Artificial Intelligence with Light Supervision

Application to Neuroimaging

Florian Dubost

Acknowledgments:
This research was funded by The Netherlands Organisation for Health Research and Development (ZonMw) Project104003005.

For financial support for the publication of this thesis, the following organisations are gratefully acknowledged: the department of Radiology of Erasmus MC and Quantib BV.

ISBN: 978-94-6375-865-9
Cover: Florian Dubost
Layout: Florian Dubost
Print: Ridderprint | www.ridderprint.nl

Artificial Intelligence with Light Supervision
Application to Neuroimaging

Licht Gecontroleerde Kunstmatige Intelligentie
Toepassing in Beeldvormend Hersenonderzoek

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Friday 8 May 2020 at 13.30 hrs

by

Florian Pierre Guy Dubost
born in Clamart, France

Erasmus University Rotterdam

Doctoral Committee:

Promotors:	Prof.dr. M. de Bruijne Prof.dr M.W. Vernooij Prof.dr. W.J. Niessen
Other members:	Dr. C. Sánchez Gutiérrez Prof.dr. A. van der Lugt Dr.ing. S. Oeltze-Jafra

Pour Viviane et Sylvain

Contents

Part A

Introduction

1 Manuscripts in this thesis

- **Chapter 1: Dubost, F.**, Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. 3D Regression Neural Network for the Quantification of Enlarged Perivascular Spaces in Brain MRI. Medical Image Analysis. 2019.

- **Chapter 2: Dubost, F.**, Bortsova, G., Adams, H., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Hydranet: Data Augmentation for Regression Neural Networks. MICCAI 2019.

- **Chapter 3: Dubost, F.**, Adams, H., Yilmaz, P., Bortsova, G., van Tulder, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Weakly Supervised Object Detection with 2D and 3D Regression Neural Networks. Submitted.

- **Chapter 4:** van Wijnen K.*, **Dubost, F*.**, Yilmaz, P., Ikram, M.A., Niessen W., Adams, H., Vernooij, M. and de Bruijne, M. Automated Lesion Detection by Regressing Intensity-Based Distance with a Neural Network. MICCAI 2019.

- **Chapter 5: Dubost, F.**, Yilmaz, P., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Enlarged Perivascular Spaces in Brain MRI: Automated Quantification in four Regions. Neuroimage. 2019.

- **Chapter 6: Dubost, F.***, Duennwald M.*, Scheumann, V., Schreiber, F., Huff, D., Vernooij, M., Niessen, W., Skalej, M., Schreiber, S., Oeltze-Jafra, S.**, de Bruijne M**. Automated Quantification of Enlarged Perivascular Spaces in Clinical Brain MRI across Sites. MICCAI workshop MLCN 2019.

- **Chapter 7:** Venkatraghavan, V.*, **Dubost, F.***, Bron, E.E., Niessen, W.J., de Bruijne, M. and Klein, S., 2019. Event-Based Modeling with High-Dimensional Imaging Biomarkers for Estimating Spatial Progression of Dementia. IPMI 2019.

- **Chapter 8: Dubost, F.**, de Bruijne, M., Nardin, M.J., Dalca, A.V., Donahue, K.L., Giese, A., Etherton, M.R., Wu, O., de Groot, M., Niessen, W., Vernooij, M.W., Rost, N.S., and Schirmer, M.D. Automated image registration quality assessment utilizing deep-learning based ventricle extraction in clinical data. Submitted.

* and ** indicate equal contribution

2 Introduction

Machine learning methods are statistical models that are optimized based on example data to find patterns in data. Thanks to their ability to handle high dimensional data with complex non-linear relationships between input and output variables, machine learning methods are especially suited to deal with the explosive growth of digital data in society. Recent advances have permitted a category of machine learning methods to emerge as state-of-the-art methods for image processing: convolutional neural networks (LeCun et al., 2015). Whereas in traditional machine learning sets of features describing the input data had to be computed prior to the training of models, convolutional neural networks compute features internally as part of the training procedure. Expert knowledge is therefore not needed anymore to design relevant features, and neural networks can be trained using raw input data.

Neural networks are a very promising technique for medical image analysis (Shen et al., 2017), where the purpose is often to quantify an imaging biomarker. An imaging biomarker is an imaging characteristic that relates to the physiological state or disease status. To assess imaging biomarkers either in medical studies or in clinical practise, radiologists mostly assess scans visually. These assessments can be time-intensive and are prone to high intraobserver and interobserver variability. Automated methods have the potential to quantify target biomarkers in fractions of seconds and with a high reproducibility. Those methods can quantify biomarkers in large datasets where performing visual assessments would be impossible due to time and resource constraints. Association between the target biomarker and other clinical variables can subsequently be determined with standard statistical models, and thus also support discoveries in medicine. In clinical scenarios, the computed biomarker values can be used to assist doctors in diagnostic and prognostic assessment and for treatment choices.

Digital medical images can be seen as a grid of pixels or voxels each having an intensity value. Most often, medical image analysis algorithms are designed and optimized to make classifications on the pixel level. For example, the most recurrent task in medical image analysis research is segmentation (Hesamian et al., 2019), which consists in classifying pixels of images into categories, such as different types of tissues. Segmentations are used to assess volumetry and to support radiotherapy planning and image guided interventions. Segmentation also enables medical researchers or clinicians to compute shape features or perform texture analysis in a region, e.g. radiomics for tumor characterization (Zhou et al., 2018). In other words, quantitative biomarker values can be derived from the pixel-wise predictions. More rarely, prediction models are optimized on the image-level, and when they are, it is most often to solve image classification tasks such as healthy versus disease state. Only a few researchers have proposed to optimize neural networks to directly regress the value of target biomarkers. While training those networks raises technical challenges in terms of optimisation and interpretability, it removes the necessity to collect pixel-wise ground truths for the training. Acquiring annotations for large datasets is indeed a costly and long process, which can considerably slow down the research.

In this thesis, I study convolutional neural networks for medical image analysis, and more specifically for the analysis of magnetic resonance images (MRIs) of the human brain. Magnetic resonance imaging is an image acquisition method mostly used for inspecting living tissue (Moore et al., 2006). MRI exploits magnetic properties of the hydrogen atoms present mostly in water and fat, and is one of the most common non-invasive imaging techniques used by radiologists to guide their clinical diagnoses. MRI is safe for the scanned individual and provides high contrast in soft tissues such as the brain, which is perfect for neurology research.

The applications presented in this thesis revolve around cerebral small vessel disease (CSVD). CSVD is an umbrella term to describe multiple pathological processes affecting small vessels in the brain. These processes are thought to be involved with the occurrence of stroke (Selvarajah et al., 2009), dementia (Mills et al., 2007), multiple sclerosis (Achiron and Faibel, 2002), and cognitive decline (Uiterwijk et al., 2016). There are several established imaging markers of CSVD, including focal lesions such as white matter hyperintensities, lacunes, and cerebral microbleeds. Microinfarcts and enlarged perivascular spaces are an emerging biomarker for CSVD. Enlarged perivascular spaces are also thought to be related to sleep and glymphatic clearance (Brown et al., 2018; Mestre et al., 2017; Rasmussen et al., 2018).

In the brain, the perivascular space is the space between penetrating blood vessels and the envelope of the brain. Perivascular spaces are filled with interstitial fluid. Because of multiple hypothesized mechanisms such as hypertension, atrophy, inflammation or glymphatic clearance, these spaces can locally enlarge and become visible on 1.5T and 3T MRI scans. Enlarged perivascular spaces (PVS) can convey information on risk of disease. For example, several studies have investigated the presence of PVS as an emerging biomarker for various brain diseases such as dementia (Mills et al., 2007), stroke (Selvarajah et al., 2009), multiple sclerosis (Achiron and Faibel, 2002) and Parkinson (Zijlmans et al., 2004). However, quantifying PVS is challenging. The enlargement of perivascular spaces is not a binary process but a continuum, and the quantification of subtly enlarged perivascular spaces remains an open research question. The size of the smallest PVS can be close to the MRI voxel resolution, and because of partial volume effects, differentiating small PVS from noise can be intractable. This introduces substantial variability in the quantitative assessment of the PVS burden. In addition, PVS can be located in different regions in the brain and can be numerous. These quantification challenges have impeded the study of etiology and clinical implications of PVS. Until now, PVS burden has mostly been quantified using visual scales where the radiologist either counts PVS in a given brain region (Adams et al., 2015) or categorises this count (Potter et al., 2015b). Because of the inherent nature of the subtle enlargement of perivascular spaces and because of their high number and small size, delineating PVS contour in large datasets is too time-consuming for radiologists. Lacking pixel-wise ground truths, very few automated methods have been developed for the quantification of PVS burden. Methods that have been developed were based on traditional image processing techniques and often suffered from a relatively poor performance. Their evaluation has also been limited to small datasets or specific brain regions. Neural networks have the ability to exploit weakly labeled datasets, such as datasets with visual scores, to optimize the prediction model end-to-end and ultimately retrieve useful information from the imaging data.

In this thesis I propose to develop neural network methods with applications in 3D brain MRI biomarker quantification. I mostly focus on PVS quantification. More specifically, I developed neural networks to predict image-level labels such a lesion count or volume, networks for weakly supervised object detection, for brain registration, and for generation of artificial brain images to model disease progression. I evaluated my quantification methods in large (more than 2000 scans) research studies and clinical datasets. When information about intrarater and interrater variability was available, I empirically demonstrated that the proposed method could reach a performance similar to that of experts. Part B

(Chapters 1 and 2) and Part C (Chapters 3 and 4) describe the methodological aspects of the work. Part D (Chapters 5 and 6) focuses more on the application of PVS quantification for neurology research. Finally Part E (Chapters 7 and 8) combines both methodological and medical research applied to other neuroimaging tasks.

In **Part B**, I study neural networks optimized to regress image-level labels. In **Chapter 1**, I take the example of the quantification of PVS burden in the basal ganglia, and empirically demonstrate (a) that neural networks optimized to regress the count of PVS achieve better results than more traditional machine learning techniques also optimized with image-level labels, (b) that these networks achieve a performance in-between the intraobserver and interobserver agreements of experts raters, and (c) that these networks focus mostly on PVS, and not on other structures in the image that might be correlated to PVS count. In **Chapter 2**, I propose a method to optimize these networks for PVS count prediction with very small training datasets (25 images, with a single label per image) and empirically demonstrate that these networks can reach a performance similar to the interobserver agreement. The analysis was realised for the quantification of PVS count in the basal ganglia and white matter hyperintensity volume.

In **Part C**, I focus on object detection with neural networks. In **Chapter 3**, I propose a weakly supervised detection method for neural networks optimized with image-level labels. While the network is trained only with image-level labels representing a count (as presented in Part B), we can compute attention maps that reveal the focus of the network during inference. I demonstrate the potential of this method on a dataset of handwritten digits and on the detection of PVS in four different brain regions. I also compare the proposed method with other weakly supervised detection methods. In **Chapter 4**, we propose a detection method based on networks optimized to predict geodesic distance maps computed from dot annotations. Obtaining these dot annotations requires more work than obtaining visual scores. We evaluate the method for the detection of PVS in the centrum semiovale and obtain a detection that is closer to that of the annotator than what was achieved with the weakly supervised method presented in Chapter 3.

In **Part D**, I propose an automated method for the quantification of PVS which could be applied in medical research and clinical practise. The evaluation of this method is more medically focused than the evaluation of the method presented in Chapter 1. The method is applied to four brain regions: the midbrain, the hippocampi, the basal ganglia, and centrum semiovale. In **Chapter 5**, I validate this method in MRI scans from a population study: the Rotterdam scan study (Ikram et al., 2017). I demonstrate empirically that associations between 20 potential determinants of PVS and visual PVS scores and associations between the same determinants and the automated PVS scores are similar. In **Chapter 6**, we deploy the methods on the brain MRI images acquired from multiple scanners in the PACS system of the university hospital of Magdeburg in Germany, and obtained results similar to the interrater agreement in the centrum semiovale.

Neural networks were not only successful for the quantification of PVS. In **Part E**, I present neural network-based methods for other neuroimaging research questions such as disease progression modelling and image registration. In **Chapter 7**, we propose an event-based method that exploits high-dimensional voxel-wise imaging biomarkers. To validate the method, we develop a framework that simulates the temporal evolution of imaging biomarkers. The method is based on variational autoencoders (Kingma and Welling, 2014) and simulates neurodegeneration in individual brain regions. In **Chapter 8**, I propose a method for ventricle segmentation in clinical scans and evaluate it in an international multi-site dataset. I use this method to automatically assess registration quality and to build a multi-atlas registration framework that uses age-specific atlases to improve registration quality.

Part B

Regression of Image-level Labels

Chapter 1

3D Regression Neural Network for the Quantification of Enlarged Perivascular Spaces in Brain MRI

Abstract

Enlarged perivascular spaces (PVS) in the brain are an emerging imaging marker for cerebral small vessel disease, and have been shown to be related to increased risk of various neurological diseases, including stroke and dementia. Automated quantification of PVS would greatly help to advance research into its etiology and its potential as a risk indicator of disease. We propose a convolutional network regression method to quantify the extent of PVS in the basal ganglia from 3D brain MRI. We first segment the basal ganglia and subsequently apply a 3D convolutional regression network designed for small object detection within this region of interest. The network takes an image as input, and outputs a quantification score of PVS. The network has significantly more convolution

operations than pooling ones and no final activation, allowing it to span the space of real numbers. We validated our approach using a dataset of 2000 brain MRI scans scored visually. Experiments with varying sizes of training and test sets showed that a good performance can be achieved with a training set of only 200 scans. With a training set of 1000 scans, the intraclass correlation coefficient (ICC) between our scoring method and the expert's visual score was 0.74. Our method outperforms by a large margin - more than 0.10 - four more conventional automated approaches based on intensities, scale-invariant feature transform, and random forest. We show that the network learns the structures of interest and investigate the influence of hyper-parameters on the performance. We also evaluate the reproducibility of our network using a set of 60 subjects scanned twice (scan-rescan reproducibility). On this set our network achieves an ICC of 0.93, while the intrarater agreement reaches 0.80. Furthermore, the automated PVS scoring correlates similarly to age as visual scoring.

1 Introduction

This chapter addresses the problem of automated quantification of enlarged perivascular spaces from MR images. The perivascular space - also called Virchow-Robin space - is the space between a vein or an artery and pia mater, the envelope covering the brain. These spaces are known to have a tendency to dilate for reasons not yet clearly understood (Adams et al., 2015). Enlarged - or dilated - perivascular spaces (PVS) can be identified as hyperintensities on T2-weighted MRI. In Figure 1.1, we show examples of PVS in T2-weighted scans. Several studies have investigated the presence of PVS as an emerging biomarker for various brain diseases such as dementia (Mills et al., 2007), stroke (Selvarajah et al., 2009), multiple sclerosis (Achiron and Faibel, 2002) and Parkinson (Zijlmans et al., 2004). In this chapter we focus on PVS located in the basal ganglia. There, the structure of PVS may for instance relate to the presence or absence of beta-amyloid, a protein that has been implicated in Alzheimer's disease (Pollock et al., 1997). Previous work on automated PVS quantification focused on the basal ganglia as well (González-Castro et al., 2016; Gonzalez-Castro et al., 2017), and clinical studies generally rate the PVS presence especially in the basal ganglia and centrum semiovale (Wardlaw et al., 2013).

Manual annotation of PVS is a challenging and very time consuming task: PVS are thin and small structures - often at the resolution limit of 1.5T and 3T MRI scanners - with much variation in their size and shape. Raters need to zoom and scroll through slices to differentiate PVS from similarly appearing brain lesions such as lacunar infarcts or small white matter lesions. Additionally, many PVS can be present within a single scan. In our dataset, for instance, there were up to 35 PVS within a single slice of the basal ganglia. Current clinical studies rely on visual scoring systems, in which expert human raters count the number of

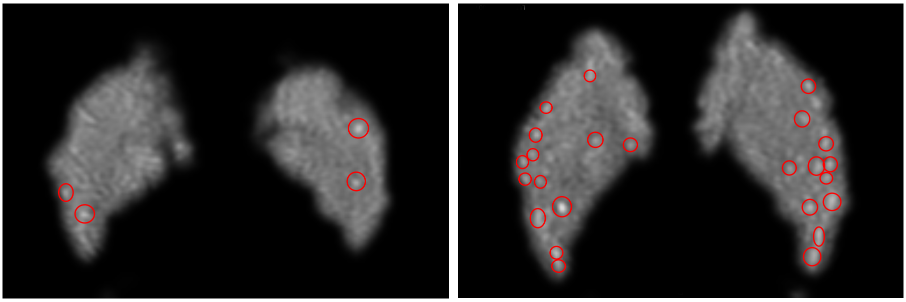


Figure 1.1: **Examples of enlarged perivascular spaces in the basal ganglia.** PVS are circled in red. The PVS have been counted in this slice (Section 2.1). Note that to correctly identify PVS, clinicians need to scroll through slices to check the 3D structure of the candidate lesions.

PVS within a given subcortical structure or region of interest (ROI) (Adams et al., 2013, 2015) or rate the PVS on a 5 point scale.

Recently several groups have addressed PVS quantification using different scenarios and techniques. Ramirez et al. (2015) developed interactive segmentation methods based on intensity thresholding. Park et al. (2016) proposed an automated PVS segmentation method based on Haar-like features. This approach was exclusively evaluated on 7 Tesla MRI scans and needed a large amount of pixel-wise annotations for training. Ballerini et al. (2016) used a Frangi filter to enhance PVS and perform segmentation of individual PVS. They evaluated their performance using a discrete 5-category PVS scoring system (Potter et al., 2015a). In González-Castro et al. (2016); Gonzalez-Castro et al. (2017), in contrast with above approaches, the same authors did not aim to segment individual PVS. They directly formulated the problem as a binary classification - few or many PVS - and used bag of words descriptors with support vector machine classification. Our work extends this by proposing, instead of a binary score, a continuous score, translating the presence PVS. Recently we published a weakly supervised method using neural networks to detect PVS in the basal ganglia (Dubost et al., 2017). Our former work targeted a detection problem, and was evaluated with manually annotated PVS, while in this work we introduce automated PVS scores without considering the location information, and focus on the evaluation of these scores.

Our proposed method relies on a 3D regression convolutional neural network (CNN). One of the main advantages of CNN in comparison to other machine learning techniques, is that the features are automatically computed to maximize the final objective function. 3D CNNs have recently received much attention in the medical imaging literature, for instance for segmentation (Chen et al., 2018; Bortsova et al., 2017; Çiçek et al., 2016), landmark detection (Ghesu et al., 2016) or lesion detection (Dou et al., 2016). CNN regression tasks have been less addressed in medical imaging. For example Miao et al. (2016) employed a set of local 2D CNN regressors for 2D/3D registration. Xie et al. (2018a) proposed a fully convolutional network to count cells by regressing their 2D density maps generated from dot-annotations.

Contributions. In this chapter we propose an automated scoring method to quantify PVS in the basal ganglia. The method is based on a 3D-CNN for regression problems and uses only visual scores labels for training. This scoring method eases the annotation effort and provides a fine scale quantification. We demonstrate the potential of our method on PVS in the basal ganglia. We show that our method correlates well with the visual scores of expert human raters and that the correlation of the automated scores with increasing age is similar to that of visual scores. It is the first time that an automated PVS quantification method is evaluated on such a large dataset (2000 MR scans).

2 Materials and Methods

The objective of our method is to automatically reproduce the PVS visual scores. Our framework consists of two steps. We first isolate the region of interest (ROI) and then apply a regression convolutional neural network (CNN) to compute the PVS presence score.

2.1 Data

In our experiments we used brain MRI scans from the Rotterdam Scan Study. The Rotterdam Scan Study is an MRI based prospective population study investigating - among others - neurological diseases in the middle aged and elderly (Ikram et al., 2015). The scans used in our experiment were acquired with a GE 1.5 Tesla scanner, between 2005 and 2011. The age of the participants ranges from 60 to 96 years old.

The scans were visually scored by a single expert rater (H. Adams), who counted - without indicating their location - the number of PVS in the basal ganglia, in the slice showing the anterior commissure (Adams et al., 2015) (see Fig 1.1 for a few examples). The number of PVS in this slice correlates with the number of PVS in the whole volume (Adams et al., 2013).

2.1.1 Size of the Datasets

In total, the visually scored dataset contains 2017 3D MRI scans from 3 different sub-cohorts. From these 2017 scans, 40 scans have also been visually scored by a second trained rater (F. Dubost), and 25 scans have been marked with dot annotations (by H. Adams) at the center of PVS to check the focus of the network. Note that only PVS in the slice showing the anterior commissure have been marked. In addition, we used 46 other scans for which 23 study participants were scanned twice within a short period (19 ± 11 days). The 46 scans of this reproducibility set are not part of the 2017 scans mentioned above and were not visually scored for PVS.

