, are thought to be associated with acceleration of cognitive decline and risk of dementia onset. In this work, we redesign the RCNN model to scale to 3D data, and to jointly detect and characterise these important markers of age-related neurovascular changes. We also propose training strategies enforcing the detection of extremely small objects, ensuring a tractable and stable training process.

1. Introduction

The vascular network that supplies the brain changes with age, inducing alterations to surrounding tissue. Macroscopic changes, hallmark of cerebral small vessel disease, can be observed on structural MR images and include white matter hyperintensities, lacunar infarcts, cerebral micro-

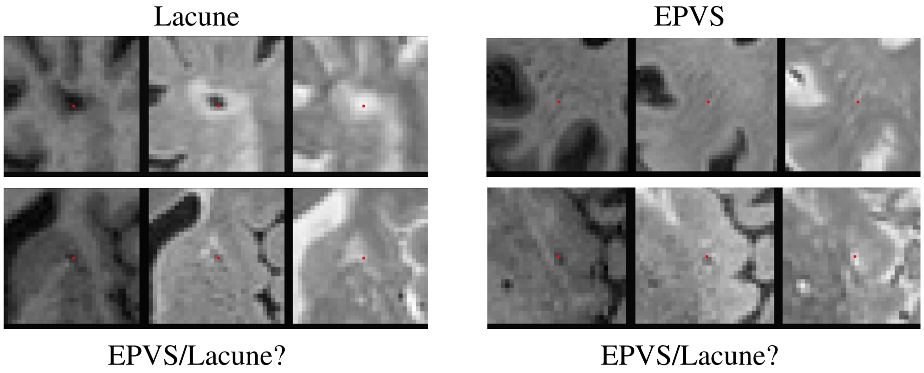

Figure 1: Examples of EPVS and lacunes on which agreement was high (top row) or low (bottom row) on the three structural modalities of interest (T1, FLAIR, T2). The red dot indicates the centre of mass of the object of interest.

haemorrhages and enlarged perivascular spaces (EPVS), among others (Wardlaw et al., 2013). More specifically, perivascular spaces are thought to be used as a lymphatic pathway in a drainage mechanism, where entrapped fluid can extend this space, making it visible in MR images, often as linearly-shaped fluid-like structure (Figure 1 top right). In clinical practice, their presence is classically assessed using visual scales on T2 MR images, described as elongated bright ellipsoids (Potter et al., 2015). The use of such visual scales requires extensive training and expertise, is prone to inter/intra rater variability, suffers from flooring/ceiling effects and is time-consuming for the operator. Some works have recently been proposed to automatically assess the EPVS burden (Boespflug et al., 2018)(Dubost et al., 2019) in clinical grade MR data, while others propose to segment EPVS at higher field (7T) (Zhang et al., 2016). In contrast, lacunar infarcts, observed with a much lower frequency, are areas of dead tissue due to complete ischemia. Their shape signature is an ovoid object of $3-15mm$ of diameter, with a cerebrospinal fluid (CSF) -like intensity in the centre. Often, on T2 weighted Fluid attenuated inversion recovery (FLAIR) images, they are surrounded by a rim of hyperintensity (see Figure 1 top left). In practice, even for trained radiologists, distinguishing between EPVS and lacunes can be very challenging (see Figure 1 bottom). This results in double counting of uncertain objects (del C. Valdés Hernández et al., 2013), and under-counting when objects branch from the same point. This task is however of clinical importance as these markers reflect tissue damage and are key to understand complex pathological pathways of age-related vascular changes (Ramirez et al., 2016).

To account for the above-mentioned challenges, we propose to adapt the 2D RCNN model presented by He et al (He et al., 2017) that allows for multiclass multi-instances simultaneous detection and segmentation to multirater 3D data, in the context of EPVS and lacune detection and size characterisation, with the perspective of a future expansion to more object classes (e.g.white matter hyperintensities) and their semantic segmentation. After a brief description of the 2D RCNN framework, we detail the challenges inherent to 3D data of such a framework in the capture of extremely small objects, and describe the introduction of multirater predictions.