2.1.2 Scans Characteristics

We used PD-weighted images for our experiments. The scans were acquired according to the following protocol: 12,300 ms repetition time, 17.3 ms echo time, 16.86 KHz bandwidth, 90-180° flip angle, 1.6 mm slice thickness, 25 cm^2 field of view, 416×256 matrix size. The images are reconstructed to a $512 \times 512 \times 192$ matrix. The voxel resolution is $0.49 \times 0.49 \times 0.8$mm^3. Note that these PD-weighted images have a contrast similar to T2-weighted images, the modality more commonly used to detect PVS.

2.1.3 Quality of the Visual Scoring

Visual PVS scores have been created according to a standard procedure proposed in the international consortium UNIVRSE (Adams et al., 2015). H. Adams established the UNIVRSE standardized PVS scoring system and had three years' experience in identifying PVS at the moment he annotated the scans for the current study. Intrarater reliability for this scoring has been computed on the Rotterdam Scan Study, and was reported to be excellent in the basal ganglia (Intraclass Correlation Coefficient (ICC) of 0.80 computed on 85 scans) and inter-rater reliability was reported to be good (ICC of 0.62 on 105 scans) (Adams et al., 2013). We plotted a histogram of the PVS distribution in Figure 1.7.

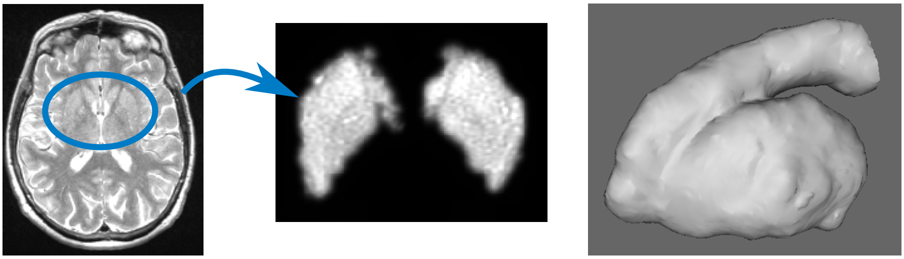


Figure 1.2: **Preprocessing: computation of a smooth mask of the basal ganglia.** From left to right: full MRI scan in axial view; basal ganglia after computation of the smooth mask; 3D rendering of the basal ganglia.

2.2 Preprocessing - Smooth ROI

We first extract a smooth ROI, which can be seen as a spatial prior and focuses the neural network to a predefined anatomical region. In case of 3D images, computing a ROI also helps avoiding the overload of GPU memory and allows to build deeper networks and to train faster.

A binary mask would arbitrarily impose a hard constraint on the input data and can lead to unwanted border effects. Therefore we propose to compute a smooth mask.

Each scan is first registered to MNI space resulting in the hypermatrix $V \in \mathbb{R}^{H \times W \times D}$. A binary mask of the ROI, $M_b \in \{0,1\}^{H \times W \times D}$, is then created using a standard algorithm for subcortical segmentation (Desikan et al., 2006a). The mask is then dilated by first applying 4 consecutive morphological binary dilations with a square connectivity equal to one (6 neighbors in 3D) and subsequently smoothed by convolving the mask with a Gaussian kernel of standard deviation σ. The dilation ensures that PVS located at the border of the ROI are not segmented out. The resulting smooth mask $M_s \in [0,1]^{H \times W \times D}$ is then multiplied element-wise with the volume V, and cropped in all 3

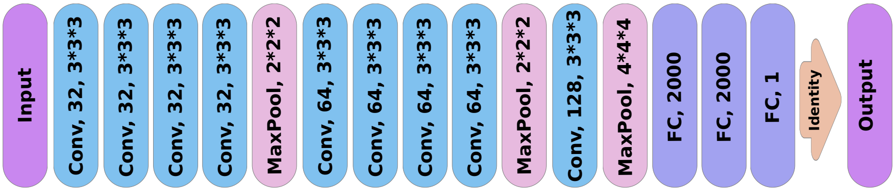


Figure 1.3: **3D Regression CNN Architecture.** The first two blocks consist of 4 3D convolutions followed by a max-pooling. The last block, before the fully connected layers, only has one convolution followed by a larger max-pooling. After each convolutional layer, we apply a rectified linear unit activation. This architecture is specifically designed to detect small lesions.

dimensions around its center of mass to get the final preprocessed image $S \in \mathbb{R}^{h \times w \times d}$, with $h \leq H$, $w \leq W$ and $d \leq D$. In the following sections we refer to S as the smooth ROI. See figure 1.2 for an illustration of the computation of the smooth ROI. We rescale S by dividing by the maximum intensity such that $S \in [0,1]^{h \times w \times d}$. This type of intensity standardization has been successfully used in other deep learning frameworks for quantification and detection of brain lesions (Dou et al., 2016).

2.3 3D Convolutional Regression Network

Once the smooth ROI S is computed we use it as input to a convolutional neural network (CNN) which proceeds to the regression task.

Our CNN architecture is similar to that of VGG (Simonyan and Zisserman, 2015a) but uses 3D convolutional kernels and a single input channel. Additionally, we adapt the architecture for better detection of small structures. We detail our architecture in the following paragraph. Please refer to Figure 1.3 for a visual representation of the network.

The network consists of two blocks of consecutively stacked convolutional layers with small filter size: $3 \times 3 \times 3$, followed by a third block containing a single convolutional layer. We could not expand the network further because of the size of our GPU memory. Note that we do not use any padding and the size of the feature maps is thus reduced after each convolution. Therefore, the input ROI should be sufficiently large to ensure that PVS located close to its border are not missed. After each convolutional layer we apply a rectified linear unit activation. Between each block of convolutions, a maxpooling layer downsamples the feature maps by 2 in each dimension (Figure 1.3). We increase the number of features maps by 2 after each pooling, following the recommendations in Simonyan and Zisserman (2015a). The last pooling layer downsamples its input by 4. The network ends with two fully connected (FC) layers of $c = 2000$ units and a final FC layer of a single unit.

As we framed the problem as a regression, the output should span $\mathbb{R}$. The last activation is then only the identity function. The network parameters are optimized using the mean squared error between $y \in \mathbb{N}^n$, the PVS visual scores, and $\hat{y} \in R^n$, the output of the network. The PVS score $\hat{y}$ is therefore optimized to predict the number of PVS inside the basal ganglia in the slice showing the anterior commissure. However, contrary to an PVS count, our PVS scoring can span $\mathbb{R}$ and not only $\mathbb{N}$. The use of a continuous scoring can reflect the uncertainty in identifying a lesion as an PVS. Besides, the network is regularized only using data-augmentation (Section 3.1).

Architecture choices can be explained as follows. In the brain there can be different type of lesions appearing similar on a given MRI modality. PVS are for instance difficult to discriminate from lacunar infarcts on our PD-weighted scans. Therefore complex features should be extracted at high image resolution, before any significant downsampling. For this reason we place the majority of the convolutional layers before and right after the first maxpooling. Once these small structures have been detected, there is no need to reach a higher level of abstraction: they only need to be counted. That is our motivation to perform only few pooling operations and finish with a large $4 \times 4 \times 4$ pooling. The role of the fully connected layers is to estimate the PVS score based on the PVS detections provided by the output of the last pooling layer. Ideally the output of the last pooling layer could be a set of low dimensional feature maps highlighting the structures of interest, in our case the PVS.

3 Experiments and Results

In order to evaluate the performance of the proposed quantification technique, we conduct seven experiments. In the two first experiments we investigate the behavior of the network and check if the network focuses on PVS. The third series of experiments compares our method with visual scores and with other automated approaches to PVS quantification. Then we investigate the influence of the number of training samples. In the fifth experiment we analyze the influence of several hyper-parameters on the performance of the network. In the sixth experiment, we assess the reproducibility of our method on short term repeat scans. Finally we show how our PVS scoring correlates with age.

3.1 Experimental Settings

In each experiment the preprocessing is the same (Section 2.2). The basal ganglia is segmented with the subcortical segmentation of FreeSurfer (Desikan et al., 2006a). All parameters are left as default, except for the skull stripping preflooding height threshold which is set to 10. Registration to MNI space is computed with the rigid registration implemented in Elastix (Klein et al., 2010) and uses default parameters with mutual information as similarity measure. The voxel size stays the same in dimensions x and y (both 0.5mm) but is different in dimension z (0.8mm before registration and 0.5 after). The Gaussian kernel used to smooth the ROI has a standard deviation $\sigma = 2$ pixel units. The cropped CNN inputs S have a size of $168 \times 128 \times 84$ voxels. We initialize the weights of the CNN by sampling from a Gaussian distribution, use Adadelta (Zeiler, 2012) for optimization and augment the training data with randomly transformed samples. The transformation parameters are uniformly drawn from an interval of 0.2

radians for rotation, 2 pixels for translation and flipping of x and y axes.

The network is trained per sample (mini-batches of a single 3D image). We implemented our algorithms in Python in Keras and Theano and ran the experiments on a Nvidia GeForce GTX 1070 GPU. This GPU has 8GB of GPU RAM, which prevents us from extending the network.

The average training time is one day. We stop the training after the validation loss converged to a stable value. Once the CNN is trained and given the smooth ROI S, the automated PVS scoring takes 440 ms on our GPU and 2 min on our CPU. We evaluate the results using four metrics: the Pearson correlation coefficient, the Spearman correlation coefficient, the intraclass correlation coefficient (ICC) and the mean square error (MSE). We compute these metrics between the visual scores of the expert rater (H.Adams) and the output of the method, the automated PVS scores. ICC is the metric most commonly used to evaluate the reliability of visual rating methods, and has also been used in previous epidemiological studies of PVS (Adams et al., 2013). We consider it as the standard metric in our experiments.

3.2 Saliency Maps

In figure 1.4, we computed 6 saliency maps using our trained model (Section 2.3). Saliency maps are computed as the derivative of the automated PVS scores (the output of the network) with respect to the input image (Simonyan et al., 2014). Saliency maps highlight regions which contributed to the PVS score and consequently we expect them to highlight PVS.

After rescaling intensities of the saliency map in $[0, 1]$, we circled the regions with a value higher than 0.5. Most strongly highlighted regions correspond to PVS, although sometimes large PVS are only slightly highlighted, while smaller-sized PVS (that do not exceed the threshold to be counted as enlarged by the expert human rater) can be highlighted as well. In most of the cases, regions with values in $[0, 0.5]$ in the saliency maps actually correspond to thin perivascular spaces.

It should be noted, however, that enlargement of perivascular space is not a 0/1 phenomenon (as a visual rating assumes) but actually happens on a continuous scale, and it is very likely that the CNN counts the PVS in a volumetric manner. Many smaller-sized PVS would thus not be counted by the expert human rater as 'enlarged' but could still slightly contribute to the total PVS burden computed by the algorithm, hence the slightly highlighted (values in $[0, 0.5]$) in the saliency maps.

Note that, while the annotator considers PVS only in a single slice, the algorithm is considering the complete 3D volume. The number of PVS in the annotated slice and in the total volume of the basal ganglia are strongly correlated (Adams et al., 2013). The algorithm most probably uses this correlation and locates PVS in the total volume and scales down its output to make it match the number of PVS in the annotated slice. We observe the same behavior in Section 3.3.

3.3 Occlusion of PVS

In this section, we perform another experiment to verify that the algorithm learns PVS. We use a set of 25 scans in which PVS have been marked with a dot in the slice showing the anterior commissure (Section 2.1).

The experiment consists of occluding marked PVS with small 3D blocks (1.5x1.5x4.8 mm) of the mean intensity of the basal ganglia. We successively occlude n PVS, with $n \in [1; 6]$, in all images and recompute, for each n, the predicted PVS score for each image. We expect the scores to decrease as we occlude more PVS.

Figure 1.5 shows the results. In the left plot, the automated scores linearly decrease as more PVS are occluded, until four PVS have been removed. Note that in the right plot, it seems that the automated score of scans with a lower amount of PVS decreases quicker than for scans with many PVS. In scans with many PVS, the PVS selected for occlusion may more frequently be a slightly enlarged PVS, considered as a limit case by the algorithm and hence having a small impact on the automated score. In the left plot, after four PVS have been removed, the slope of the curve decreases. At that point, most of PVS have been removed from the images, only remains images with many PVS.

One could expect the scores to decrease by n as we occlude n PVS. The scores decrease instead by a smaller amount. The automated PVS scores are indeed computed across the volume and scaled down to match the visual scores that were based on a single slice. Removing a single PVS slightly affects the automated PVS score.

In Figure 1.6, we performed additional experiments to verify this hypothesis. As expected, we notice that occluding a lesion in the input image reduces the intensity at that location in the saliency map. However we also notice, that the more lesions are occluded in a single slice, the lower the influence on the saliency map is, and the less the automated PVS score decreases. After removing the most obvious lesions, we actually start to occlude only slightly enlarged ones, that have a lower impact on the quantification. If we now occlude more enlarged lesions in other slices, the saliency map and automated PVS scores are again more impacted. This confirms the hypothesis that the algorithm considers PVS across the volume of the basal ganglia.

For comparison, we also occluded the image of Figure 1.6 at random locations. We occluded 1-5 random locations in the basal ganglia, and repeated the experiment 100 times. With no occlusion, the PVS score was 7.14. One random occlusion led to an PVS score of 7.12 +/- 0.1 (standard deviation). This decrease is negligible in comparison to the change in PVS score after occluding one PVS: 6.86. Occluding five random locations led to an PVS score of 7.10 +/- 0.28. Thus, occluding PVS has a significant impact on the PVS score in contrast to occluding random locations. We can therefore conclude that the algorithm focuses on PVS.

3.4 Comparison to visual scores and to other automated approaches

Table 1.1: **Correlation with expert's visual scores for the proposed method and four other more conventional approaches.** We also report the mean square error (MSE). Best performance in each column is indicated in bold.

Method	Pearson	Spearman	ICC	MSE
Intensity (a)	0.38	0.19	0.37	18.36
Volume (b)	0.47	0.34	-0.27	116.2
Components (c)	0.63	0.48	0.63	9.88
SIFT-BOW (d)	0.57	0.59	0.55	10.05
3D Regression CNN	**0.75**	**0.61**	**0.74**	**6.14**

In this section we compare the automated scores to visual scores and demonstrate the effectiveness of our method in comparison to four other automated approaches.

For the first series of experiments, the dataset is randomly split into the following subsets: 1289 scans for training, 323 for validation and 405 for

Table 1.2: **Intraclass Correlation Coefficent for Interrater Reliability.** A stands for the rater H. Adams, B1 is the first rating of the rater F. Dubost, and B2 the second rating of rater F. Dubost. See end Section 3.4 for more details.

	A	B1	B2
B1	0.70		
B2	0.68	0.80	
Proposed Method	0.80	0.62	0.70

testing. The first three methods (a,b and c) quantify hyperintense regions in the MRI scans. The last method (d) is a machine learning approach similar to a state-of-the-art technique for PVS quantification in the basal ganglia (González-Castro et al., 2016; Gonzalez-Castro et al., 2017). These four baseline methods are particularly interesting as they cover a wide range of complexity.

The output of method (a) is simply the average of all voxels intensity values inside the ROI S. Both the second (b) and third (c) method first thresholds S to keep only high intensities. This threshold is optimized on the training set, without applying the intensity standardization described in Section 3.1. We denote S_t the thresholded image S. The output of (b) is the volume - the count of non-zero values - of the threshold image S_t. The output of (c) is the number of connected components in S_t. The method (d) computes bag of visual words (BoW) features using SIFT (Lowe, 2004) as descriptors and uses a regression forest. SIFT parameters are tuned - by visual assessment - to highlight PVS on the training set. 2D SIFT are computed in each of the 15 slices surrounding the slice annotated by clinicians. In our experiments, using more surrounding slices proved to be too complex for the model, which would then fail to learn the aimed correlation. The number of words in the BoW dictionary was set to 100 for each slice. Concatenating the feature vectors of each slice yielded better results than averaging these vectors. The BoW features for the entire volume are therefore vectors of $15 * 100 = 1500$ elements. The regression forest has 3000 trees and a maximum depth of 50 nodes.

For all these other automated approaches, the regression results need to be rescaled to be able to compute the ICC. We apply a linear transformation to the outputs. The predicted values can consequently become negative. The parameters of this transformation are optimized to maximize the ICC on the validation set.

We report the results of this experiment in Table 1.1. The regression network performs best for all measures and outperforms the other methods by a large margin - more than 0.10 - for both Pearson correlation and ICC. Our method performs significantly better than all four baselines (William's test, p-value $<$ 0.00001 for baselines (a), (b), (d) and $<$ 0.01 for baseline (c)). Methods (c) and (d) are the strongest baselines

Figure 1.7 presents scatter plots of the estimated outputs for each method. We notice that method (c) sometimes strongly overestimates the number of PVS in scans with no PVS. Such errors do not happen with our regression network. On the other hand, method (d), and to a lesser extent the proposed method, have a tendency to underestimate PVS in scans with the largest amounts of PVS. A possible explanation for this underestimation is that in case of a larger number of PVS, the chance of having lesions close to each other is higher. This makes the detection more challenging. Several very close PVS may appear similar to a single larger PVS in other scans.

Note that despite its simplicity, method (c) performs reasonably well, especially in comparison with the random forest (d), which is much more complex (more parameters). However, note that the performance metrics of method (c) as displayed in Table 1.1 are strongly influenced by few scans having many PVS (see Figure 1.7). If we ignore these scans and recompute the ICC for scans with only 20 PVS or less, method (c) drops to 0.48 ICC and 11.02 MSE while method (d) gets to 0.59 ICC and 9.22 MSE and the proposed method is at 0.68 ICC and 6.74 MSE.

In the experiments described above, we have demonstrated that the scores predicted by our algorithm have a good to excellent (according to Cicchetti (1994) guidelines) correlation with the scores of a single expert rater (H. Adams). However, as the algorithm is trained with the scores of this same rater, its predictions may be biased.

To verify this, we evaluated the performance of our algorithm on a smaller set annotated by two raters (H. Adams and F. Dubost) (see Section 2.1). For this experiment we trained the algorithm on a training set (training + validation) of 1600 scans and a test set of 400 scans. Table 1.2 shows the results.

3.5 Learning Curve

In this section we study how the number of annotated scans used for optimization influences the performance of our automated quantification method.

We train our network using different subsets of the 2017 MRI scans described in Section 2. We perform experiments using 5 different sizes for the training set. For a fixed number of training scans, we repeat the experiment 5 times with different randomly drawn train/test splits of the data. This results in $5 * 5 = 25$ experiments with different random train/test splits of the data. Figure 1.8 shows the results of the experiment. In the training set size, we count both training (80%) and validation (20%) sets.

Even with a relatively small training set size (200 scans) our method performs well: the correlation between the automated and visual scores reaches an ICC of 0.66. Our model reaches its best performance (ICC of 0.74 ± 0.044) with 1000 training scans. Using more scans does not bring further improvement. Using only a few training scans (40) leads to a significant drop in performance (ICC of 0.30) with higher standard deviation.

3.6 Analysis of Network Parameters

In this section, we investigate the influence of several parameters of the model. Table 1.3 summarizes a set of experiments performed on the same split of training, validation and testing set, which sizes are 1289 scans, 323 and 405 respectively. In this series of experiments the varying parameters are: registration to MNI space (MNI); number of features in the first layer (Feat1stL); for the data augmentation, flipping scans in the direction of the sagittal axis (FlipX), the left-right axis (FlipY), the longitudinal axis (FlipZ); the layout of the fully connected layer (FC), where e.g. 2*2000 means 2 layers of 2000 neurons each; the loss (Loss), where MSE stands for mean square error, MCE for mean cubic error, MQE for mean quartic error, Tukey for Tukey's biweight and RSME for root mean square error. Blocks is the number of convolutional blocks as described in Section 2 and Conv/Block is the number of convolutional layers per block. ICC and MSE are the metrics we computed on the test set. Note that we conducted these experiments a posteriori and did not use these results to tune the parameters of the method for the experiments in sections 3.4, 3.5, 3.7 and 3.8.

Table 1.3 is separated in several categories of experiments. The first line shows the algorithm implemented in this chapter. On the second line we notice that registering to MNI spaces does not provide a large improvement. In the third category, we investigate several loss functions. MSE provides a better performance. In the fourth category we investigate different architectures. Reducing the number of convolutional layers or fully connected layers does not bring a large difference, neither does changing the number of features in the first layer. To perform the experiment with three blocks, we halved the number of features maps in each layer. This architecture yields worse results than shallower architectures. The last category investigates different levels of data augmentation. The most important augmentation is flipping the images in the y-axis, which is an anatomically plausible augmentation. Other forms of data augmentation bring no improvement in this scenario and can make the training process more difficult and slower.

Overall, in this problem setting, registering to MNI is not necessary, MSE is the loss of choice, architecture changes do not bring significant differences but one could prefer using a smaller network for faster training, and the best augmentation is flipping in anatomically plausible directions.

We noticed in Table 1.3 that, considering the ICC, shallower networks perform similar to deeper ones in this problem (in regards to the MSE, the proposed deep network performs slighted better though). We investigate the behavior of these shallower models for smaller amount of training samples. Figure 1.9 shows a comparison of the learning curves of a deep network (as implemented in this chapter) and a shallow network with two blocks and a single convolutional layer per block (see Table 1.3 and Section 2). The deep network performs slightly better and the difference in performance is larger for smaller training sets.

3.7 Reproducibility

In order to evaluate the reproducibility of our automated PVS scoring method, we run our algorithm on the reproducibility set described in Section 2.1. In this experiment we consider two versions of our model. For each version, we trained a set of 5 networks with randomly selected training sets of scans. For both versions, we actually use the same networks as in the learning curve experiments (Figure 1.8). In the first version, the networks have been optimized using 1000 scans and yields a ICC of 0.740 ± 0.044 with visual scores from the human rater. In the second version, the networks have been optimized only with 40 scans and yields an ICC of 0.298 ± 0.062 with the visual scores. On the reproducibility set, the first model yields an ICC of 0.93 ± 0.02 between the first and second sets of scans. The second model yields an ICC of 0.83 ± 0.011. According to Cicchetti (1994) guidelines , both models have an excellent correlation. Adams et al. (2013) reported an intrarater agreement of 0.80 ICC for PVS visual scoring in the basal ganglia. In our study, the second rater also had an intrarater agreement of 0.80 ICC (Section 3.4). From this comparison we can conclude that our automated PVS scoring appears to be more reproducible than visual scoring.

3.8 Correlation with Age

Now that we have demonstrated the performance of our approach in comparison with other automated approaches and human visual scores, we investigate the correlation of our automated PVS scores with clinical factors. PVS have been shown to correlate with age (Potter et al., 2015b). We consider correlations between age and visual PVS scores from human raters (a), and between age and automated PVS scores (b). We split our dataset into a training set of 1000 scans and a testing set of the remaining 1000 scans. We use the training set to optimize the parameters of our automated scoring algorithm. For (a) and (b), we perform a zero-inflated negative binomial regression. The model is zero-inflated to take into account the over-representation of participants with no PVS (see PVS distribution across participants in Figure 1.7). The per-decade odds ratio and 95% confidence interval are for (a) 1.30 ± 0.08 and for (b) 1.34 ± 0.07. Figure 1.10 shows the trends of increasing PVS scores with age, which are very similar for automated and visual scores.

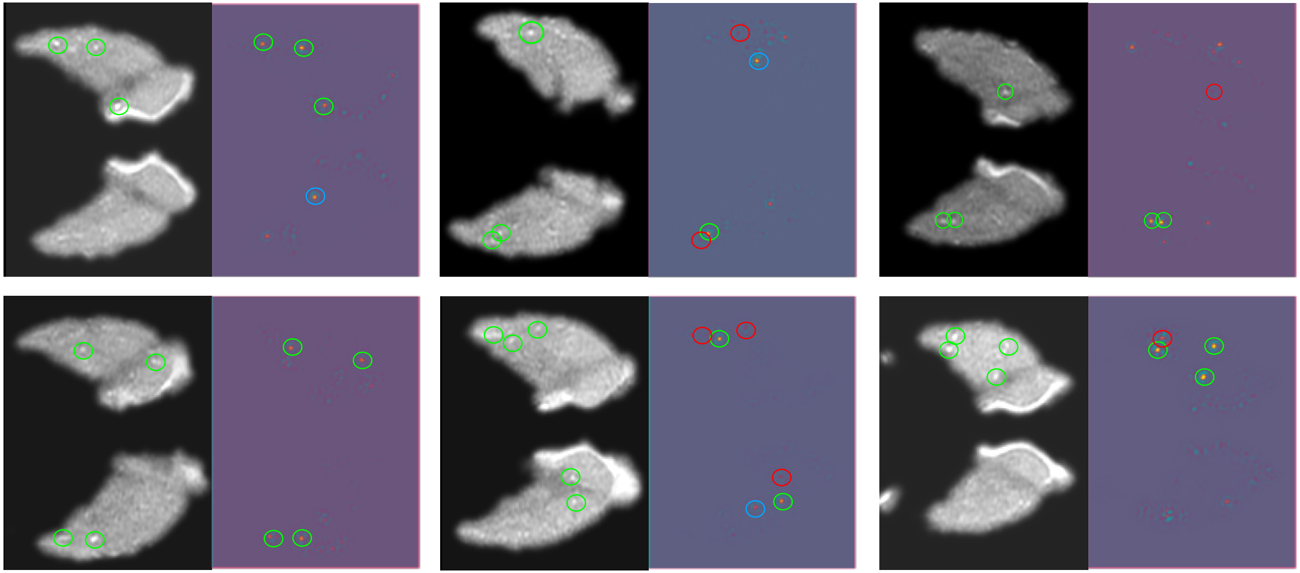


Figure 1.4: **Examples of saliency maps.** We display the middle slices of 6 scans on the left, and the corresponding rescaled saliency maps produced by the network (Simonyan et al., 2014) in right. On the scans, green circles highlight PVS. On the saliency maps, regions of high activation matching with an PVS in the scan are circle in green. When these do not match any PVS, they are circled in blue. If a region is not activated by the presence of an PVS, it is circled in red.

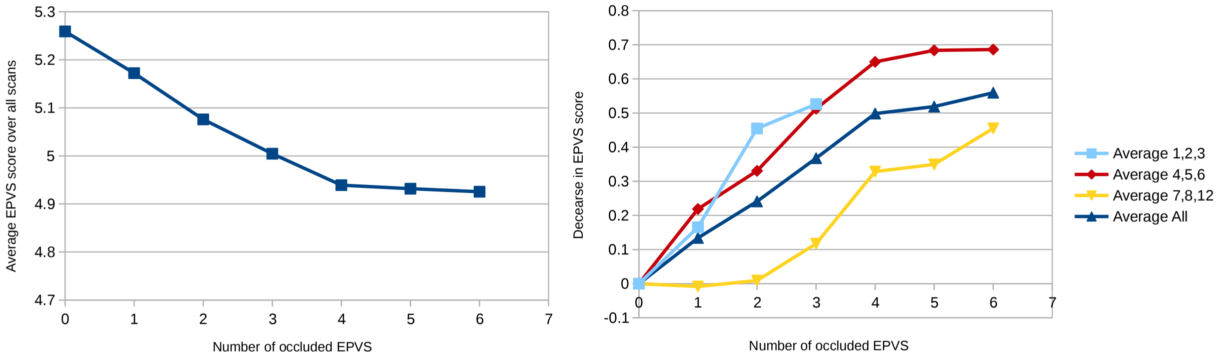


Figure 1.5: **Predicted scores after PVS occlusion for increasing number occluded PVS.** On the right plot, the scores are averaged among group of scans having similar initial numbers of PVS. For instance the light blue label stands for scans having either 1, 2 or 3 marked PVS in the slice showing the anterior commissure. Once a scan has no PVS to remove in the annotated slice, the predicted score stays the same. For the light blue curve, as no scans has more than 3 PVS to occlude, the curve would stay constant after 3 PVS. We do not plot these points.

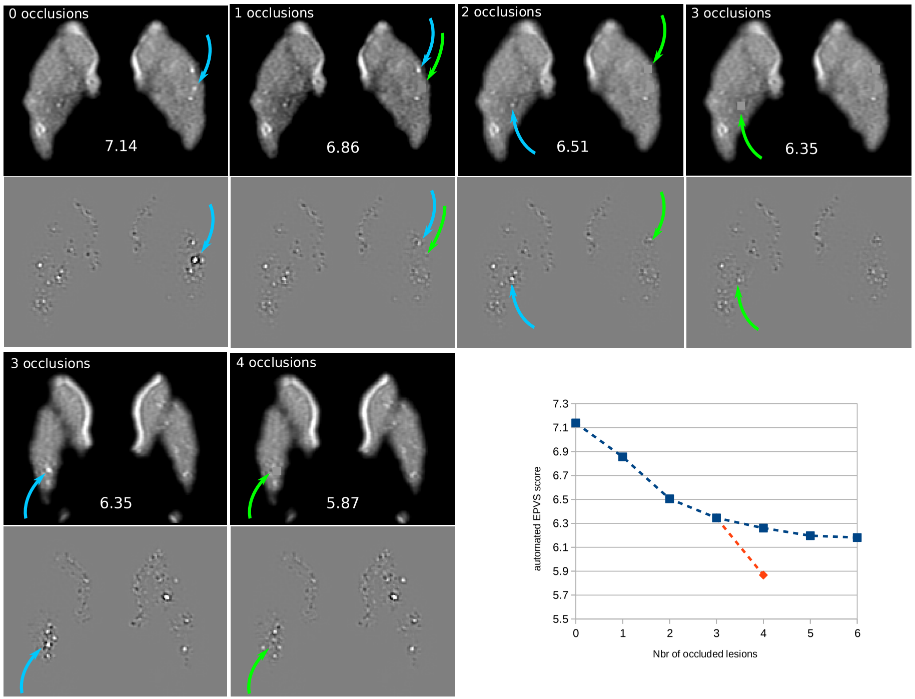


Figure 1.6: **Occlusion in a single image.** Several lesions are progressively occluded in the same image. In the first row, we occlude lesions in the slice annotated by the expert rater. In the second row, we occlude an additional lesion in an upper slice of the same 3D image. The top image is the input image, and the bottom one is the corresponding saliency map (see Section 3.2). We indicated the number of occluded lesions at the top of each image, and the updated automated PVS score in the middle. Blue arrows indicate lesions which will be occluded next. Green arrows indicate the location of lesions that have just been occluded. In the bottom-left of the figure, we also plot the evolution of the automated PVS score while removing lesions. Blue is removing lesions in the annotated slice. Orange is removing the lesion in the upper slice (second row of images). Results are interpreted in Section 3.3.

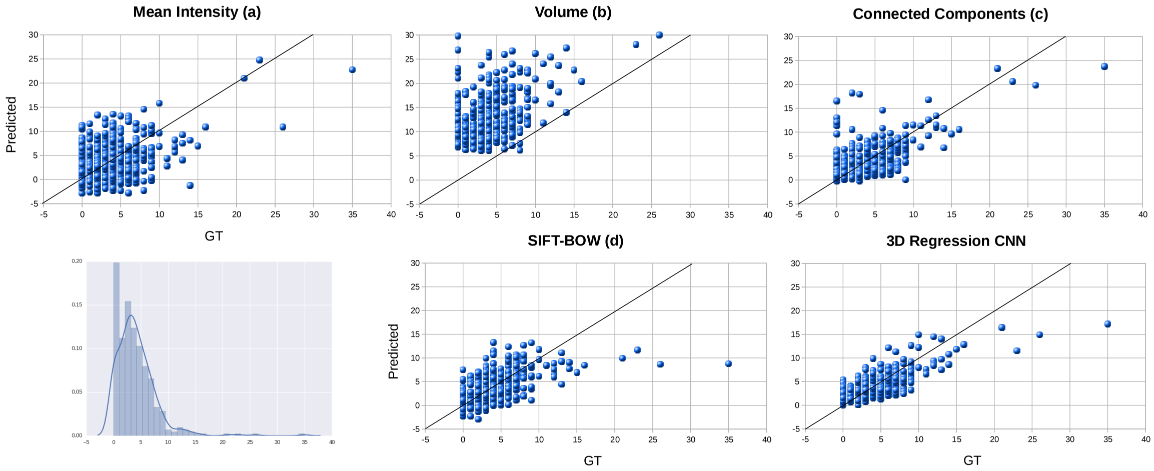


Figure 1.7: **Regression results on the test set.** The different methods are detailed in Section 3.4. The ground truths are represented on the x-axis. The predicted outputs of the methods are on the y-axis. See Table 1.1 for correlation coefficients. On the bottom-left, we plot a histogram of the distribution of PVS visual scores across scans.

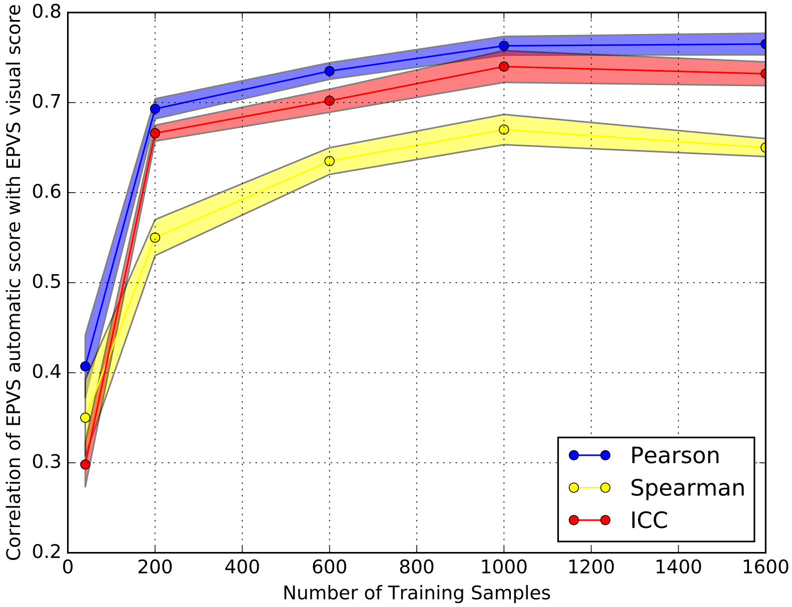


Figure 1.8: textbfLearning Curve. The number of scans for training (80% training set and 20% validation set) is represented on the x-axis. Three different correlation coefficients (Pearson, Spearman, Intraclass) with visual scores are represented on the y-axis. For a given number of training samples, we average the results over 5 experiments. For each experiment, the data is randomly split into non-overlapping train, validation and test sets. Across experiments, the sets overlap (Monte Carlo cross-validation). For each point, we plot the 95% confidence interval related to the corresponding 5 experiments.

Table 1.3: **Network parameters and Corresponding Results.** See Section 3.6 for details. ⋆ indicates the proposed method. In ⋆⋆, the network has only half of the features of the other variants in this table. The best results per category of experiment are in bold.

MNI	Feat1stL	FlipX	FlipY	FlipZ	FC	Loss	Blocks	Conv/Block	ICC	MSE
⋆ 1	32	1	1	1	2*2000	MSE	2	4	0.783	4.37
0	32	1	1	1	2*2000	MSE	2	4	0.771	4.99
1	32	1	1	1	2*2000	MCE	2	4	**0.751**	6.11
1	32	1	1	1	2*2000	MFE	2	4	0.708	**5.76**
1	32	1	1	1	2*2000	tukey	2	4	did not	converge
1	32	1	1	1	2*2000	RMSE	2	4	did not	converge
1	32	1	1	1	2*2000	MSE	1	4	0.807	4.76
1	32	1	1	1	2*2000	MSE	2	3	0.805	4.93
1	32	1	1	1	2*2000	MSE	2	2	0.808	5.03
1	32	1	1	1	2*2000	MSE	2	1	0.803	4.85
⋆⋆1	16	1	1	1	2*2000	MSE	3	4	0.767	5.64
1	16	1	1	1	2*2000	MSE	2	1	0.776	5.17
1	32	1	1	1	2*2000	MSE	2	1	**0.803**	**4.85**
1	64	1	1	1	2*2000	MSE	2	1	0.780	5.14
1	32	1	1	1	1*2000	MSE	2	4	0.781	5.06
1	32	1	1	1	0	MSE	2	4	0.788	4.76
1	32	0	1	0	2*2000	MSE	2	4	**0.787**	**4.65**
1	32	0	0	0	2*2000	MSE	2	4	0.742	5.88
1	32	No	Data	Augm	2*2000	MSE	2	4	0.742	6.23

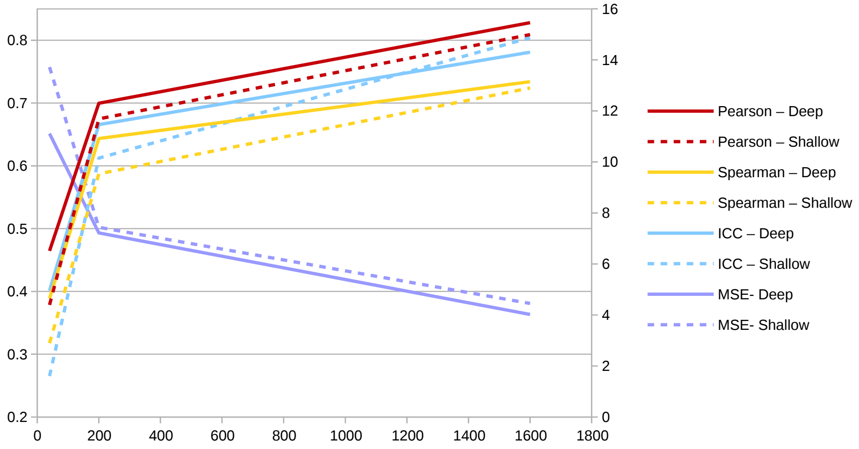


Figure 1.9: **Learning Curve of shallow and deep networks.**The number of training and validation scans is displayed on the x-axis. The correlations coefficients (Pearson, Spearman and ICC) between automated and visual scores are displayed on the left y-axis (the scale ranges from 0.2 to 0.85). The MSE between automated and visual scores is displayed on the right y-axis. Solid lines are used for the deep network, and dotted lines for the shallow network.

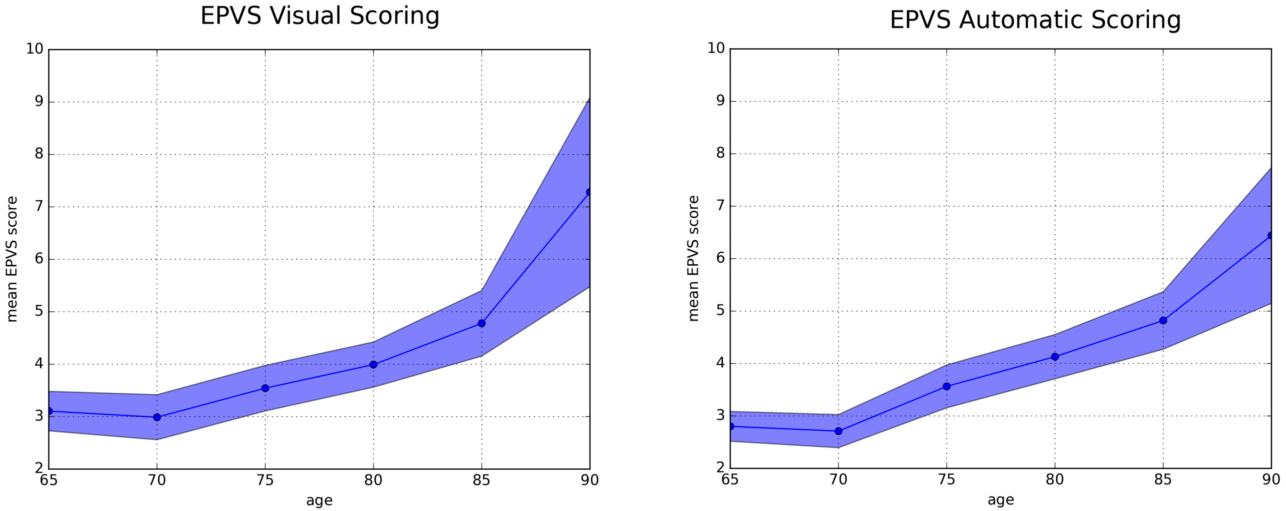


Figure 1.10: **PVS scores as a function of age.** We show the mean PVS scores and 95% confidence interval per 5 years, for visual (left) and automated (right).

4 Discussion

We showed that our regression network indeed focuses on PVS to compute the automated scores, although no information about the location of these lesions had been given during training. This automated scoring has a good agreement with the visual scoring performed by a single expert rater, is highly reproducible, and significantly outperforms the scoring of the four more conventional methods we compared to.

Few other papers addressed PVS quantification. In contrast with our approach, Gonzalez-Castro et al. (2017) formulated the problem as a binary classification where a threshold is set to $t = 10$ PVS to differentiate between the severe or mild presence of PVS. The authors use bag of visual words and SIFT features (Lowe, 2004), similar to our baseline method (d), and achieve an accuracy of 82% on a test set of 80 scans. The regression approach as presented in our paper provides a much finer - and therefore likely more relevant - quantification than this binary classification. In addition, in our experiments, the regression network yields much better results than the bag of words with SIFT approach (Table 1.1). In Figure 1.7, the bag of word approach (d) is also more spread along the second principal component, meaning that this method is on average less precise in its quantification (high mean square error). This matches with the mean square errors reported in Table 1.1.