2. Methods

2.1. Two dimensional RCNN

In the original RCNN framework, a backbone network is trained to extract generic features. This initial training is then complemented by two stages: a region proposal network and a final classification network applied to selected boxes whose shapes have been modified to fit a specified mask. In the 2D setting, the region proposal network is based on the classification as positive or negative of a series of predefined boxes created based on anchors, regularly spaced on the 2D grid with different ratios of height and width. All selected grid are then resampled (pooled) to a user-specified shape and fed to the final segmentation classification branch of the framework.

2.2. Challenges and strategies for a multirater 3D extension

The main challenges related to the extension of the successful RCNN framework to 3D data lay in the memory and data requirements, as well as an extreme class imbalance. In terms of memory, the generation of grid anchors become notably prohibitive in 3D. Additionally, when dealing with ESOs, any interpolation induced by the region pooling may obscure relevant features and render the segmentation meaningless. In order to account for these challenges, the following strategies were adopted at the different stages of the framework:

Backbone network The 3D HighResNet proposed by Li et al. (Li et al., 2017) was used as backbone network to extract features. This architecture has a large contextual field of view at reduced parameter cost. This network uses three levels of residual convolutional networks with dilated convolutions with increasing dilation factor, each level consisting of three dilated convolutions with fixed dilation factor alternating with batch normalisation and ReLu activation. In the presented setting, the network was applied to regress a distance map with a root mean square error loss. The distance map is calculated from each given element's segmentation.

Region Proposal Network (RPN) In order to alleviate the memory burden of having to explicitly describe anchors and associated boxes, the RPN, consisting of one classification and one regression branch, was applied in a convolutional fashion to every voxel. The features extracted at the backbone level were fed into a small convolutional network with a single common 3^3 kernel, followed by either a classification layer or a regression layer. The classification layer establishes if the centre of the patch is likely to be the centre of mass of the target object, while the regression part outputs four values: three values representing the distance to the closest object centre of mass, and the fourth representing the scale of the targeted object. Classification and regression were learnt from 300 samples on the patch, with a 50/50 balance between positive and negative samples. To avoid any impact on the regression branch, negative samples did not bear any weight on the regression loss. A cross-entropy loss was used for the classification branch while a smooth distance loss was applied on the regression branch for the estimation of the distance to the closest element centre of mass. Denoting r_n the absolute error between predicted value and ground truth for a given sample n, the smooth distance loss DL is expressed as:

$$DL = \frac{1}{N}\sum_{n=1}^{N} f(r_n) \text{ where } f(r_n) = \begin{cases} 0.5r_n^2 & \text{if } r_n < 0.5 \\ (r_n - 0.125)^2 - 2 & \text{if } r_n > 2.125 \\ r_n - 0.125 & \text{otherwise} \end{cases}$$

Refinement/Classification Network (RCN) From the location of proposed ESO centres-of-mass, boxes were associated with ground-truth objects, and extracted masks are directly fed so as to classify the boxes and adjust the regression of the centre of mass.

The branch jointly classifying the element and regressing centre of mass and object scale consisted of a convolutional layer of kernel size 7, followed by a fully connected layer. After average pooling, classification and bounding box regression were established. For the regression branch, the target prediction was the residual between the RPN prediction and the ground truth for the three location elements, and a scale correction factor for the size. A similar smooth distance loss was applied as a cost function. In contrast to the original RCNN framework, selected boxes were neither resized nor pooled to a predefined shape. This is in order to avoid interpolation that would be detrimental, given that many of the targeted elements are one voxel wide.

Multirater encoding For each of the manually-segmented elements, the raters were asked to attribute one of the following class: 1)Nothing; 2) Lacune; 3) EPVS; 4) Undecided between lacune and EPVS; Instead of a crisp classification, a soft probability label was obtained as the average of the multiple raters involved in the classification and used as target. For each rater, a fully connected layer was added in order to directly infer the classification of each individual. The architecture framework is displayed in Figure 2.

2.3. Implementation

Sampling and data normalisation The existence of two types of imbalance (foreground vs background, and between EPVS vs lacunes) required a purpose-specific sampling scheme. A probabilistic weight sampling was adopted as suggested by Ronneberger et al (Ronneberger et al., 2015) to extract patches of size 64^3 over the images. For this purpose, the inverse of the distance maps from segmented EPVS and lacunes were smoothed and linearly combined using a ratio of 1/100 reflecting the relative frequency of occurrence of these two classes. These maps were clipped to

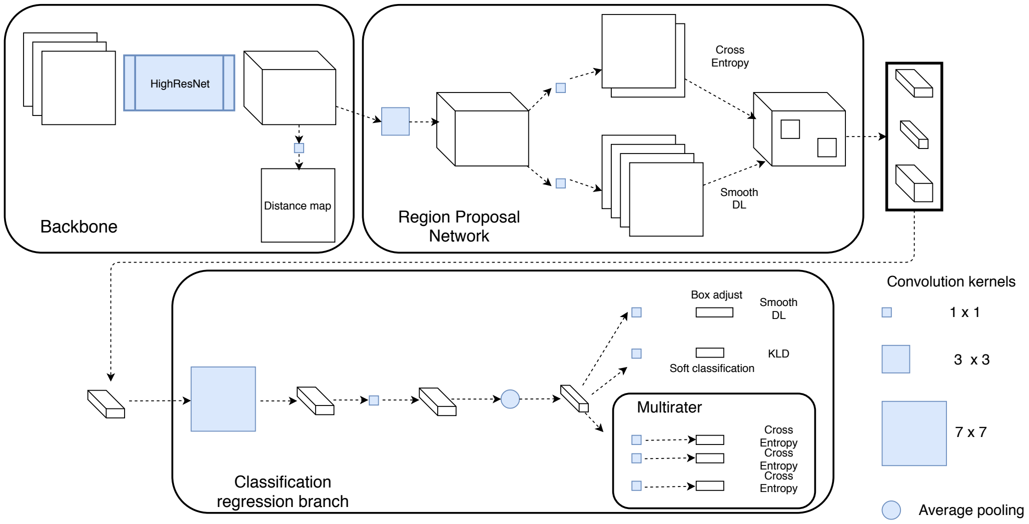

Figure 2: Architecture of the 3D multirater RCNN for extremely small objects.

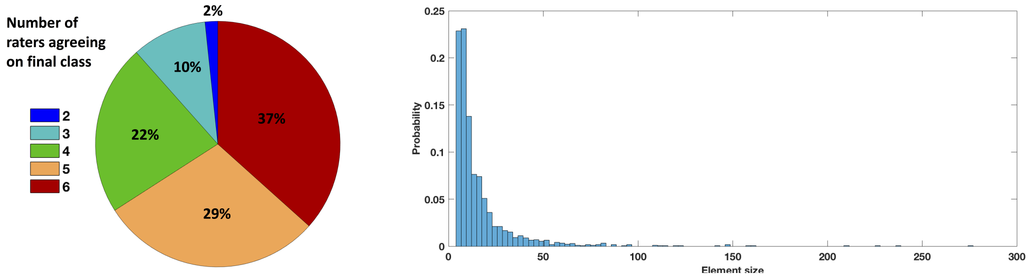

Figure 3: Repartition of agreement between raters responsible for the final crisp classification (left) and distribution of the size of the targeted elements (right).

a minimum of 10^{-5} to reflect the overall background/foreground ratio. All input data (T1, T2 and FLAIR images) was bias field corrected, skull stripped, and then z-scored to the white matter region statistics, segmented independently (Sudre et al., 2015).

Training scheduling and loss functions The framework was implemented within NiftyNet (Gibson et al., 2018) (`niftynet.io`) and will be merged into the main codebase at the time of publication. The network was trained progressively per stage to mitigate training stability issues. Sections where classification and regression were combined (RPN and RCN) were trained in two steps: the first one consisted of the classification training with a sigmoid applied to the regression loss, and the second step was the sum of the two losses. Each of the steps was trained for 1000 iterations with learning rate of 0.0001. In order to account for scale differences observed across combined loss functions, notably between classification and positioning regression losses, empirical weights were chosen and progressively modified throughout the training of the network in order to always ensure a balance between classification accuracy and box positioning.