More recently, the same authors (Ballerini et al., 2016) used methods based on vessel enhancement filtering, and reported a Spearman correlation of 0.75 with a 5-category PVS ranking (the Potter scale, Potter et al. (2015a)) in the centrum semiovale. Our method achieves a Pearson correlation of 0.763 ± 0.026 and a Spearman correlation of 0.670 ± 0.042 with visual scoring in the basal ganglia. These results cannot directly be compared as the regions, visual scoring systems, and datasets are different. A possible advantage of the visual PVS score used in our work (Adams et al., 2013) with respect to the Potter scale (Potter et al., 2015a), is that it provides a finer quantification. In our study population, the majority of images would fall into the first 2 categories of the Potter scale (0 PVS and 1-10 PVS), while the score of Adams et al. (2013) allows further separation.

Ramirez et al. (2015) developed interactive segmentation methods based on intensity thresholding. The authors show good results but need the intervention of a human rater, which in large datasets is an important drawback. Our method is fully automated. Park et al. (2016) proposed an automated PVS segmentation method based on Haar-like features. This method reaches up to 64% Dice coefficient with ground truth annotations. This approach was exclusively evaluated on 7 Tesla MRI scans, needs a large amount of pixel-wise annotations for training, and was only evaluated on a dataset of 17 young healthy subjects. We evaluated our method on the Rotterdam Scan Study (Ikram et al., 2015), a population-based study in middle aged and elderly subjects. The elderly subjects are more prone to cerebral small vessel diseases, and may have other types of brain lesions, similar to PVS (e.g. lacunar infarcts). This makes the exclusive quantification of PVS more challenging on our dataset, but also closer to the clinical need.

Several other learning-based approaches to counting objects in images have been proposed in the literature, mostly in case of 2D images. These techniques also often need labels about the location of the target objects. Lempitsky and Zisserman (2010) proposed a supervised learning method to count objects in images. However their method is based on density map regression and relies on dot annotations for training. More recently, Walach and Wolf (2016) proposed a convolutional neural network with boosting and selective sampling for cell and pedestrian counting. Their method is also base on density map regression and needs dot annotations. Ren and Zemel (2017a) proposed a method to jointly count and segment instances in 2D images. They combined a recurrent neural network with an attention model. However the method needs a pixel-wise ground truth for its segmentation component. Seguí et al. (2015) proposed a convolution neural network for counting handwritten digits and pedestrians. The network are optimized for classification with weak global labels: the number of instances of the target object. This work is closer to our method, as we also use weak global labels. However, we use regression networks. All these method were evaluated only on 2D tasks. For instance, overcoming occlusions is one of the main difficulties tackled in pedestrian counting, a problem which does not occur in case of 3D volumes.

Our method is both reproducible (0.93 ICC) and agrees well with the visual scores of the expert human rater it has been trained on: the correlation between the automated and visual scores is 0.74 ICC, which is in between interrater agreement (0.62 in Adams et al. (2013), and 0.68 and 0.70 in our study (Table 1.2)) and intrarater agreement (0.80 in both in Adams et al. (2013) and our study (Table 1.2)). Furthermore, the correlation between the automated scores and the visual scores of a second expert human rater - which have not been seen during training - is similar to that of the interrater agreeement (Table 1.2). Therefore, we believe our method is sufficiently precise and robust to perform automated PVS quantification in large scale clinical research. The processing time stays low enough: 440 ms on GPU per scan given to the regression network. However, as all images in our database were acquired with a single scanner, for application in different data it would need to be evaluated on a multi-center dataset to further verify its robustness. Additionally, our method was exclusively evaluated in the basal ganglia, as perivascular spaces in this region are suggested to be most clinically relevant (Potter et al., 2015a). In other PVS research studies (Adams et al., 2015; Ikram et al., 2015; Maillard et al., 2016; Hilal et al., 2013), PVS can also be visually scored in other brain regions such as centrum semiovale, hippocampus and midbrain (Adams et al., 2013). This is particularly relevant as the location of PVS is thought to differ with etiology and even relate to different clinical outcomes (Banerjee et al., 2017; Charidimou et al., 2017). We expect our method to perform similarly in other brain regions.

Contrary to PVS visual scoring, we quantify the PVS in the entire ROI volume and not only in a single slice. However it has been shown (Adams et al., 2015) that the visual PVS score in a slice of the basal ganglia is highly correlated to the PVS visual score in the entire volume. The results from experiments with occlusion suggest that our method uses this correlation by detecting PVS in the whole volume and scaling the score down to match the visual scores done in a single slice. The automated scores are more robust than visual ones in this regard. Training a classifier on visual scores of the whole basal ganglia volume could provide an even more robust approach and could prove itself useful to investigate more subtle correlations with clinical factors.

In this work, we did not limit our input to the visually scored slice. The human rater indeed uses information from more than just one slice to discriminate PVS from similarly appearing brain lesions, and we expect the network to benefit from this information as well. Besides, we expect that quantifying PVS in the entire basal ganglia, fusing information from multiple slices, is more reliable than only quantifying them in a single slice.

In Table 1.2, while the correlation between the automated scores and the visual scores of the second rater (F. Dubost) is slightly lower than the correlation between both raters (F. Dubost and H.Adams), it is still higher than the interrater ICC reported in Adams et al. (2013). Overall, we believe that this table shows that we automatized the first rater (H. Adams), with interrater and intrarater reliabilities similar to that of expert human raters.

Looking at the learning curve (Fig 1.8), it seems that the performance of the network does not improve when training on more than 1000 images. This could mean that either this is the maximum achievable performance using this ground truth or that increasing the complexity of the network (by adding layers and feature maps) could still lead to an increase in performance. However the experiments conducted in section 3.1 suggest that a similar performance can be achieve by shallower networks. Though, shallower networks seem to perform worse for small training sets. More regularization (Dropout, L1 or L2) may help to reduce the drop in performance (for both deep and shallow networks) when training on small amount of samples.

In theory, we think that the performance of the network could be further boosted with e.g. attention mechanisms (Mnih et al., 2014), given highly accurate ground truth labels. However, we cannot expect any methods trained on ratings of a single rater to perform better than intra-rater agreement (here ICC of 0.8). In several cases (see Table 1.3) our prediction reaches this level of agreement with the expert's scores. That is why we did not experiment with more complicated methods: with the current ground truth based on visual assessment, we can not expect nor would we be able to meaningfully evaluate any further performance gain.

The large size of the required training set could be seen as an obstacle to the clinical application of the automated scoring method. However, although our best performance is achieved with a training set of 1000 scans, training with 200 scans already provides a good performance. We believe this method can be extended to and would be useful for other large clinical and population-based studies such as ADNI (Jack et al., 2008), UK Biobank (Sudlow et al., 2015) and

German National Cohort (Ahrens et al., 2014).

5 Conclusion

We presented a novel regression method to automatically quantify the amount of enlarged perivascular spaces in the basal ganglia in brain MRI. We validated our approach on 2000 brain MRI scans (using different sizes for the testing set, up to a maximum of 1960 scans). Our method significantly outperforms four other more conventional automated approaches. The agreement with visual scoring (ICC of 0.74) is higher than the inter-observer agreements (ICC of 0.68 and 0.70). The scan-rescan reproducibility is very high (ICC of 0.93), compared to intra-observer agreement (ICC of 0.80). Our result are relatively robust across network architectures. We also demonstrated that the automated PVS scores correlate with age, similarly to the visual PVS scores. We believe that this method can replace visual scoring of PVS in epidemiological and clinical studies.

Acknowledgments

This research was funded by The Netherlands Organisation for Health Research and Development (ZonMw) Project 104003005.

Chapter 2

Hydranet: Data Augmentation for Regression Neural Networks

Abstract

Deep learning techniques are often criticized to heavily depend on a large quantity of labeled data. This problem is even more challenging in medical image analysis where the annotator expertise is often scarce. We propose a novel data-augmentation method to regularize neural network regressors that learn from a single global label per image. The principle of the method is to create new samples by recombining existing ones. We demonstrate the performance of our algorithm on two tasks: estimation of the number of enlarged perivascular spaces in the basal ganglia, and estimation of white matter hyperintensities volume. We show that the proposed method improves the performance over more basic data augmentation. The proposed method reached an intraclass correlation coefficient between ground truth and network predictions of 0.73 on the first task and 0.84 on the second task, only using between 25 and 30

scans with a single global label per scan for training. With the same number of training scans, more conventional data augmentation methods could only reach intraclass correlation coefficients of 0.68 on the first task, and 0.79 on the second task.

1 Introduction

Deep learning techniques are getting increasingly popular for image analysis but are often dependent on a large quantity of labeled data. In case of medical images, this problem is even stronger as data acquisition is administratively and technically more complex, as data sharing is more restricted, and as the annotator expertise is scarce.

To address biomarker (e.g. number or volume of lesions) quantification, many methods propose to optimize first a segmentation problem and then derive the target quantity with simpler methods. These approaches require expensive voxel-wise annotations. In this work, we circumvent the segmentation problem by optimizing our method to directly regress the target quantity (Cole et al., 2017; González et al., 2018; Wang et al., 2019; Lee and Kim, 2018). Therefore we need only a single label per image instead of voxel-wise annotations. Our main contribution is that we push this limit even further by proposing a data augmentation method to reduce the number of training images required to optimize the regressors. The proposed method is designed for global image-level labels that represent a countable quantity. Its principle is to combine *real* training samples to construct many more *virtual* training samples. During training, our model takes as input random sets of images and is optimized to predict a single label for each of these sets that denotes the sum of the labels of all images of the set. This is motivated by the idea that adding a large quantity of virtual samples with weaker labels may reduce the over-fitting to training samples and improve the generalization to unseen data.

1.1 Related Work

Data augmentation can act as a regularizer and improve the generalization performance of neural networks. In addition to simple data-augmentations such as rotation, translation and flipping, the authors of Unet (Ronneberger et al., 2015) stress for instance that random elastic deformations significantly improved the performance of their model. Generative adversarial networks have for instance also been used to generate training samples, and hence reduce the over-fitting (Sixt et al., 2018).

Recently, data augmentation methods using combinations of training samples have been published. Zhang et al. (2018a) proposed to construct virtual training samples by computing a linear combination of pairs of real training samples. The corresponding one-hot labels are summed with the same coefficients. The authors evaluated their method on classification datasets from computer vision and on a speech dataset, and demonstrate that their method improves the generalization of state-of-the-art neural networks. Simultaneously, Inoue (2018) and Tokozume et al. (2018) reached similar conclusions. In case of grayscale volumetric inputs, summing image intensity values could overlay the target structures, confuse discriminative shapes, and thus harm the performance of the network. With our method, training samples can be combined without overlaying the intensity values. The other difference with the above-mentioned approaches is that our method is also not designed for classification, but for regression of global labels, such as volume or count in an image. With the proposed combination of samples, our method computes plausible augmentation.

2 Methods

The principle of the proposed data augmentation method is to create many new (and weaker) training samples by combining existing ones (see Figure 2.1). In the remainder, the original samples are called *real samples*, and the newly created samples are called *virtual samples*.

2.1 Proposed Data Augmentation.

During training, the model is not optimized on single real samples I with label y, but on sets S of n random samples $I_1, I_2, ..., I_n$ with label $y_s = \sum_{i=1}^{n} y_i$, with y_i the label of sample I_i. These sets S with labels y_s are the virtual samples. Consequently, the loss function L is computed directly on these virtual samples S and not anymore the individual real samples I_i. This approach is designed for labels describing a quantitative element in the samples, such as volume or count in an image.

To create the sets S, the samples I_i are drawn without replacement from the training set at each epoch. To create more combinations of samples, and to allow the model to use the real samples for its optimization, the size of the sets S can randomly vary in $\{1, n\}$ during training. If the training set contains m samples, with our method, we can create $\sum_{i=1}^{n} \binom{m}{i}$ possible different combinations (the order of the samples I_i in S has no effect on the optimization).

2.1.1 Difference with mini-batch stochastic gradient descent (SGD)

In mini-batch SGD, the model is also optimized on sets of random samples, but the loss function L is computed individually for each sample of the batch, and then summed (averaged). For the proposed method, the predictions are first

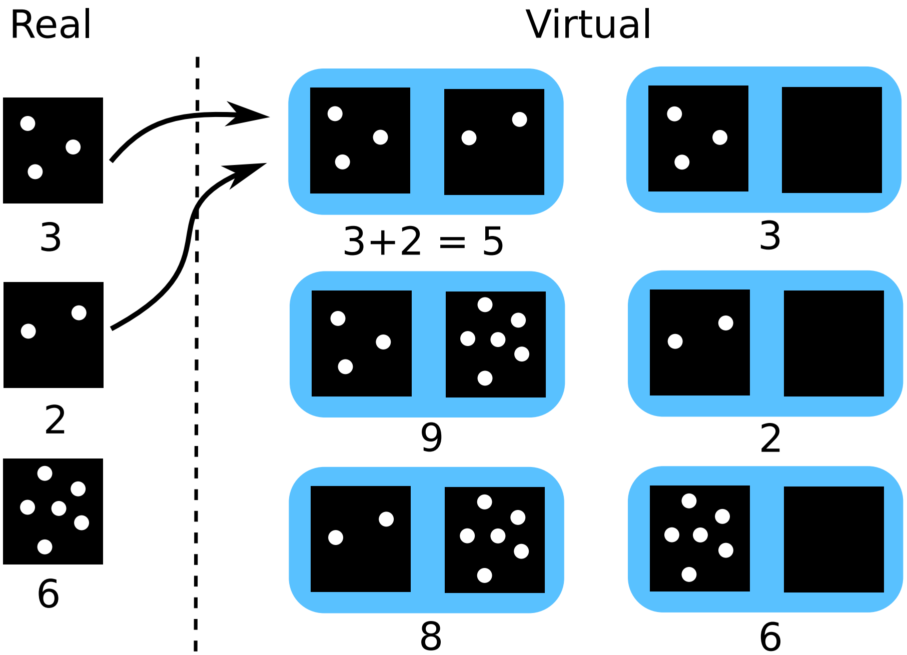


Figure 2.1: **Creating virtual training samples by recombining real training samples for regression tasks.** The real training samples are displayed on the left, and the virtual samples on the right. The label is indicated under each sample, and corresponds to the number of white blobs. By recombining samples, we can significantly increase the size of the training dataset. For example, by recombining the real samples with labels 3 and 2, we can create a new sample with label 5 (arrows). All possible combinations are shown in blue. For the illustration, we show only combinations of two samples, but any number of samples can be combined. In our experiments, we used combinations of maximum 4 samples.

summed, and the loss function is then computed a single time. For non-linear loss functions, this is not equivalent: $\sum_{i=1}^{n} L(\hat{y}_i, y_i) \neq L(\sum_{i=1}^{n} \hat{y}_i, \sum_{i=1}^{n} y_i)$, with $\hat{y}_i$ the model's prediction for sample I_i.

2.1.2 Regularization Strength

The regularization strength can usually be modulated by at least one parameter, for instance the degree of rotation applied to the input image, or the percentage of neurons dropped in Dropout (Srivastava et al., 2014). In the proposed method, the regularization effect can be controlled by varying the average number of samples used to create combinations.

2.2 Implementation

We optimize a regression neural network with a 3D image for input, and global label representing a volume or count for output. There are at least two possible implementations of the proposed method. The first implementation could consist of modifying the computation of the loss function across samples in a mini-batch, and provide mini-batches of random size. Alternatively the model's architecture could be adapted to receive the set of images. We opted for the second approach.

2.2.1 Base Regressor

Figure 2.2 left shows the architecture of the base regression neural network. It is both simple (196 418 parameters) and flexible to allow fast prototyping. There is no activation function after the last layer. The output $\hat{y}$ can therefore span $\mathbb{R}$ and the network is optimized with the mean squared error (MSE). We call this regression network f, such that $f(x) = \hat{y}$, with x the input image.

2.2.2 Combination of Samples

To process several images simultaneously, we replicate n times the regressor f during training (Figure 2.2 right), resulting in n different branches $f_1, f_2, ..., f_n$

that receive the images $I_1, I_2, ..., I_n$. The weights of each head f_i are shared such that $f_i = f$. A new network g is constructed as:

$$g(S) = g(I_1, I_2, ..., I_n) = \sum_{i=1}^{n} f_i(I_i) = \sum_{i=1}^{n} f(I_i) = \sum_{i=1}^{n} \hat{y}_i. \tag{2.1}$$

To allow the size of the sets S to randomly vary in $\{1, n\}$ during training, each element of S has a chance p to be a black image B of zero intensities only (Figure 2.1 right column). With $f(B) = 0$, the following situation becomes possible:

$$g(S) = f(I_j) + \sum_{i=1, i \neq j}^{n} f_i(B) = f(I_j) + (n-1)f(B) = f(I_j). \tag{2.2}$$

For this implementation, the batch size b has to be a multiple of the number of branches n. We chose $b = n$ due to constraints in GPU memory. The regularization strength is controlled by the averaged number of samples used to create combinations, hence depends on n and p. During inference, to predict the label for a single input image, the input of all other branches is set to zero.

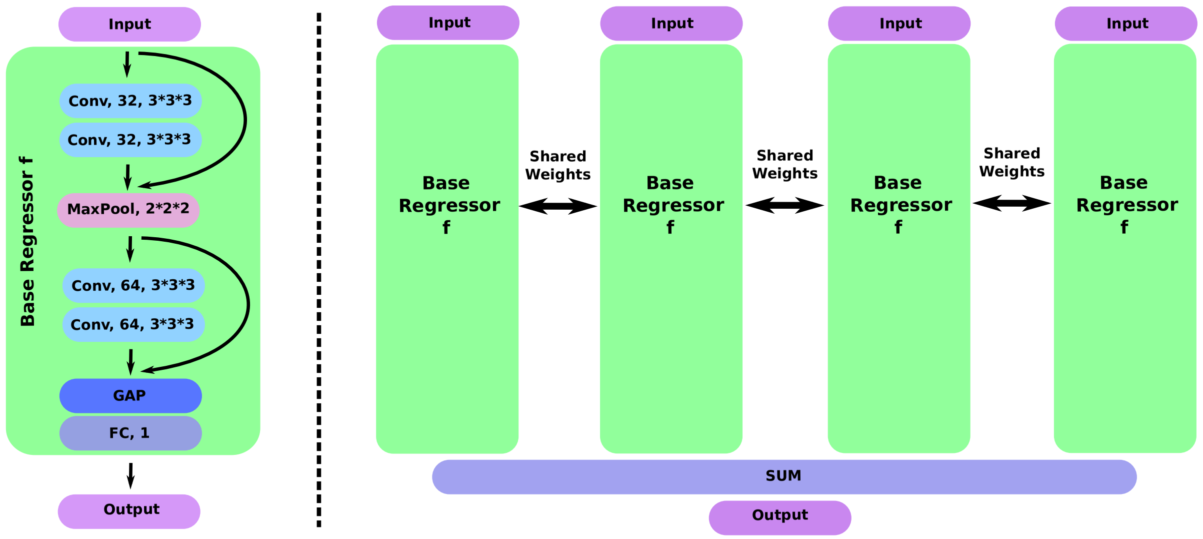


Figure 2.2: **Architectures**. On the left, architecture of the base regressor f. 'Conv' stands for 3D convolutions, followed by the number of feature maps, and the kernel size. After each convolution, there is a ReLU activation. The round arrows are skip connections with concatenated feature maps. GAP stands for Global Average Pooling layer, and FC for Fully Connected layer. On the right, example of our data augmentation method with $n = 4$ replications. Each replication f_i is a copy of the base regressor f on the left. Once the training is done, all f_i but one can be removed, and the evaluation is performed using the original architecture.

3 Experiments

Enlarged perivascular spaces (PVS) and white matter hyperintensities (WMH) are two types of brain lesions associated with small vessel disease. The method is evaluated for the estimation of number PVS in the basal ganglia, and estimation of WMH volume. We compare the performance of our method to that of the base regressor f with and without and Dropout, and for different sizes of training set.

The PVS dataset contains T2-weighted scans, from 2017 subjects, acquired from a 1.5T GE scanner. The scans were visually scored by an expert rater who counted the PVS in the basal ganglia in a single slice. The WMH dataset is the training set of the MICCAI2017's WMH challenge (Kuijf et al., 2019). We use the available 2D multi-slice FLAIR-weighted MRI scans as input to the networks. Scans were acquired from 60 participants from 3 centers: 20 scans from Amsterdam (GE scanner), 20 from Utrecht (Philips) and 20 from Singapore (Siemens). Although the ground truths of the challenge are pixel-wise, we only used the number of WMH voxels as ground truth during training.