Inference At inference, a similar patch size was used as for the training step in order to expect a similar number of proposals (limited to 300). In order to prune the potential positions of centre of mass, the information from the score map and the distance map were combined. The score map was thresholded at 0.25 and the morphological skeleton of the underlying distance map were extracted. The corresponding distance score maxima were then taken as potential proposed centres of mass. Centres of mass closer than 2mm were pruned as a form of non-maximum suppression.

3. Data and experiments

3.1. Data

16 subjects were selected out of a longitudinal tri-ethnic cohort of elderly subjects aiming at investigating the relationship between cardiovascular risk factors and brain health (Tillin et al., 2012). At the third wave of investigation, subjects of this cohort underwent an MR session including the acquisition of 3D 1mm^3 isotropic T1 weighted, T2 weighted and T2-weighted FLAIR images (Sudre et al., 2018). The 16 subjects were chosen for their elevated vascular burden visually assessed by a trained radiologist. EPVS and lacunes were manually segmented on the three available structural MR sequences using ITKSnap (Yushkevich et al., 2006). Performed by a rater accustomed to the use of the segmentation software, the delineation of EPVS and lacunes for a single subject required

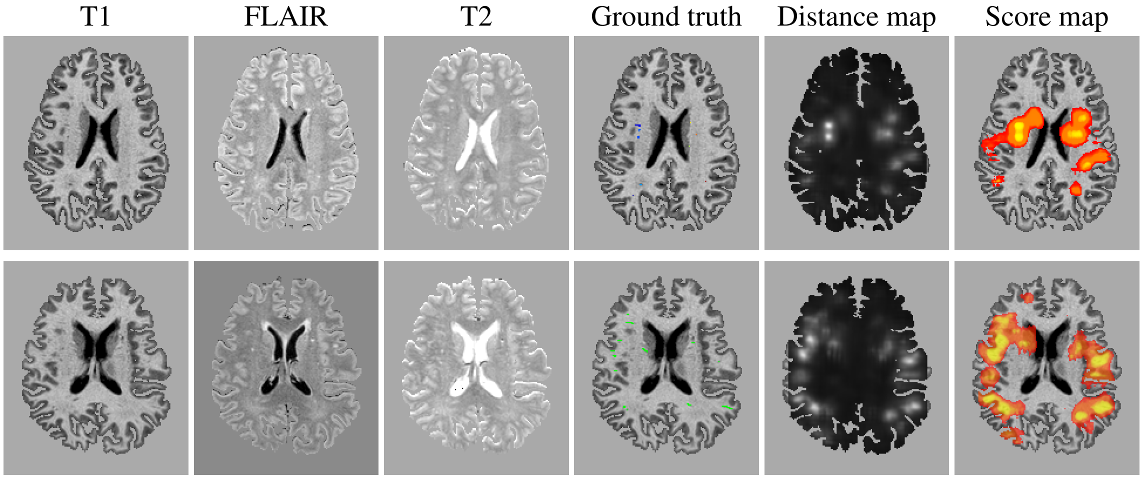

Figure 4: Two holdout cases with the three input channels (T1, FLAIR, T2), gold standard segmentation, inferred distance maps and score map.

an average of 20h. Segmentations were done in a multi-view manner to ensure geometrical consistency, with all images aligned to the T1 sequence as a geometrical reference. Segmentation masks were then automatically corrected and voxels with inappropriate signal identity signature were removed. Individual EPVS and lacunes were further classified at the level of connected components by six operators with a varied range of expertise using an in house dedicated viewer. Only elements with a volume of more than 5 voxels were considered in this work, resulting in a database of 2442 elements. The volumes of segmented elements ranged thus from 5 to 350, with 48.8% with a size below 10 voxels. Perfect agreement among raters was reached only in 36.6% of the cases, and only 2.8% of the elements were ultimately classified as lacunes. Figure 3 presents an histogram of element size, and a pie chart representing the proportion or rater agreement. The poor inter-rater classification agreement hints at the complexity of the task. Uncertainty over the segmentation would have to be evaluated over multiple raters before envisioning moving the proposed object RCNN detection model to a full Mask-RCNN, also performing segmentation. Due to the lack of more training data, 14 of the subjects were used for training and 2 were hold-out for testing.