For the regression of PVS in the basal ganglia, a mask of the basal ganglia is created with the subcortical segmentation algorithm from FreeSurfer (Desikan et al., 2006b), and smoothed with a gaussian filter (standard deviation of 2 voxels) before being applied the image. The result is subsequently cropped around the basal ganglia. For the WMH dataset, we only crop each image around its center of mass, weighted by the voxel intensities. For both tasks the intensities are then rescaled between 0 and 1.

During training, for all methods, the images are randomly augmented on-the-fly with standard methods. The possible augmentations are flipping in x, y or z, 3D rotation from -0.2 to 0.2 radians and random translations in x, y or z from -2 to 2 voxels. Adadelta (Zeiler, 2012) is used as optimizer. The networks are trained with batch-size $b = 4$. For the proposed method, the network's architecture has then four branches ($n = b = 4$). During an epoch, the proposed method gets as input m/n different combinations of n training samples, were m is the total number of training images. During the same epoch, the base regressor f simply gets the m images separately (in batches of size $b = 4$). For the proposed method p was set to 0.1. In some experiments with Dropout (Srivastava et al., 2014) we included a dropout layer after each convolution and after the global pooling layer. The code is written in Keras with Tensorflow as backend, and the experiments were run on a Nvidia GeForce GTX 1070 GPU.

For the PVS dataset, we experiment with varying size of training set, between 12 and 25 scans. The validation set always contains the same 5 scans. All methods are evaluated on the same separated test set of 1977 scans. For the WMH dataset, the set is split into 30 training scans and 30 testing scans. Six scan from the training set are used as validation scans. In both cases, the dataset is randomly (uniform distribution) split into training and testing sets. For the PVS dataset, once the dataset has been split into 30 training scans and 1977 testing scan, we manually sample scans to keep a pseudo-uniform distribution of the lesion count when decreasing the number of training scans.

To compare the automated predictions to visual scoring (for PVS) or volumes (for WMH), we use two evaluation metrics: the mean squared error (MSE), and the intraclass correlation coefficient (ICC).

3.1 Results

3.1.1 Enlarged Perivascular Spaces (PVS)

Figure 2.3 compares the proposed method to the base regressor f on the PVS datasets, and for an increasing number of training samples. Their performance is also compared to the average interrater agreement computed for the same problem and reported in Chapter 1. The proposed method always reaches a better MSE than the conventional methods for all training set sizes. The proposed method also significantly outperforms the base regressor in ICC (Williams' test p-value < 0.001) when averaging the predictions of the methods across the four points of their learning curve.

3.1.2 White Matter Hyperintensities (WMH)

We conducted three series of experiments, and trained in total five neural networks (Table 2.1). When using small training sets, the proposed method outperforms the base network f, when optimized either for MSE or for mean absolute error. With larger training sets, the difference of performance reduces, and the base regressor performs slightly better on the ICC.

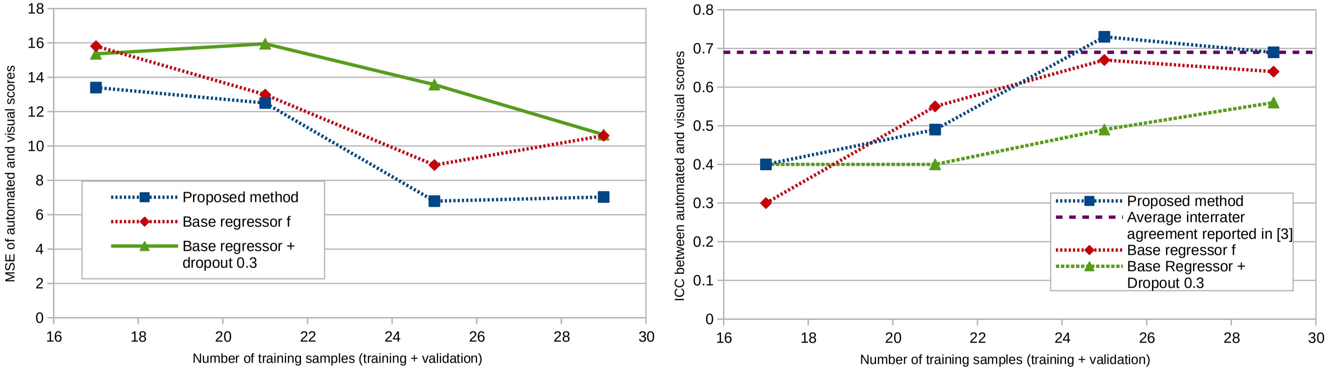


Figure 2.3: **Comparison between the proposed method with $n = 4$ and the base regressor on the PVS dataset.** MSE is displayed on the left, and ICC on the right.

Table 2.1: **Results on the WMH dataset** We conducted three series of experiments with different training set sizes and loss functions. In the two first rows, we repeated the experiments with three random initializations of the weights (on the same split), and report mean and standard deviation. MAE is an acronym for mean absolute error.

Method	Training scans	Testing scans	Loss	Performance (ICC)
Base Network f	30	30	MSE	0.79 ± 0.12
Proposed Method	30	30	MSE	$\mathbf{0.84} \pm \mathbf{0.02}$
Base Network f	30	30	MAE	0.78
Proposed Method	30	30	MAE	**0.87**
Base Network f	40	20	MSE	**0.89**
Proposed Method	40	20	MSE	0.86

4 Discussion and Conclusion

With the proposed data augmentation method, we could reach the inter-rater agreement performance on PVS quantification reported in Chapter 1 with only 25 training scans, and without pretraining.

In Chapter 1, we also regressed the number of PVS in the basal ganglia with a neural network. We achieve a similar result (0.73 ICC) while training on 25 scans instead of 1000. Zhang et al. (2018a) also proposed to combine training samples as a data augmentation method. In their experiments, combining more than $n = 2$ images does not bring any improvement. With the proposed method, training with combinations of four images brought improvement over only using pairs of images. We did not experiment with values of n larger than 4 due to GPU memory constraints. Contrary to the expected gain in generalization, on both PVS (Figure 2.3) and WMH datasets, using Dropout (Srivastava et al., 2014) worsened the results when training on very little data, even with low dropout rates such as 0.3. As dropout already did not improve the performance of the baseline, we do not expect improvement by including dropout in the proposed method.

To create combination of images for the proposed method, images where drawn without replacement for the sake of implementation simplicity. The regularization strength could be increased by drawing samples with replacement, which could be beneficial for small training sets. We also mentioned two possible implementations of the proposed method: (1) changing the computation of the loss over mini-batches, (2) replicating the architecture of network. In this work we used the second approach, as it was simpler to implement with our library (Keras). However with this approach, all samples used in a given the combination have to be simultaneously processed by the network, which can cause GPU memory overload in case of large 3D images or large values of n. The first approach does not suffer from this overload, as the samples can be successively loaded, while only saving the individual scalar predictions in the GPU memory. In case of large 3D images, we would consequently recommend implementing the first approach.

Acknowledgments

This research was funded by the Netherlands Organisation for Health Research and Development (ZonMw) Project104003005, with additional support of Netherlands Organisation for Scientific Research (NWO), project NWO-EWVIDI 639.022.010 and project NWO-TTW Perspectief Programme P15-26. This work was partly carried out on the Dutch national e-infrastructure with the support of SURFCooperative.

Part C

Object Detection

Chapter 3

Weakly Supervised Object Detection with 2D and 3D Regression Neural Networks

Abstract

Finding automatically multiple lesions in large images is a common problem in medical image analysis. Solving this problem can be challenging if, during optimization, the automated method cannot access information about the location of the lesions nor is given single examples of the lesions. We propose a new weakly supervised detection method using neural networks, that computes attention maps revealing the locations of brain lesions. These attention maps are computed using the last feature maps of a segmentation network optimized only with global image-level labels. The proposed method can generate attention maps at full input resolution without need for interpolation during preprocessing, which allows small lesions to appear in attention maps. For comparison, we modify state-of-the-art methods to compute attention maps for weakly supervised object detection, by using a global regression objective instead of the more conventional classification objective. This regression objective optimizes the number of occurrences of the target object in an image, e.g. the number of brain lesions in a scan, or the number of digits in an image. We study the behavior of the proposed method in MNIST-based detection datasets, and evaluate it for the challenging detection of enlarged perivascular spaces – a type of brain lesion – in a dataset of 2202 3D scans with point-wise annotations in the center of all lesions in four brain regions. In MNIST-based datasets, the proposed method outperforms the other methods. In the brain dataset, the weakly supervised detection methods come close to the human intrarater agreement in each region. The proposed method reaches the best area under the curve in two out of four regions, and has the lowest number of false positive detections in all regions, while its average sensitivity over all regions is similar to that of the other best methods. The proposed method can facilitate epidemiological and clinical studies of enlarged perivascular spaces and help advance research in the etiology of enlarged perivascular spaces and in their relationship with cerebrovascular diseases.

1 Introduction

Weakly supervised machine learning methods are designed to be optimized with limited amounts of labelled data and are very promising for a large number of medical image analysis problems. As medical expertise is scarce and annotation time expensive, unsupervised (Schlegl et al., 2017) and weakly supervised methods (Qi et al., 2017; Bortsova et al., 2018) are most suited to extract information from large medical databases, in which labels are often either sparse or non-existent. In this article, we use attention maps for weakly supervised detection of brain lesions. Attention maps can be computed to reveal discriminative areas for the predictions of neural networks that process images such MRI, CT or X-ray. Most attention maps computation methods have originally been designed to make deep networks more explainable (Zhang et al., 2018b; Oktay et al., 2018; Zhang and Zhu, 2018; Hwang and Kim, 2016). As those methods do not require annotations for the optimization of the networks but only global labels such as biomarkers or phenotypes (Wang et al., 2019), they can also be optimized using only counting objectives such as the number of lesions in a brain region, and subsequently predict the location of these lesions during test time.

We propose a novel weakly supervised detection method, using attention maps computed from the feature maps of a segmentation network architecture optimized with global labels. By using the last feature maps of such an architecture, attention maps can be computed at full input resolution, and small structures can be detected more accurately. In this article, we focus on weak supervision with regression neural networks for counting. Regression networks have widely been optimized with local labels such as voxel coordinates (Redmon et al., 2016), distance maps (Xie et al., 2018a,b) or depth maps (Laina et al., 2016).

Less frequently, regression networks have been used to predict global labels, such as age (Cole et al., 2017; Wang et al., 2019), brain lesion count (Dubost et al., 2017), pedestrian count (Seguí et al., 2015), or car count (Mundhenk et al., 2016). Other researchers have also optimized neural networks to infer count. Ren and Zemel (2017b) combined a recurrent network with an attention model to jointly count and segment the target objects, but need pixel-wise ground truths for the optimization. In bioimaging, methods inferring count have often been applied to cell counting in 2D images (Lempitsky and Zisserman, 2010; Walach and Wolf, 2016; Xie et al., 2018a; Tan et al., 2018; Alam and Islam, 2019). These approaches are often optimized to regress distance or density maps computed from dot annotations at the center of the target objects. Instead of regressing density maps, Paul Cohen et al. (2017) performed cell counting by regressing pixel-wise labels that represent the count of cells in the neighborhood. In our approach, pixel-wise labels are not needed for training: only the image-level count are used. Earlier, Seguí et al. (2015) have also optimized networks using image-level count labels alone for digit and pedestrian count and visualized the attention of the networks. However, they did not quantify the performance of the resulting weakly supervision detection. Xue et al. (2016) performed cell counting also using regression network optimized with patch-wise cell count, computed density maps, but did not quantify the performance on the pixel level. In this article, we optimize regression networks using image-level count labels, but use this as a means for detection.

We compare the proposed method to four state-of-the-art methods (Simonyan et al., 2014; Springenberg et al., 2015; Schlemper et al., 2018; Selvaraju et al., 2017). Other weakly supervised detection methods have been proposed relying, for example, on latent support vector machines (SVMs) (Felzenszwalb et al., 2010), a reformulation of the multiple instance learning mi-SVMs (Andrews et al., 2003), or more recently, on multiple instance learning with attention-based neural networks (Ilse et al., 2018), and on iterative learning with neural networks classifiers, where the training set is made of subsets of most reliable bounding boxes from the last iteration Sangineto et al. (2018).

We evaluate the methods using two datasets: a MNIST-based detection dataset and a dataset for the detection of enlarged perivascular spaces, a type of brain lesion that is associated with cerebral small vessel disease. On 1.5T scans, perivascular spaces become visible when enlarged. Following the neuroimaging standards proposed by Wardlaw et al. (2013), we use the consensus term perivascular space (PVS) throughout the manuscript without always referring to their enlargement. PVS is an emerging biomarker, and ongoing research attempts to better understand their etiology and relation with neurological disorders (Adams et al., 2014; Duperron et al., 2019; Gutierrez et al., 2019). Most of the research on perivascular spaces is based on quantification of PVS burden using visual scores based PVS counts (Adams et al., 2014; Potter et al., 2015c). Next to overall PVS burden, the location of PVS can have a clinical significance that varies depending on the brain region (midbrain, hippocampi, basal ganglia and centrum semiovale) and also within a brain region. For example PVS are thought to be benign when observed where perforating vessels enter the brain region (Jungreis et al., 1988), such as PVS in the lower half of the basal ganglia. Understanding more precisely how the specific locations of PVS can relate with determinants of PVS and outcomes can aid neurology research. Automatically quantifying and detecting PVS is challenging, because PVS are very small (at the limit of the scan resolution) and can easily be confused with several other types of lesions (Dubost et al., 2019b; Adams et al., 2013; Sudre et al., 2018; Brown et al., 2018). Recently, automated methods have been developed to address PVS quantification (Ballerini et al., 2018; Sudre et al., 2018; Sepehrband et al., 2019; Boespflug et al., 2018), but these methods were not evaluated in large datasets or for the detection of individual PVS. The proposed method only requires PVS visual scores for its optimization and is evaluated for the detection of individual PVS. In most of the large imaging studies, PVS are quantified using visual scores based on counts. Considering the generalizability issues of neural networks, using networks that require only PVS count for their optimization can consequently be considered to have more practical impact than networks that require annotations for their optimization.

1.1 State-of-the-art for attention map computation

All state-of-the-art methods investigated in this article are based on convolutional neural networks (CNNs) that compute a pseudo-probability map which indicates the locations of the target objects in the input image. In the rest of the article, we call this map the *attention map*. The methods can be divided into three categories: methods using class activation maps (CAMs), methods based on the gradient of the output of the network, and methods using perturbations of the input of the network.

1.1.0.1 CAM methods This category consists of variants of the class activation maps (CAMs) method proposed by Zhou et al. (2016). CAMs are computed from the deepest feature maps of the network. These feature maps are followed by a global pooling layer, and usually one or more fully connected layers to connect to the output of the network. CAMs are computed during inference as a linear combination of these last feature maps, weighted by the parameters of the fully connected layers learnt during training. If the last feature maps have a much lower resolution than the input – as is the case in deep networks with multiple pooling layers – the resulting attention maps can be very coarse.

This is suboptimal when small objects need to be localized, or when contours need to be segmented precisely. To alleviate this issue, Dubost et al. (2017); Schlemper et al. (2018) proposed to include finer-scale and lower-level feature maps in the computation of the attention maps. Dubost et al. (2017) combined higher and lower level feature maps via skip connections and concatenation similarly to U-Net (Ronneberger et al., 2015), while Schlemper et al. (2018) used gated attention mechanisms, which rely on the implicit computation of internal attention maps. Selvaraju et al. (2017) proposed to generalize CAM to any network architecture, using weights computed with the derivative of the output. Unlike other CAM methods, the method by Selvaraju et al. (2017) does not require the presence of a global pooling layer in the network, and can be computed for any layer of the network.

1.1.0.2 Gradient methods Simonyan et al. (2014) proposed to compute attention maps using the derivative of a classification network's output with respect to the input image. These attention maps are fine-grained, but often noisy. Springenberg et al. (2015) reduced this noise by masking the values corresponding to negative entries of the top gradient (coming from the output of the network) in the ReLU activations. Gradients methods can be applied to any CNN.

1.1.0.3 Perturbation methods Perturbation methods compute attention maps by applying random perturbations to the input and observe the changes in the network output. These methods are model-agnostic, they can be used with any prediction model, not even necessarily restricted to neural networks. One of the simplest and most effective implementations of such methods was recently proposed by Petsiuk et al. (2018) with masking perturbations. The input is masked with a series of random smooth masks, before being passed to the network. Using a linear combination of these masks weighted by the updated network classification scores, the authors could compute attention maps revealing the location of the target object. This method relies on a mask sampling technique, where the masks are first sampled in a lower dimensional space, and then rescaled to the size of the full image. Earlier, Fong and Vedaldi (2017) proposed several other perturbation techniques including replacing a region with a constant value, injecting noise, and blurring the image. Perturbation methods are the most general as they can also be applied to other classifiers than CNN. We do not study perturbation models in this paper, because their optimization was more challenging than that of other methods, especially for the detection of small objects.

1.2 Contributions

The contribution of this work is fourfold. First, we propose a novel weakly-supervised detection method, named *GP-Unet*. The principle of the method is to use a segmentation architecture with skip connections to compute attention maps at full input resolution to help the detection of small objects. A preliminary version of this work was presented in (Dubost et al., 2017).

Second, the proposed method is compared to five previously published methods (Dubost et al., 2017; Schlemper et al., 2018; Selvaraju et al., 2017; Simonyan et al., 2014; Springenberg et al., 2015).

Third, we assess in MNIST-based (LeCun et al., 1998) datasets whether a classification or regression objective performs best for the weakly supervised detection.

Fourth, we evaluate the methods both in MNIST-based detection datasets

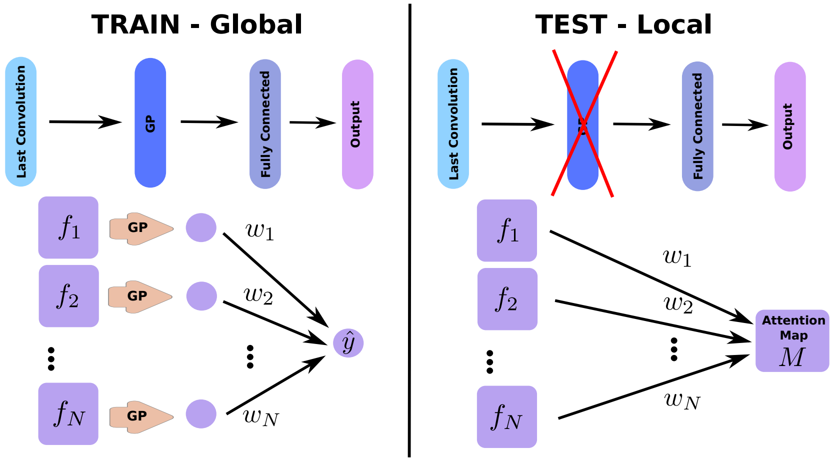


Figure 3.1: **Principle of CAM methods for regression.** GP stands for Global Pooling. f_k correspond to the feature maps of the last convolutional layer. Disks correspond to scalar values. w_k are the weights of the fully connected layer. Left: the architecture of the network during training. Right: the architecture at inference time, where the global pooling is removed. During training, the network outputs a scalar value which is compared to the image level label to compute the loss and update the network's parameters. During testing, the global pooling layer is removed. Consequently, the network outputs an image. This image is computed as the linear combination of feature maps of the layer preceding the global pooling layer using the weights of the following fully connected layer.

and in the 3D detection of enlarged perivascular spaces. The MNIST datasets is used as a faster and more controlled experimental setting to study methodological differences between attention map computation methods, optimization objectives, and architectures. We evaluate the best methods in a real-world practical task with clinical relevance: the detection of PVS. The current work is the largest study to date to evaluate automated PVS detection in a large dataset (four regions and 2202 scans) using center locations of PVS.

2 Methods

We implemented seven methods for weakly supervised detection with CNNs: (a) *GP-Unet* (this article) , (b) *GP-Unet no residual* (Dubost et al., 2017) the first proposed version of GP-Unet, (c) *Gated Attention* (Schlemper et al., 2018), (d) *Grad-CAM* (Selvaraju et al., 2017), (e) *Grad* (Simonyan et al., 2014), (f) *Guided-backpropgation* (Springenberg et al., 2015), and (g) an intensity thresholding method for brain datasets only. For all methods, the CNNs are designed to output a single scalar $\hat{y} \in \mathbb{R}$ and are trained with mean squared error using only global labels: the number of occurrences of target objects $y \in \mathbb{N}$. Then for a given input image I the attention map M is computed at inference time. Below, we detail the computation of these attention maps for each method.

2.1 Computation of the attention maps

2.1.1 CAM methods

The principle of all CAM methods is to use the feature maps – or activation maps – of the network to compute attention maps. CAM methods usually exploit the feature maps of the last convolutional layer of the network, as they are expected to be more closely related to the target prediction than feature maps of intermediate layers. Zhou et al. (2016) first proposed to introduce a global pooling layer after the last convolution. The global pooling layer projects each feature map f_k to a single neuron, resulting in a vector of N scalar values, where N is the number of feature maps f_k in the last layer. The global pooling layer is followed by a fully connected layer to a number of neurons corresponding to the number of classes (for classification), or to a single neuron representing the output $\hat{y} \in \mathbb{R}$ (for regression). The network can then be trained with image-level labels using, for example, a cross-entropy or mean squared error loss function. During inference the global pooling layer can be removed, and the attention map is then computed as a linear combination of the feature maps f_k (before global pooling) using the weights of the fully connected layer w_k:

$$M_{CAM} = \sum_{k}^{N} w_k f_k. \tag{3.1}$$

The computation of CAM attention maps is illustrated in Figure 3.1.

2.1.1.1 GP-Unet In the approach by Zhou et al. (2016) the attention map is computed from the last feature maps of the network, which are often downsampled with respect to the input image due to pooling layers in the network. To alleviate this problem, we use the same principle with the architecture of a segmentation network (U-net from Ronneberger et al. (2015)), i.e. with an upsampling path, where the feature maps f_k of the last convolution layer - before global pooling (GP) - have the same size as the input image I (see architectures in Figure 3.2 and section 2.2). The attention maps are still computed with Equation 3.1.

2.1.1.2 GP-Unet no residual In our earlier work, we proposed another version of GP-Unet (Dubost et al., 2017) based on a deeper architecture without residual connections (see architectures in Figure 3.2 and section 2.2). Experiments showed that such deep architecture was not needed (Dubost et al., 2019a), and could slow the optimization. We refer to this approach as *GP-Unet no residual* in the rest of the paper. To detect hyperintense brain lesions in MRI data Dubost et al. (2017) also rescaled the attention map values to $[0, 1]$ and summed them pixel-wise with rescaled image intensities. This is not needed in the new version of GP-Unet above because residual connections between the input and output of two successive convolutional layers allow the network to learn this operation.

2.1.1.3 Gated Attention While we proposed to upsample and concatenate features maps of different scales (Dubost et al., 2017) as advised for segmentation networks by Ronneberger et al. (2015), Schlemper et al. (2018) proposed instead a more complex gated attention mechanism to combine information from different scales. This gated attention mechanism relies on attention units – also called attention gates – that compute soft attention maps and use these maps to mask irrelevant information in the feature maps. Here, global pooling is applied at every scale s and the results are directly linked to the output by a fully connected layer aggregating information across scales. Schlemper et al. (2018) proposed three aggregation strategies: concatenation, deep supervision (Lee et al., 2015), and fine-tuning by training the network for each scale separately. With the fine tuning strategy, the authors reached a slightly higher performance than concatenation and deep supervision. For the sake of simplicity, we employed the concatenation strategy in our experiments. See Figure 3.2 for an illustration of the architectures of Gated Attention and of GP-Unet. The attention maps M_{Gated} of the gated attention mechanism method are computed as:

$$M_{Gated} = \sum_{s} \sum_{k}^{N_s} w_k^s f_k^s, \tag{3.2}$$

where w_k^s are the weights of the last fully connected layer for the neurons computed from the feature maps f_k^s at scale s.

2.1.1.4 Grad-CAM Finally, Grad-CAM (Selvaraju et al., 2017) is a generalization of CAM Zhou et al. (2016) to any network architecture. The computation of the attention map is similar to Equation 3.1, but instead of the weights w_k, uses new weights α_k in the linear combination. The weights α_k are computed with the backpropagation algorithm. With this technique the global pooling layer is not needed anymore, and attention maps can be computed from any layer in any network architecture. More precisely, each weight α_k is computed as the average over all voxels of the derivative of the output $\hat{y}$ with respect to the feature maps f_k of the target convolution layer. In our case, we use the feature maps of the last convolution layer preceding global pooling, and the weights are computed as:

$$\alpha_k = \frac{1}{Z} \sum \frac{\partial \hat{y}}{\partial f_k}, \tag{3.3}$$

where Z is the number of voxels in the feature map f_k. The attention map $M_{Grad-CAM}$ is then computed as a linear combination of the feature maps weighted by the α_k, and upsampled with linear interpolation to compensate the maxpooling layers:

$$M_{Grad-CAM} = \sum_{k}^{N} \alpha_k f_k. \tag{3.4}$$

In their original work, Selvaraju et al. (2017) proposed to compute attention maps from any layer in the network. While this approach has the advantage of generating several explanations for the network's behavior, choosing which layer should be used to compute the global attention of network becomes less obvious and objective. In our experiments, we observed that attention maps computed from the first layers of the network highlight large brain structures, and are not helpful for the detection tasks. To be more comparable to the other approaches, we used the feature maps f_k of the last convolution layer.

2.1.2 Gradient methods

2.1.2.1 Grad Simonyan et al. (2014) proposed to compute attention maps by estimating the gradient of the output with respect to the input image. Gradients are computed with the backpropagation algorithm. This method highlights pixels for which a small change would affect the prediction $\hat{y}$ by a large amount. The attention map M_{Grad} is computed as

$$M_{Grad} = \frac{\partial \hat{y}}{\partial I}. \tag{3.5}$$

2.1.2.2 Guided-backpropagation The attention maps obtained by Grad can highlight fine detail in the input image, but often display noise patterns. This noise mostly results from negative gradients flowing back in the rectified linear unit (ReLU) activations. In theory these negative gradients should relate to negative contributions to the network prediction, in practice they deteriorate attention maps and are believed to interact with positive gradients according to an interference phenomenon (Korbar et al., 2017). With the standard backpropagation algorithm, during the backward pass, ReLU nullifies gradients corresponding to negative entries of the bottom data (input of the ReLU coming from the input to the CNN), but not those that have a negative value in the top layer (which precedes the ReLU during the backward pass). Springenberg et al. (2015) proposed to additionally mask out the values corresponding negative entries of the top gradient in the ReLU activations. This is motivated by the deconvolution approach, which can been seen as a backward pass through the CNN where the information passes in reverse direction through the ReLU activations (Simonyan et al., 2014; Springenberg et al., 2015). Masking out these negative entries from the top layer effectively clears the noise in the attention maps.

2.1.3 Intensity method – for brain datasets only

PVS appear as hyperintense areas in the T2-weighted images. In some regions – especially midbrain, and to some extent basal ganglia – the image intensity can often be discriminative enough and can be used as a crude attention map. We therefore include the raw image intensity as one of the attention maps in our comparison, and, after non-maximum suppression, use the lesion count n predicted using the base architecture (see Section 2.2) to select the threshold.

2.2 Architectures

In total, four architectures were implemented to evaluate all six methods. These architectures are illustrated in Figure 3.2. Grad, Guided-backpropagation, and Grad-CAM use the same neural networks (same architecture and weights), but differ in the computation of the attention maps during inference. The other methods require different architectures, and are trained separately. In the following section, we detail the components of each architecture in 3D.

We perform experiments on 2D CNNs for the MNIST dataset and on 3D CNNs for the brain dataset. The 3D CNNs use 3D convolutional layers with 3x3x3 filters with zero-padding, and 3D maxpooling layers of size 2x2x2. Similarly, the 2D CNNs use 2D convolutional layers with 3x3 filters with zero-padding, and 2D maxpooling layers of size 2x2. The 2D CNNs always use four times fewer features maps than their 3D counterpart to allow faster experimentation. After the last convolution layer, each feature map is projected to a single neuron using global average pooling. These neurons are connected with a fully connected layer to a single neuron indicating the output of network $\hat{y} \in \mathbb{R}$. Rectified linear unit

(ReLU) activations are used after each convolution. We use skip connections by concatenating the feature maps of different layers (and not by summing them).

2.2.0.1 GP-Unet architecture (A in Figure 3.2) GP-Unet architecture is that of small segmentation network, with an encoder and a decoder part. The architecture starts with two convolutional layers with 32 filters each. The output of these two layers is concatenated with the input. Then follows a maxpooling layer and two convolutional layers with 64 filters each. The feature maps preceding and following these two layers are concatenated. In order to combine of features at different scales, these low dimension feature maps are upsampled, concatenated with features maps preceding the maxpooling layer, and given to a convolutional layers of 32 filters. Then follows a global average pooling layer, from which a fully connected layer maps to the output. This architecture is simple (308 705 parameters for the 3D version), fast to train (less than one day on 1070 Nvidia GPU), and allows computing attention maps at the full resolution of the input image.

2.2.0.2 GP-Unet no residual architecture (D in Figure 3.2) The architecture of GP-Unet no residual was proposed by (Dubost et al., 2017). In this work, we only changed the global pooling layer from maximum to average to make comparisons between methods more meaningful. This network is a segmentation network with a downsampling and upsampling path. The downsampling path has two convolutional layers of 32 filters, a maxpooling layer, two convolutional layers of 64 filters, a maxpooling layer, and one convolutional layer of 128 filters. The upsampling path starts with an upsampling layer, concatenates the upsampled feature maps with the features maps preceding the maxpooling layer in the downsampling path, computes a convolutional layer with 64 filters, and repeat this complete process for the last scale of feature maps, with a convolutional layer of 32 filters. After that, comes the global pooling layer, and fully connected layer to a single neuron.

The difference with architecture (A) (Dubost et al., 2017) is that the feature maps are downsampled twice instead of once, and that there are no skip connections between sets of two consecutive convolutions (blockwise skip connection in red in Figure 3.2). Consequently, the last convolution layer does not have access to the input image intensities. We believe these residual connections make the design of GP-Unet more flexible than this architecture, by facilitating for instance the network to directly use the input intensities and locally adjust its predictions. This can be crucial for the correct detection of brain lesions. This architecture has twice more parameters (637 185 parameters for the 3D version) than that of GP-Unet.

2.2.0.3 Gated Attention architecture (B in Figure 3.2) We adapted the architecture of the Gated Attention network proposed by Schlemper et al. (2018) to make it more comparable to the other approaches presented in the current work. Here, the Gated Attention architecture is the same as GP-Unet architecture (A) except for two differences: to merge the feature maps between the two different scales, instead of upsampling, concatenation and convolution, we use the attention gate as described by Schlemper et al. (2018). The other difference is that, in this architecture (B), the downsampled feature maps are also projected to single neurons with global pooling. The neurons corresponding to the two different scales are then aggregated (using concatenation) and connected to the single output neuron with a single fully connected layer. This architecture has 198 580 parameters for the 3D version.

The attention gate computes a normalized internal attention map. In their implementation, Schlemper et al. (2018) proposed a custom normalization to prevent the attention map from becoming too sparse. We did not experience such problems and opted for the standard sigmoid normalization.

Similarly to GP-Unet, Gated Attention computes attention maps at the resolution of the input image. However it combines multi-level information with a more complex process than GP-Unet.

2.2.0.4 Base architecture (C in Figure 3.2) The network architecture used for Grad, Guided-backpropagation, and Grad-CAM is kept as similar as possible to that of GP-Unet for better comparison of methods. It starts with two convolutional layers with 32 filters each. The output of these two layers is concatenated with the input. Then follows a maxpooling layer and two convolutional layers with 64 filters each. The output of these two layers is concatenated with the feature maps following the maxpooling layer, and is given directly to the global average pooling layer. In other words, we apply global pooling to the original image (after maxpooling) and the feature maps after the second convolution at each scale - so on 1+32+64 feature maps. This architecture has shown competitive performance on different types of problems in our experiments (eg. in brain lesions in (Dubost et al., 2019b)). With this architecture, unlike GP-Unet, Grad-CAM produces attention maps at a resolution twice smaller than that of the input image, and could miss small target objects. This architecture has 196 418 parameters for the 3D version.

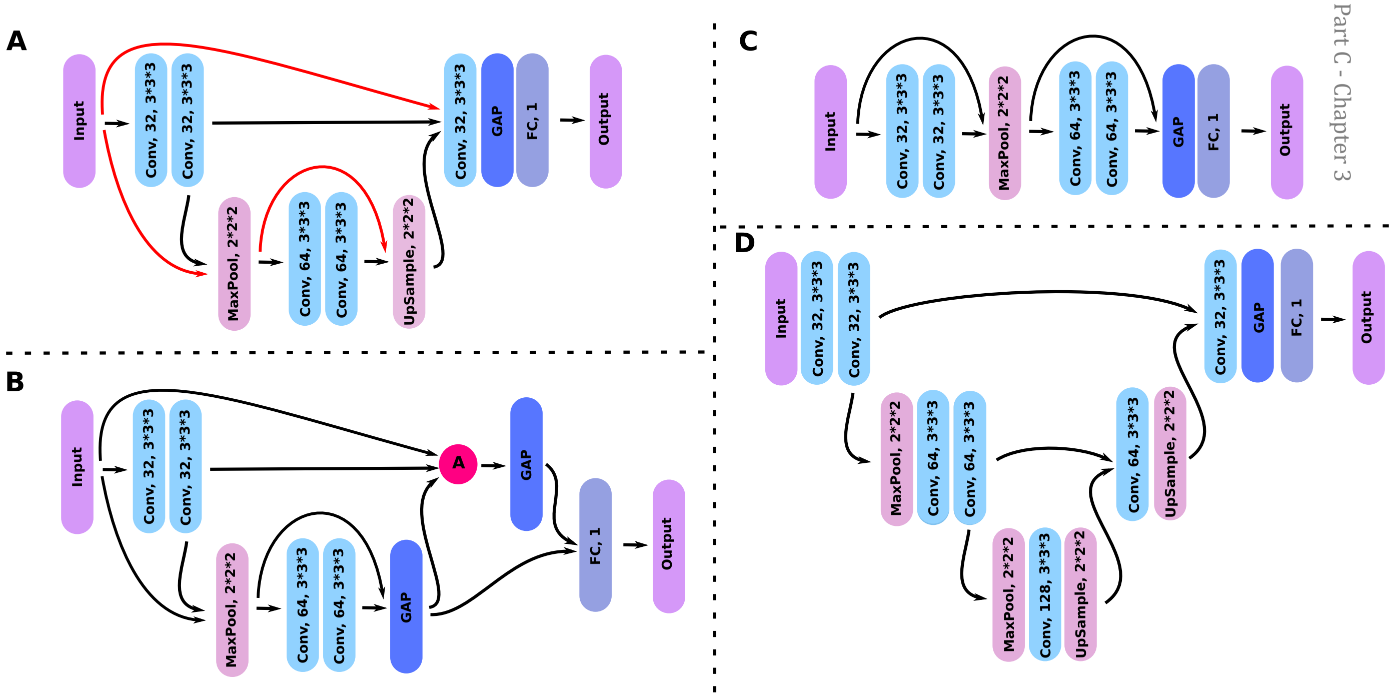


Figure 3.2: **Architectures.** A is GP-Unet's architecture. B is Gated Attention architecture. C is the base architecture used for Grad, Guided-backpropagation, and Grad-CAM. D is GP-Unet no residual architecture. GAP stand for global average pooling layer, FC for fully connected layer, and A for attention gate. All architectures are detailed in Section 2.2. In architecture A, we showed in red the blockwise skip connections.

3 Experiments

In this work, we compare our proposed method to five weakly supervised detection methods. We use the MNIST datasets (LeCun et al., 1998) to compare regression against classification for weak supervision. We compared performance of the different methods – using regression objectives – on weakly supervised lesion detection in a large brain MRI dataset.

3.1 MNIST Datasets

We construct images as a grid of 7 by 5 randomly sampled MNIST digit images. Examples are shown in Figures 3.4 and 3.5. Each digit is uniformly drawn from the set of all training/validation/testing digits, hence with a probability 0.1 to be a target digit d. To avoid class imbalance, we adapt the dataset to each target digit d by sampling 50% of images with no occurrence of d, and 50% of images with at least one occurence of d, resulting in ten different datasets.

3.2 Brain Datasets

Brain MRI was performed on a 1.5-Tesla MRI scanner (GE-Healthcare, Milwaukee, WI, USA) with an eight-channel head coil to obtain 3D T2-contrast magnetic resonance scans. The full imaging protocol has been described by Ikram et al. (2015). In total, our dataset contains 2202 brain scans, each scan being acquired from a different subject.

An expert rater annotated PVS in four brain regions: in the complete midbrain and hippocampi, and in a single slice in axial view in the basal ganglia (the slice showing the anterior commissure) and the centrum semiovale (the slice 10 cm above the top of the lateral ventricle). The annotation protocol follows the

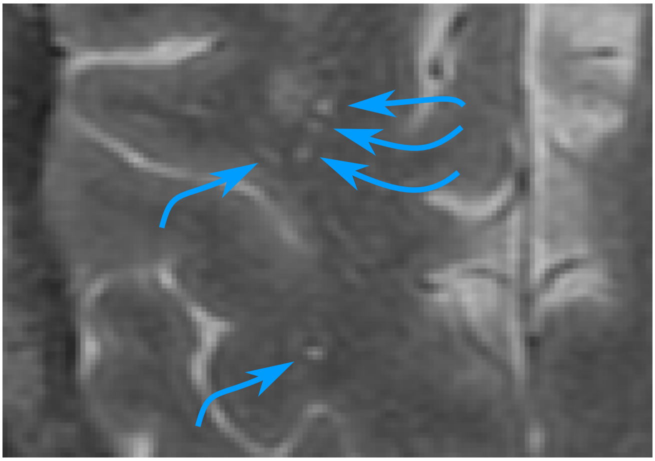


Figure 3.3: **Examples of PVS in the centrum semiovale.** This is a crop of a T2-weighted image in axial view. PVS are indicated with blue arrows.

guidelines by Adams et al. (2014) and Adams et al. (2013) for visual scoring of PVS, with the difference that Adams et al. (2014) only counted the number of PVS, while in the current work, all PVS have been marked with a dot in their center. Figure 3.3 shows examples of PVS in the centrum semiovale.

3.3 Aim of the experiments

In the MNIST datasets, the objective is to detect all occurrences of a target digit d. During optimization, the regression objective is to count the number of occurrences of d, while the classification objective is to detect the presence of at least one occurence of d.

In the experiments on 3D brain MRI scans, the objective is to detect enlarged perivascular spaces (PVS) in the four brain regions described in section 3.2. For these datasets we investigate only regression neural networks. These networks are optimized using the number of annotated PVS in the region of interest as the weak global label, as proposed in our earlier work Dubost et al. (2019b). The location of PVS are only used for the evaluation of the detection during inference.

3.4 Preprocessing

3.4.0.1 MNIST data We scale the image intensity values in the MNIST grid images between zero and one to ease the learning process.

3.4.0.2 Brain scans We first apply the FreeSurfer multi-atlas segmentation algorithm (Desikan et al., 2006a) to locate and mask the midbrain, hippocampi, basal ganglia and centrum semiovale in each scan. For each region, we then extract a fixed volume centered on the center of mass of the region. For midbrain (88x88x11 voxels), hippocampi (168x128x84 voxels) and basal ganglia (168x128x84 voxels) these cropped volumes contain the full region. The centrum semiovale is too large to fit in the memory of our GPU (graphics processing unit), so for this region we only extract the slices surrounding the slice that was scored by the expert rater (250x290x14 voxels). Consequently, we apply a smooth region mask to nullify values corresponding to other brain regions. Finally, we scale the intensity values between zero and one to ease the learning process. The preprocessing and extraction of brain regions is presented in more details in previous work (Dubost et al., 2019b).

3.5 Training of the networks

All regression networks are optimized with Adadelta (Zeiler, 2012) to minimize the mean squared error between their prediction $\hat{y} \in \mathbb{R}$ and the ground truth count $y \in \mathbb{N}$. The classification networks in our MNIST experiments were optimized with Adadelta and the binary cross-entropy loss function.

Weights of the convolution filters and fully connected layers are initialized from a Gaussian distribution with zero mean and unit variance, and biases are initialized to zero.

A validation set is used to prevent over-fitting. The optimization is stopped at least 100 epochs after the validation loss stopped decreasing. We select the model with the lowest validation loss. For the MNIST datasets, the models are trained on a set of 500 images (400 for training and 100 for validation). For the brain datasets, the models are trained on a set of 1202 scans (1000 for training and 202 for validation). During training, we use on-the-fly data augmentation with a random combination of random translations of up to 2 pixels in all directions, random rotations up to 0.2 radians in all directions, and random flipping in all directions. For the MNIST datasets, the batch size was set to 64. For the brain datasets, because of GPU memory constraints, the networks are trained per sample: each mini-batch contains a single 3D image. As the convergence can be slow in some datasets, we first trained the networks on the smallest and easiest region (midbrain), and fine-tune the parameters for the other regions, similarly to Dubost et al. (2019b).