3.2. Experiments

In order to compare the performance of a standard segmentation approach to the proposed multi-class detection framework, we trained semantic segmentation models with multiple combinations of architectures, loss functions (e.g Generalised Dice Loss), learning rates (from 10^{-6} to 10^{-3}) and regularisation. Parameter choice was similar to the one used for the backbone network, with ranges that have been shown to perform well on unbalanced data. Unfortunately, no network was able to segment any foreground class.

We present hereafter the results obtained at the different stages of the model in terms of distance regression, score map, RPN and multirater classification.

4. Results

Each step of the framework was assessed on the two held out test subjects using the same metrics as the loss functions. Figure 4 presents the input data for the three modalities along with the ground truth segmentation, the regressed distance map and the inferred score map.

Interestingly, some elements not present in the gold standard segmentation but detected as per the score map were a posteriori considered as valid enlarged perivascular spaces as can be seen on Figure 5.

Given the limitations of the available gold standard in terms of inter-rater element classification, and potential missing objects, the validation focused on the sensitivity of the trained model and the relationship of the results with the multi-rater uncertainty. A sensitivity of 72.7% was observed across the two test subjects with a significant difference in element size between false negatives and true positives (Wilcoxon ranksum test $p<0.00001$). Investigating the relationship between the ratio of overlap between best matching detected box and ground truth proposal, a significant association between agreement of raters and overlap was observed ($p=0.002$) with a median overlap of 59% when all raters agreed and an overlap of 30% for the more uncertain cases (at least one rater considering the element not to be relevant). Note that overlap is measured on the predicted box, which can vary widely in its size. Figure 6 presents boxplots of relationship between ESO scale and detection (left), and overlap ratio with rater uncertainty (right).

Figure 7 presents the ground truth and matching predicted boxes where the color reflects the probability of belonging to each of the classes (nothing - lacune - EPVS - undecided).

5. Discussion and conclusion

In this work we proposed a 3D deep learning model for the detection and characterisation of extremely small objects incorporating multi-rater labels and agreement. In this context, two types of extreme class imbalance were found, with a very low ratio of foreground to background, as well as a strong imbalance between the estimated classes where the prevalence of enlarged perivascular spaces being much higher than the number of lacunes.

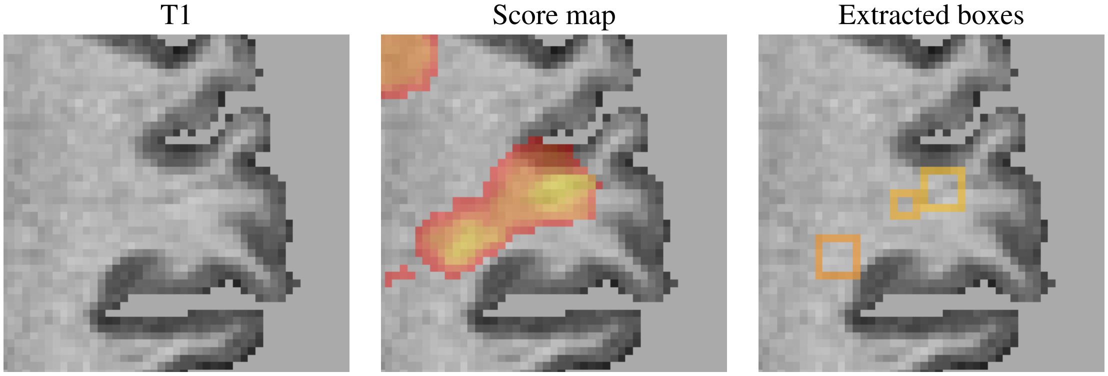

Figure 5: ESOs rightly detected by the network but missed during manual labelling. From left to right, T1, predicted score map and predicted boxes.