We implemented our algorithms in Python in Keras (Chollet et al., 2015) with TensorFlow as backend, and ran the experiments on a Nvidia GeForce GTX 1070 GPU and Nvidia Tesla K40 [1]. The average training time was one day.

3.6 Negative values in attention maps

Attention maps can have negative values, which meaning can differ for CAM methods and gradient methods. For CAM methods, negative values could highlight objects in the image which presence is negatively associated with the

[1] We used computing resources provided by SurfSara at the Dutch Cartesius cluster.

target objects. For gradient methods, they correspond to areas where increasing the intensity would decrease the predicted count (or where decreasing the intensity would increase the predicted count, these are the same areas).

For image understanding, keeping negative values in attention maps seems most appropriate as the purpose is to discover which parts of the image contributed either negatively or positively to the prediction, and how a change in their intensity could affect the prediction. For detection, the purpose is to find to find all occurrences of the target object in the image and ignore other objects. In the literature, two approaches have been proposed to handle negative values for object detection: either setting them to zero, or taking the absolute value. CAM methods (Zhou et al., 2016; Selvaraju et al., 2017) nullify negative values of the attention maps to mimic the behavior of ReLU activations. Gradient methods (Simonyan et al., 2014; Springenberg et al., 2015) focus on the magnitude of the derivative and thus compute the absolute value.

In our case, we aim to solve a detection problem in datasets where the target objects are among the highest intensity values in the image. For gradient methods, this implies that negatives values in the attention maps do not indicate the location of the target object in our case. We can therefore ignore negative values, and decided to nullify them. For CAM methods, we follow the recommendation of the literature, and also nullify negative values in attention maps. Consequently, we nullified negative values for all methods. Nullifying negative values actually only impacts the visualization of the attention maps, and not the detection metrics, as we select only candidates with highest values in the attention maps (Section 3.7). On the contrary taking the absolute value could increase the number of detections and would impact our detection metrics.

3.7 Performance evaluation

The output of all weakly-supervised detection methods presented in Section 2 are attention maps. We still need to obtain the coordinates of the detections, and evaluate the matching with the ground truth.

After setting negative values to zero (Section 3.6), we apply non-maximum suppression on the attention maps using a 2D (MNIST, centrum semiovale and basal ganglia) or 3D (hippocampi and midbrain) maximum filter of size 6 voxels (which corresponds to 3 mm in axial plane, the maximum size for PVS as defined by Adams et al. (2013) – we used the same value for the MNIST datasets) with 8 neighborhood in 2D or 26 neighborhood in 3D. This results in a set of candidates that we order according to their value in the attention map. The candidates with highest values are considered the most likely to be the target object.

For the basal ganglia and the centrum semiovale, our dataset does not contain full 3D annotations, but only provides annotations for a single 2D slice per scan (see Section 3.2). As annotations were only available in a single slice, we evaluated the attention maps only in the annotated slice, although we can compute attention maps for the complete volume of these regions. For our evaluation we extract the corresponding 2D slice from the attention map prior to post-processing and compute the metrics only for this slice. In case no lesion was annotated, we selected the middle slice of the attention map as a reasonable approximation of the rated slice.

As we aim to solve a detection problem, we need to quantify the matching between two sets of dots: the annotators dots, and the algorithms' predictions. We used the Hungarian algorithm (Kuhn, 1955) to create an optimal one-to-one match between each detected lesion or digit to the closest annotation in the ground truth. For the brain dataset, we counted a positive detection if a detection was within at most 6 voxels from the corresponding point in the ground truth. This corresponds to the maximum diameter of PVS in the axial view, as defined in Adams et al. (2013). For the MNIST datasets, we counted a positive detection if a detection fell inside the 28*28 pixels wide original MNIST image of the target digit.

As the algorithms output candidates with confidence scores, we can compute free-response receiver operating characteristic (FROC) curves (Bandos et al., 2009) that show the trade-off between high sensitivity and the number of false positives, in our case more precisely the average number of false positives per scan (FPavg). To draw these curves, we varied the number of selected candidates. For each network in our experiments, we report the area under the FROC curve (FAUC) computed from 0 to 5 FPavg for MNIST and from 0 to 15 FPavg for brain lesion detection. We also show the standard deviation of the FAUC, computed by bootstrapping the test set.

In addition to the attention maps, the regression networks also predict the number of target objects in the image. For the detection of brain lesions, we use this predicted count rounded to an integer n to select the top-n candidates with highest scores, and compute the corresponding sensitivity and FPavg, and the average number of false negative per scan (FNavg). For statistical significance of difference of FAUCs, we performed a bootstrap hypothesis testing and consider statistical significance for p-value lower than 0.05. For FPavg, FNavg and Sensitivity we performed Wilcoxon tests using p-value lower than 0.05.

3.8 Intra-rater variability of the lesion annotations

Intra-rater variability has been measured in each region using a separate set of 40 MRI scans acquired and annotated with the same protocol. The rater annotated

PVS twice in each scan with two weeks of interval, and in a different random order.

To compute the sensitivity and FPavg for the Intra-rater variability, one of the two series of annotations has to be set as reference to define true positives, positives and false positives. We successively set the first and second series of annotations as reference, leading to two different results. All results for all regions are displayed next to the FROC curves in Figure 3.7.

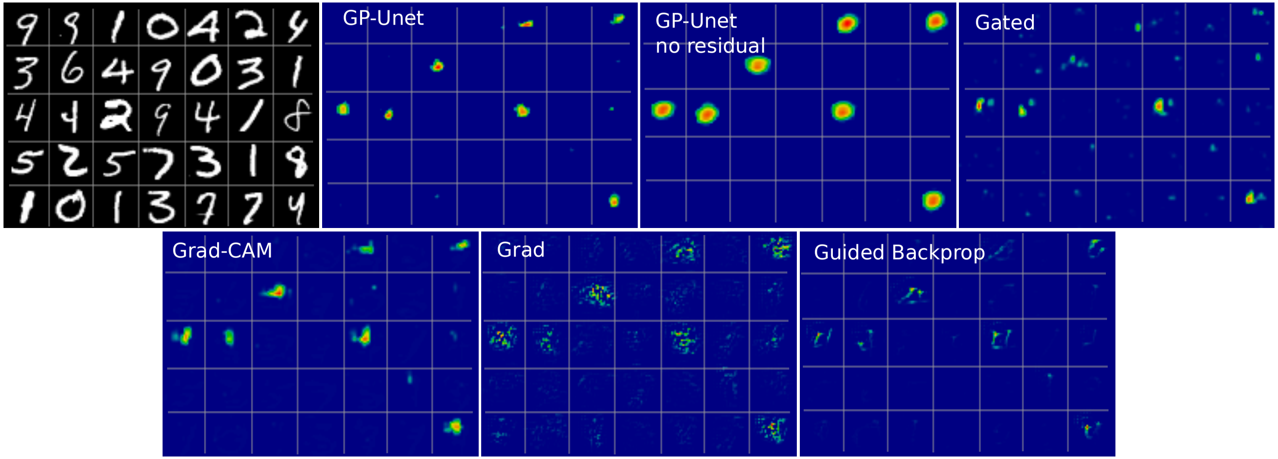


Figure 3.4: **Examples of attention maps of the different weakly supervised detections methods for the detection of digit 4.** Top-left: MNIST image. All methods with optimized with regression objectives.

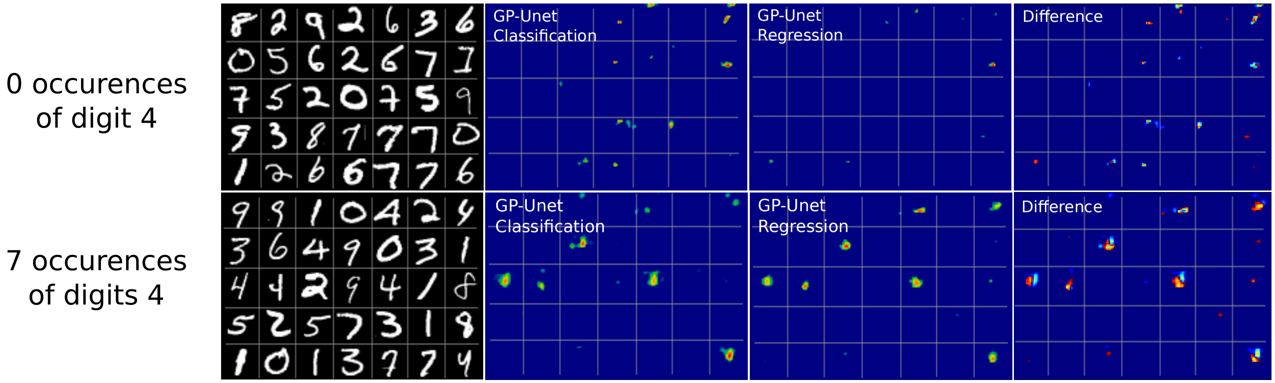


Figure 3.5: **Examples of attention maps of GP-Unet for the detection of digit 4 and optimized with classification and regression objectives.** Left: MNIST image, middle: attention map generated from a classification network, right: attention map generated from a regression network. The first row displays an image without digit 4. The second row displays an image with seven occurences of the digit 4. For the classification method, in the first row we notice more false positives than for the regression method. On the second row, the two digits 4 at the top are less highlighted than the other digits 4 in the image. It is not the case for the regression attention map. This observation supports the hypothesis that attention maps computed from classification objectives tend to focus more on the most obvious occurence of the target object, instead of equally focusing on all occurrences. On the right, we show the difference between the attention maps for regression and classification.

4 Results

4.1 Regression vs classification objectives - MNIST datasets

The methods were evaluated on left-out test sets of 500 images, balanced as described in section 3.1. Figure 3.6 compares the FAUC of regression and classification networks, for all MNIST digits, and for all methods. Additional results such as FROC curves, sensitivity, FPavg and FNavg are given in Appendix A and Appendix B. Overall, regression methods reach a higher detection performance than classification methods. For all digits, regression GP-Unet no residual reaches the best performance. The second best method for all digits is regression GP-Unet. Both GP-Unet regression methods are consistently better than any other method for all digits. Regression Grad-CAM comes third, and regression Guided-backpropagation fourth. Grad and Gated Attention come last. The ordering of best classification methods is different than that of the best (regression) methods: Guided-backpropagation comes first, Grad-CAM second and GP-Unet no residual third.

Figure 3.4 shows an example of the attention maps obtained for all weakly supervised methods optimized with regression objectives. As expected, Grad produces noisy attention maps with many high values, for both classification and regression objectives, and Guided-backpropagation corrects these mistakes. Gradient methods seems to highlight multiple discriminating features of the digit 4 (e.g. its top branches), while CAM methods highlight a single larger, less detailed region. This may suggest that gradients methods may be more suited to weakly supervised segmentation, although judging from the figure, none of the methods seems capable of correctly segmenting digits.

Figure 3.5 compares attention maps of GP-Unet optimized with regression and classification. We noticed two interesting differences. First, when the target digit is present on the image, the regression attention map highlights each occurrence of the target digits with a similar intensity, while the classification attention map highlights more strongly the most obvious occurrences of the target digit. Second, when the target digit is not present in the image, contrary to the regression attention map, the classification attention map may highlight many false positives, possibly resulting in a significant drop in the detection performance.

4.1.0.1 Regression Guided-backpropagation vs Grad. Regression Guided-backpropagation detects of all digits more accurately than regression Grad. The same comparison holds for classification Guided-backpropagation versus classification Grad. However Regression Grad sometimes performs as well (digits 4, 6, 7) or better (digits 0, 9) than Classification Guided-backpropagation, which underlines the added-value of optimizing weakly supervised detection methods with regression objectives instead of classification objectives.

4.2 Variations of the architecture of GP-Unet - MNIST datasets

In this section we studied the influence of the skip connections between sets of two consecutive convolutions (blockwise skip connections, in red in Figure 3.2) in GP-Unet's architecture and the influence of the type of global pooling in GP-Unet's architecture on the detection performance. Removing the blockwise skip connections did not make the detection worse for most digits (except digit 1 and 7 where having the blockwise skip connections helped). Using global max pooling instead of global average pooling led to worse detection performance for all digits. For all digits the optimization was better with the proposed architecture. Removing skip connections or using global max pooling made the optimization take longer to converge, made loss curves not as smooth and made the loss converged to a higher value. The corresponding FROC curves, FAUC barplot, and FAUC, FPavg, FNavg and Sensitivity Tables are given in Appendix C.

4.3 Detection of brain lesions

In the brain dataset, we compare the performance of the weakly supervised methods for the detection enlarged perivascular spaces (PVS) by evaluating them on the left-out test set of 1000 scans, and in four brain regions: midbrain, hippocampi, basal ganglia, and centrum semiovale.

Figures 3.8 - 3.11 show attention maps for all methods in the four regions. Figure 3.7 shows FROC curves for all methods in the brain datasets. Table 3.1 shows the corresponding FAUCs. Table 3.2 and 3.3 show the sensitivity and FPavg measured at the operating point chosen for each method as described in Section 3.7.

Judging from Tables 3.1, 3.2 and 3.3, the methods achieving the best results are GP-Unet, Grad-CAM and Guid-backpropagation. Unlike the results on MNIST datasets, there is no method consistently better than others for all regions. In the midbrain and basal ganglia, Guided-backpropagation reaches the best results of all methods, and in all three metrics, with the exception of FPavg in the basal ganglia. In the hippocampi, GP-Unet reaches the best results of all methods, and in all four metrics. In the centrum semiovale, GP-Unet and Grad-CAM achieve the best results, and have a similar performance. Intensity thresholding reaches a competitive performance in the midbrain and basal ganglia, but completely fails in the hippocampi and centrum semiovale because it highlights many false positives, corresponding to other hyperintense structures. Surrounding cerebrospinal fluid, white matter hyperintensities, and sulci are examples of these structures.

In Figure 3.7, the sensitivity and FPavg between two series of annotations of the same scans from the same rater (green triangle) gives an idea of the difficulty of detecting PVS in each region. In the midbrain and hippocampi, PVS are relatively easy to identify, as they are the only hyperintense lesions visible on T2 images. On the contrary, the detection of PVS in the basal ganglia and centrum semiovale is much more challenging, because in those regions other hyperintense structures that look similar to enlarged perivascular spaces. In all regions, the performance of the automated methods come close to the intra-rater agreement. This intrarater agreement was however computed on a substantially smaller set – 40 vs 1000 scans – and shorter annotation period – 1 week vs several months. Interestingly, several methods highlight the same false positives. After visual checking by experts, many of these false positives appear to be PVS annotated by the rater. In the set of 40 scans used the the intrarater measures, 68 percent of false positive detections of GP-Unet in the centrum semiovale were PVS. More precisely, 39 percent of false positives were enlarged PVS and 29 percent were slightly enlarged PVS.

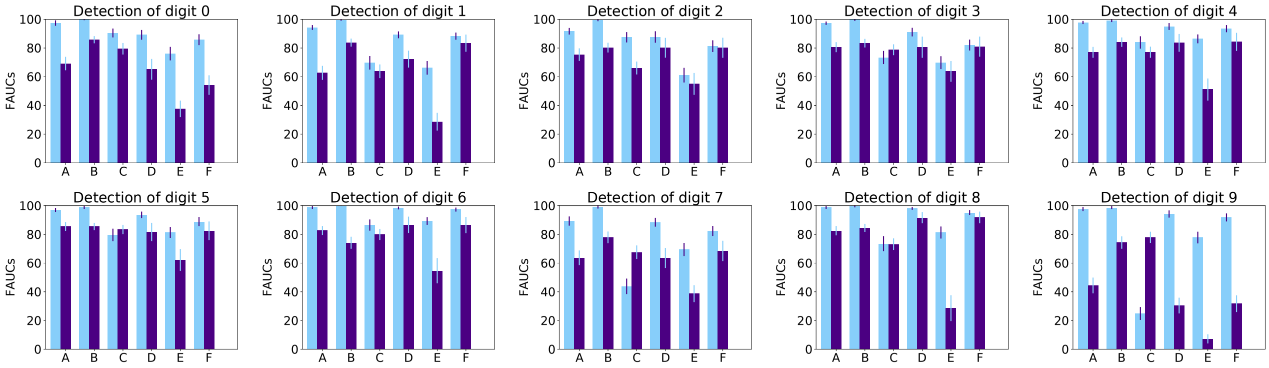


Figure 3.6: **FAUCs (Section 3.7) on the MNIST dataset for all methods.** Each subplot corresponds to the detection of a different digit. Results for regression networks are displayed in light blue, and results for classification networks are displayed in indigo. FAUCs are displayed with standard deviations computed by bootstrapping the test set. A is GP-Unet, B GP-Unet no residual, C Gated Attention, D Grad-CAM, E Grad and F Guided-backpropagation.

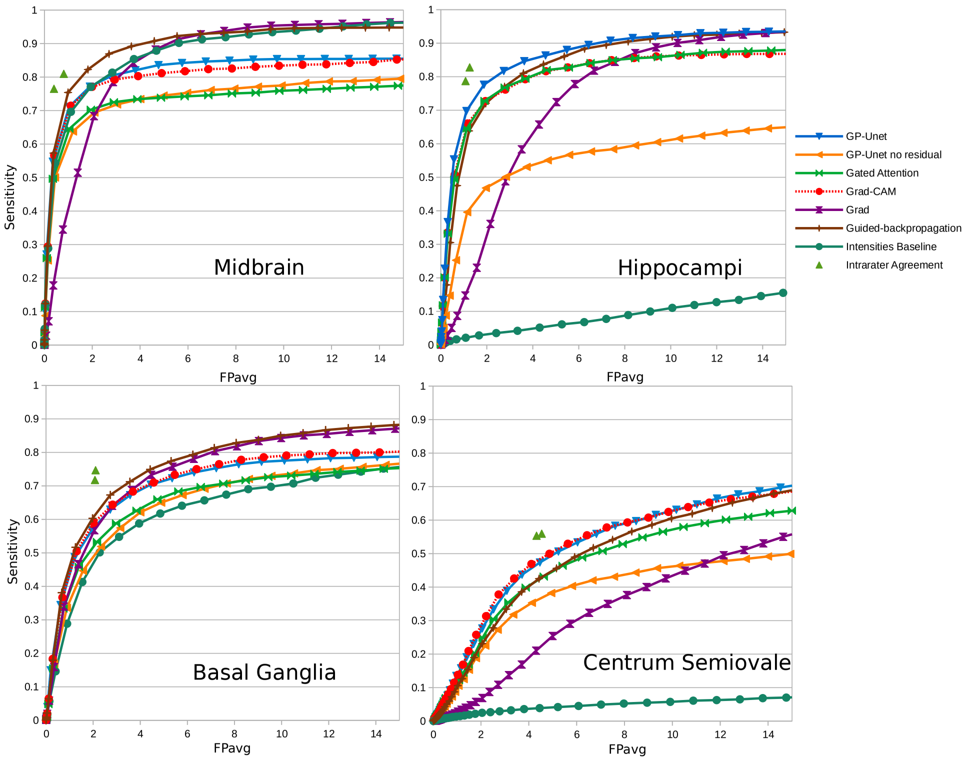


Figure 3.7: **FROC curves of enlarged perivascular spaces detection in the brain MRI in four different regions.** The average number of false positives per scan is displayed on the x-axis, and the sensitivity on the y-axis. Axes have been rescaled for better visibility. The green triangles indicate intra-rater agreement (on a smaller set) as described in Section 3.8 .

Table 3.1: **FAUCs for the detection of brain lesions.** To compute the these FAUCs, we integrate the FROC (Figure 3.7) between 0 and 15 (Section 3.7). The best performance in each region is indicated in bold.

	GP-Unet	GP-Unet no residual	Gated Attention	Grad-CAM	Grad	Guided-backprop	Intensities
	(this paper)	Dubost et al. (2017)	Schlemper et al. (2018)	Selvaraju et al. (2017)	Simonyan et al. (2014)	Springenberg et al. (2015)	Section 4.3
Midbrain	81.5 (80.1 - 82.8)	73.4 (72.0 - 74.8)	72.7 (71.1 - 74.4)	79.8 (78.5 - 81.1)	84.5 (83.5 - 85.4)	**89.2 (88.3 - 90.2)**	87.1 (86.1 - 88.1)
Hippocampi	**85.8 (84.8 - 86.7)**	55.1 (53.5 - 56.7)	80.2 (79.1 - 81.3)	80.1 (78.9 - 81.3)	71.5 (70.4 - 72.6)	83.3 (82.2 - 84.3)	8.3 (7.5 - 9.0)
Basal Ganglia	69.6 (68.1 - 71.2)	64.4 (63.0 - 65.9)	64.8 (63.4 - 66.4)	70.6 (69.3 - 72.0)	**73.5 (72.2 - 74.9)**	**75.6 (74.3 - 76.8)**	61.7 (59.9 - 63.5)
Centrum Semiovale	**51.3 (50.1 - 52.6)**	37.9 (36.8 - 39.2)	46.2 (45.0 - 47.5)	**51.5 (50.2 - 52.7)**	31.9 (30.7 - 33.2)	48.1 (46.9 - 49.3)	4.7 (4.2 - 5.2)
Average	72.0 +/- 13.3	57.7 +/- 13.1	66.0 +/- 12.7	70.5 +/- 11.6	65.4 +/- 19.9	74.1 +/- 15.7	40.5 +/- 35.2

Table 3.2: **Sensitivity in the brain datasets.** Best performance are indicated in bold.