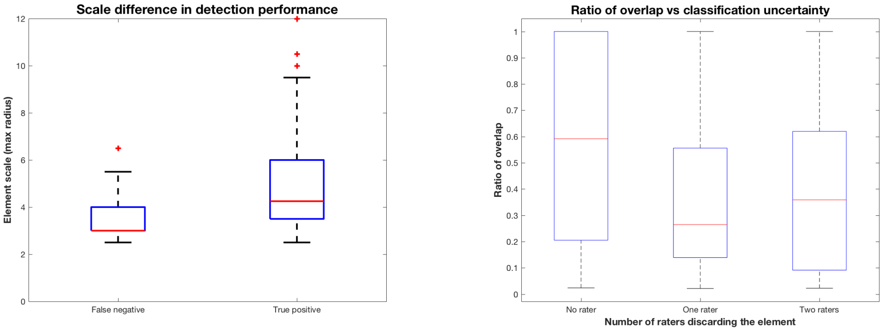

Figure 6: Left: Gold standard scale of ESO versus detection (False negative/True positive). Right: Relationship between multi-rater disagreement and box overlap performance. Note that overlap ratio is higher for more certain objects.

The different steps of the framework were evaluated, showing a good sensitivity of the region proposal network. Specificity was not ideal, probably limited by the missing annotation of individual branching elements (currently considered as a single ESO). Future work will use the multi-rater gold standard to better guide network updates by penalising classification errors made on definite classifications more strongly. Additionally, the segmentation, currently only used to obtain the original distance map, could enrich the model by defining a soft labelling at the edges and/or obtaining additional manual ratings. Furthermore, it must be noted that the training accuracy heavily depends on the quality of the initial co-registration of the different modalities, as one voxel of shift may lead to an aberrant intensity signature. At this stage, proposal boxes are cuboid, since a single scale factor is regressed at training. Future work will also involve transforming the scale regression of the

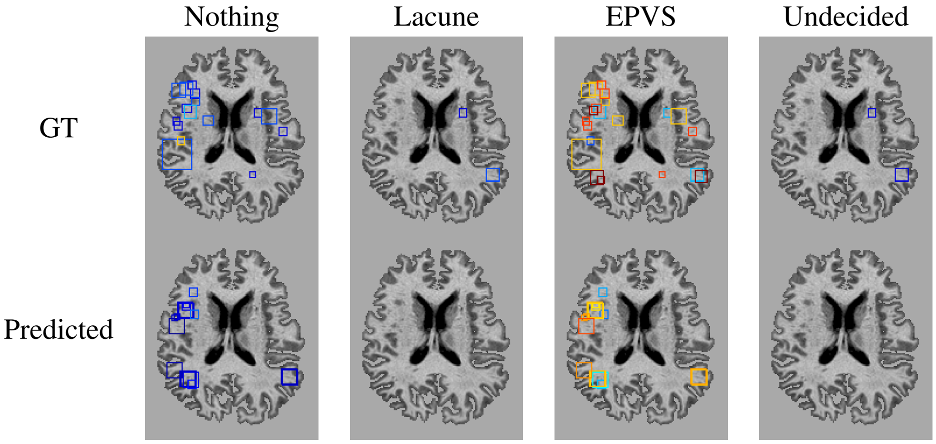

Figure 7: Probabilistic ground truth (GT) and predicted boxes for the different classes. All blue boxes correspond to low classification probabilities (p<0.5), and illustrate rating variability. Yellow to red boxes represent probabilities ranging from 0.5 to 1, and represent confident ESOs classifications.

RCNN into a multi direction scale factor transformation thus providing further information on the shape of the enclosed object.

Acknowledgments

We are extremely grateful to all the participants of the SABRE study, and past and present members of the SABRE team. This work was supported by an Alzheimer's Society Junior Fellowship (AS-JF-17-011), the Wellcome/EPSRC Centre for Medical Engineering [WT 203148/Z/16/Z], IMI2 grant AMYPAD [115952], the MSCA-ITN-Demo [721820], and the Wellcome Flagship Programme in High-Dimensional Neurology. The SABRE study was funded at baseline by the Medical Research Council, Diabetes UK, and the British Heart Foundation. At follow-up, the study was funded by the Wellcome Trust (067100, 37055891 and 086676/7/08/Z), the British Heart Foundation (PG/06/145, PG/08/103/ 26133, PG/12/29/29497 and CS/13/1/30327) and Diabetes UK (13/0004774). We gratefully acknowledge NVIDIA corporation for the donation of a GPU Tesla K40 that was used in the preparation of this work.

References

Erin L. Boespflug, Daniel L. Schwartz, David Lahna, Jeffrey Pollock, Jeffrey J. Iliff, Jeffrey A. Kaye, William Rooney, and Lisa C. Silbert. MR Imaging–based Multimodal Autoidentification of Perivascular Spaces (mMAPS): Automated Morphologic Segmentation of Enlarged Perivascular Spaces at Clinical Field Strength. *Radiology*, 286(2):632–642, feb 2018.

Maria del C. Valdés Hernández, Rory J. Piper, Xin Wang, Ian J. Deary, and Joanna M. Wardlaw. Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: A systematic review. *Journal of Magnetic Resonance Imaging*, 38(4): 774–785, oct 2013.

Florian Dubost, Hieab Adams, Gerda Bortsova, M Arfan Ikram, Wiro Niessen, Meike Vernooij, and Marleen de Bruijne. 3D regression neural network for the quantification of enlarged perivascular spaces in brain MRI. *Medical image analysis*, 51:89–100, jan 2019.

Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. Niftynet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158:113–122, 2018.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M Jorge Cardoso, and Tom Vercauteren. On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. In *International Conference on Information Processing in Medical Imaging (IPMI)*, 2017.

Gillian M Potter, Francesca M Chappell, Zoe Morris, and Joanna M Wardlaw. Cerebral perivascular spaces visible on magnetic resonance imaging: development of a qualitative rating scale and its observer reliability. *Cerebrovascular Diseases*, 39(3-4):224–231, jan 2015.

Joel Ramirez, Courtney Berezuk, Alicia A McNeely, Fuqiang Gao, JoAnne McLaurin, and Sandra E Black. Imaging the Perivascular Space as a Potential Biomarker of Neurovascular and Neurodegenerative Diseases. *Cellular and molecular neurobiology*, mar 2016. ISSN 1573-6830. doi: 10.1007/s10571-016-0343-6.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Carole Sudre, M Jorge Cardoso, Willem Bouvy, Geert Biessels, Josephine Barnes, and Sébastien Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Transactions on Medical Imaging*, 34(10):2079–2102, apr 2015. ISSN 1558-254X. doi: 10.1109/TMI.2015.2419072.

Carole H. Sudre, Lorna Smith, David Atkinson, Nish Chaturvedi, Sébastien Ourselin, Frederik Barkhof, Alun D. Hughes, H. Rolf Jäger, and M. Jorge Cardoso. Cardiovascular Risk Factors and White Matter Hyperintensities: Difference in Susceptibility in South Asians Compared With Europeans. *Journal of the American Heart Association*, 7(21), nov 2018.

Therese Tillin, Nita G Forouhi, Paul M McKeigue, Nish for the SABRE group Chatuverdi, and Nish Chaturvedi. Southall And Brent REvisited: cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. *International Journal of Epidemiology*, 41(1):33–42, feb 2012.

Joanna M Wardlaw, Eric E Smith, G J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, Sandra E Black, Carol Brayne, Monique M B Breteler, Hugues Chabriat, Charles DeCarli, Frank-Erik de Leeuw, Fergus Doubal, Marco Duering, Nick C Fox, Steven Greenberg, Vladimir Hachinski, Ingo Kilimann, Vincent Mok, Robert van Oostenbrugge, Leonardo Pantoni, Oliver Speck, Blossom C M Stephan, Stefan Teipel, Viswanathan Anand, David Werring, Christopher Chen, Colin Smith, Mark A van Buchem, Bo Norrving, Philip B Gorelick, and Martin Dichgans. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurology*, 12:822–838, 2013.

Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

Jun Zhang, Yaozong Gao, Sang Hyun Park, Xiaopeng Zong, Weili Lin, and Dinggang Shen. Segmentation of Perivascular Spaces Using Vascular Features and Structured Random Forest from 7T MR Image. *Machine learning in medical imaging. MLMI (Workshop)*, 10019:61–68, oct 2016.