	GP-Unet (this paper)	GP-Unet no residual Dubost et al. (2017)	Gated Attention Schlemper et al. (2018)	Grad-CAM Selvaraju et al. (2017)	Grad Simonyan et al. (2014)	Guided-backprop Springenberg et al. (2015)	Intensities Section 4.3
Midbrain	71.1 (69.5 - 72.7)	63.8 (62.1 - 65.5)	64.6 (62.8 - 66.3)	71.5 (69.8 - 73.1)	51.5 (49.6 - 53.3)	**75.4 (73.8 - 77.0)**	69.6 (67.9 - 71.4)
Hippocampi	**69.8 (68.2 - 71.3)**	46.8 (45.2 - 48.4)	64.6 (62.9 - 66.2)	66.1 (64.5 - 67.6)	36.1 (34.5 - 37.6)	63.8 (62.2 - 65.5)	4.2 (3.6 - 4.8)
Basal Ganglia	56.8 (55.0 - 58.5)	51.9 (50.1 - 53.6)	53.3 (51.6 - 55.0)	58.9 (57.2 - 60.6)	56.8 (55.1 - 58.5)	**60.3 (58.6 - 62.0)**	50.1 (48.3 - 52.0)
Centrum Semiovale	50.6 (49.3 - 52.0)	42.0 (40.7 - 43.4)	48.8 (47.5 - 50.2)	**53.0 (51.6 - 54.3)**	35.0 (33.9 - 36.1)	49.0 (47.7 - 50.3)	5.7 (5.2 - 6.3)
Average	62.1 +/- 8.7	51.1 +/- 8.1	57.8 +/- 6.9	62.4 +/- 7.0	44.8 +/- 9.5	62.1 +/- 9.4	32.4 +/- 28.3

Table 3.3: **Average number of false positives per scan in the brain datasets.** Best performances are indicated in bold.

	GP-Unet (this paper)	GP-Unet no residual Dubost et al. (2017)	Gated Attention Schlemper et al. (2018)	Grad-CAM Selvaraju et al. (2017)	Grad Simonyan et al. (2014)	Guided-backprop Springenberg et al. (2015)	Intensities Section 4.3
Midbrain	**1.03 (0.99 - 1.07)**	1.19 (1.15 - 1.24)	**1.04 (0.99 - 1.09)**	1.10 (1.05 - 1.15)	1.40 (1.34 - 1.45)	**0.99 (0.94 - 1.03)**	1.11 (1.06 - 1.15)
Hippocampi	**1.12 (1.06 - 1.17)**	1.96 (1.88 - 2.03)	**1.13 (1.06 - 1.19)**	**1.16 (1.10 - 1.22)**	2.16 (2.06 - 2.25)	1.23 (1.16 - 1.29)	3.34 (3.22 - 3.45)
Basal Ganglia	**1.95 (1.88 - 2.01)**	2.33 (2.27 - 2.39)	2.16 (2.10 - 2.23)	2.02 (1.95 - 2.09)	2.06 (1.98 - 2.13)	**1.98 (1.91 - 2.04)**	2.28 (2.21 - 2.35)
Centrum Semiovale	**5.24 (5.04 - 5.43)**	6.66 (6.46 - 6.86)	6.23 (6.02 - 6.44)	5.63 (5.44 - 5.82)	7.30 (7.03 - 7.57)	5.92 (5.71 - 6.12)	9.91 (9.62 - 10.21)
Average	2.33 +/- 1.71	3.04 +/- 2.13	2.64 +/- 2.12	2.48 +/- 1.86	3.23 +/- 2.37	2.53 +/- 1.99	4.16 +/- 3.41

Table 3.4: **Average number of false negatives per scan in the brain datasets.** Best performances are indicated in bold.

	GP-Unet (this paper)	GP-Unet no residual Dubost et al. (2017)	Gated Attention Schlemper et al. (2018)	Grad-CAM Selvaraju et al. (2017)	Grad Simonyan et al. (2014)	Guided-backprop Springenberg et al. (2015)	Intensities Section 4.3
Midbrain	0.77 (0.71 - 0.83)	0.98 (0.91 - 1.05)	0.94 (0.87 - 1.00)	0.77 (0.71 - 0.82)	1.06 (1.00 - 1.12)	**0.65 (0.60 - 0.71)**	0.77 (0.72 - 0.83)
Hippocampi	**1.14 (1.07 - 1.22)**	2.12 (2.01 - 2.23)	1.33 (1.25 - 1.41)	1.32 (1.24 - 1.41)	2.32 (2.21 - 2.43)	1.39 (1.31 - 1.47)	3.50 (3.36 - 3.64)
Basal Ganglia	2.00 (1.85 - 2.14)	2.11 (1.97 - 2.25)	2.08 (1.94 - 2.21)	**1.92 (1.78 - 2.06)**	1.96 (1.82 - 2.09)	**1.88 (1.74 - 2.01)**	2.18 (2.03 - 2.33)
Centrum Semiovale	5.83 (5.50 - 6.17)	6.67 (6.30 - 7.03)	5.98 (5.64 - 6.32)	**5.63 (5.30 - 5.96)**	7.30 (6.92 - 7.68)	5.92 (5.58 - 6.26)	9.91 (9.44 - 10.38)
Average	2.44 +/- 2.01	2.97 +/- 2.18	2.58 +/- 2.00	2.41 +/- 1.90	3.16 +/- 2.43	2.46 +/- 2.04	4.09 +/- 3.50

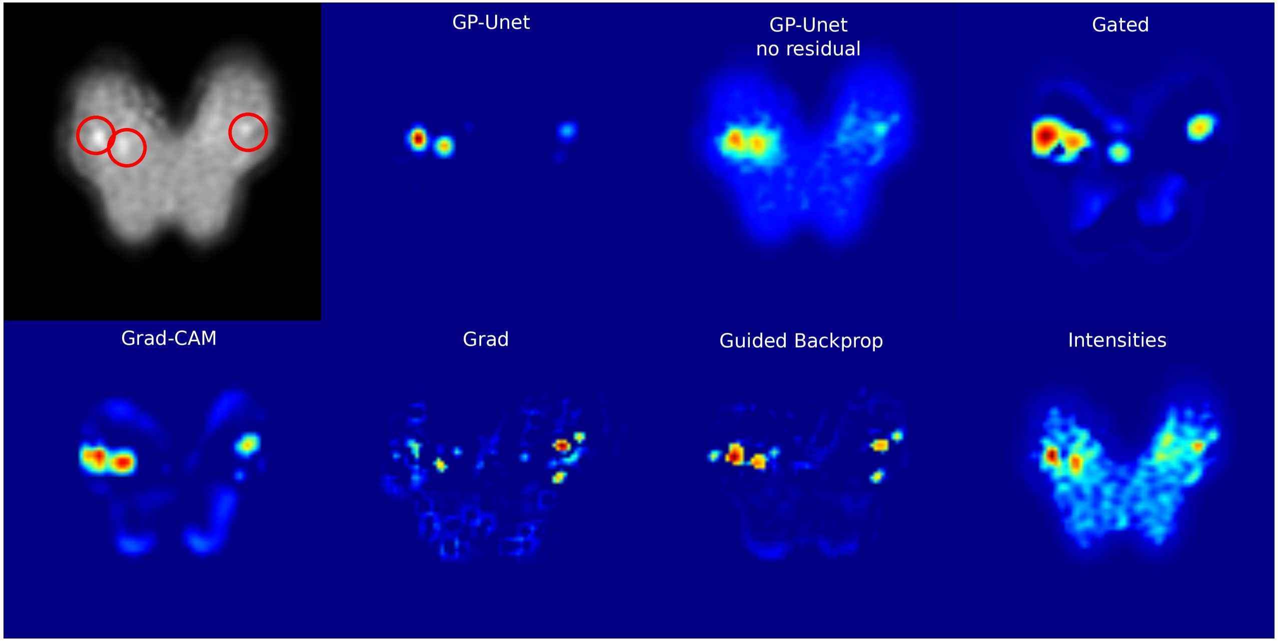


Figure 3.8: **Attentions maps in the midbrain.** The top left image shows the slice of an example image of the midbrain after preprocessing, with PVS indicated with red circles. The other images correspond to attention maps computed for that same slice. Red values correspond to high values in the attention maps. The intensity baseline method in the bottom right corner is actually the same as the image in the upper left corner but with a different color map. Values in attention maps are not bounded, and the maximum varies between images and methods. For the visualization, we chose the scaling of attention maps to best show the range of values in each image.

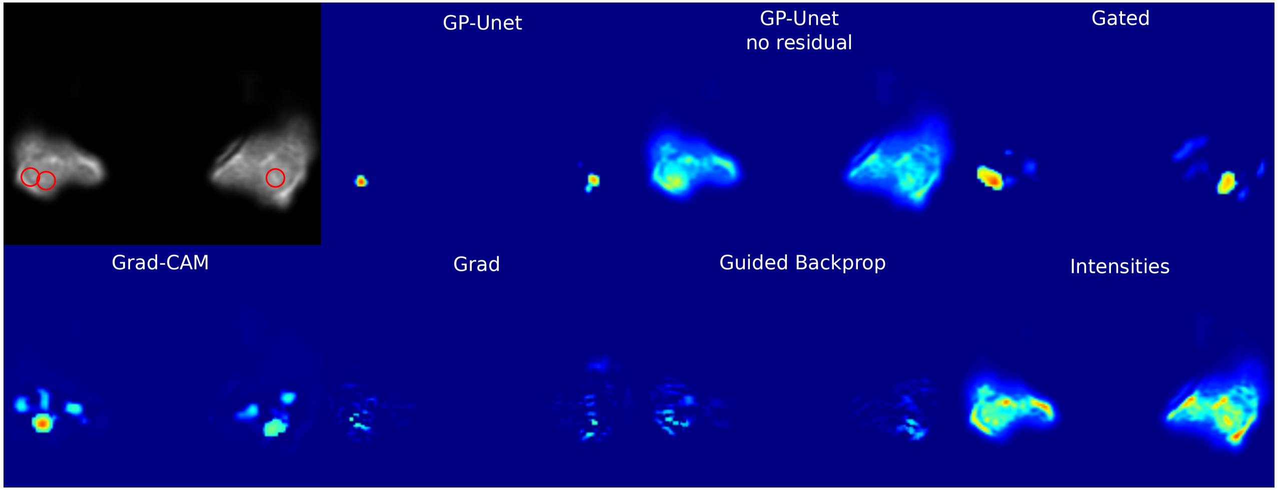


Figure 3.9: **Attentions maps in the hippocampi.**

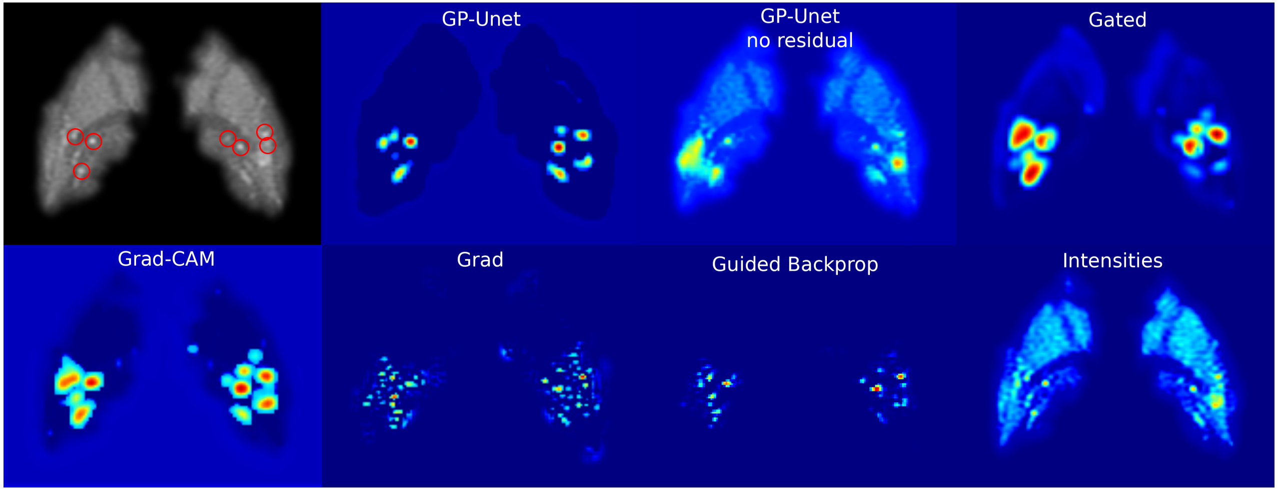


Figure 3.10: **Attentions maps in the basal ganglia.**

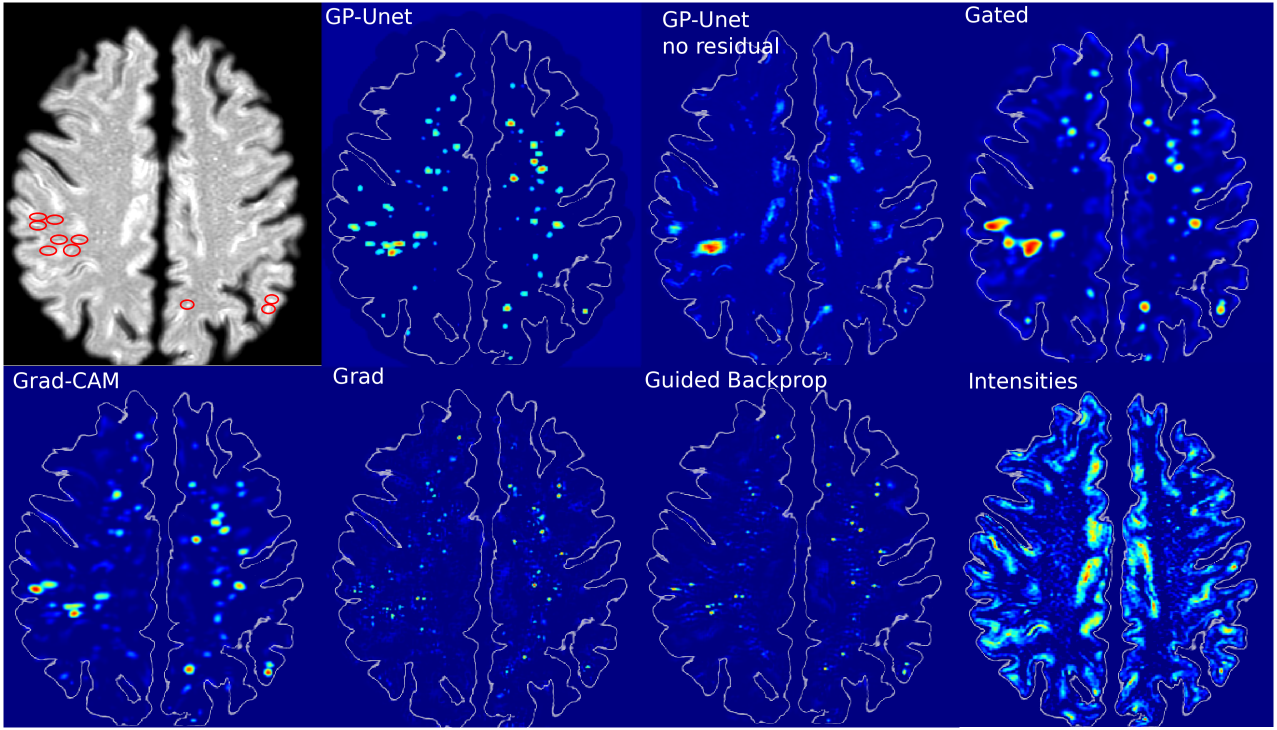


Figure 3.11: **Attentions maps in the centrum semiovale.** Contours of the brain have been delineated in white for better visualization.

5 Discussion

Overall, results showed that weakly supervised methods can detect PVS almost as well as expert raters. The performance of the best detection methods was close to the intrarater agreement. The interrater agreement is also probably lower than this intrarater agreement. Finally, further visual inspection also revealed that many of the false positives correspond to PVS that were not annotated by the human rater. We especially noticed that annotating all PVS was difficult for the expert rater in scans with many PVS.

We compared six weakly supervised detection methods in two datasets. We showed that the proposed method could be used with either 2D or 3D networks. For all methods, 2D networks in the MNIST datasets converged substantially faster (hours) than the 3D networks in the brain dataset (days). In MNIST datasets for regression, GP-Unet no residual (Dubost et al., 2017) and GP-Unet (this article) perform significantly better than all other methods, probably because they can combine the information of different scales more effectively than other methods. For GP-Unet no residual, part of this performance difference can also be explained by the larger number of parameters and larger receptive field (Section 2.2). On the contrary, for GP-Unet, the number of parameters is comparable to that of the other methods. In the brain dataset, the best methods are Guided-backpropagation (Springenberg et al., 2015) with 74.1 average FAUC over regions, GP-Unet with 72.0 average FAUC, and Grad-CAM Selvaraju et al. (2017) with 70.5 average FAUC. As GP-Unet performs either similarly to or better than Grad-CAM depending on the region, given a new weakly supervised detection task, we would consequently recommend Guided-backpropagation and GP-Unet.

Grad-CAM and GP-Unet reach similar FAUCs (Table 3.1) in the basal ganglia and centrum semiovale. However, GP-Unet outperforms Grad-CAM in the midbrain and by a large margin in the hippocampi. In these two regions, at the operating point Grad-CAM suffers from more false positives than GP-Unet, while having a similar or worse sensitivity (Table 3.3 and 3.2). The attention maps of the hippocampi (Figure 3.16) – and to some extent those of the midbrain (Figure 3.8) – show that GP-Unet is less distracted by the surrounding cerebrospinal fluid than Grad-CAM – or the methods emphasizing intensities (GP-Unet no residual, Intensities). The attention maps of Grad-CAM and GP-Unet share most of the false positive detections. Most of these false positives are PVS that were not annotated by the rater. Overall, the attention maps of GP-Unet are also sharper than the ones of Grad-CAM, probably because GP-Unet can compute attention maps at a higher resolution: the resolution of the input image.

The motivation of Gated Attention (Schlemper et al., 2018) is similar to that of GP-Unet: combining multiscale information in the computation of attention maps. In the MNIST datasets, while Gated Attention and GP-Unet reach a similar detection performance when optimized with classification objectives, contrary to GP-Unet, Gated Attention rarely benefits from the regression objective. More generally, Gated Attention seems to benefit less often from the regression objective than the other methods. These results suggest that gate mechanisms may harm the detection performance for networks optimized with regression objectives, and that a simple concatenation of feature maps should be preferred. In the brain datasets, Gated Attention works better than the intensity baseline, Grad (Simonyan et al., 2014), and GP-Unet no residual, but performs significantly worse than Grad-CAM, Guided-backpropagation, and GP-Unet. One should also keep in mind that Gated Attention was originally proposed for deeper networks. In case of shallow networks, this method may not reach its full potential, as it benefits only from few (two on our case) different feature scales.

We mentioned above that the attention maps of GP-Unet are sharper than those of Grad-CAM. In Appendix C, we investigate the influence of the architecture and compare attention maps of GP-Unet, GP-Unet without blockwise skip connections (GP-Unet No Skip) and GP-Unet with global max pooling instead of global average pooling (GP-Unet Max Pool). Removing the skip connections does not seem to make the attention less compact. Using global max pooling does make the attention maps more compact but increases the number of false negatives. GP-Unet may have more compact attention maps than Grad-CAM on the basic architecture thanks to the upsampling path in GP-Unet. To compute the attention at full input resolution with Grad-CAM, the attention maps need to be interpolated, resulting in les compact attention maps. GP-Unet may have more compact attention maps than Gated Attention because concatenating feature maps might be more efficient (maybe easier to optimize) in combining multiscale features than using the gated attention.

Due to the special properties of the PVS detection problem in the brain datasets, intensity thresholding provides a simple approach to solving the same problem. Although intensity thresholding yields the worst results in hippocampi, basal ganglia, and centrum semiovale, it achieves the second best FAUC in the midbrain. This high performance results from the effective region masking specific to the midbrain: because PVS are almost always in the center of this region, we can erode the border of the region mask, and eliminate the hyperintense cerebrospinal fluid surrounding the midbrain. As there are no other visible lesions in the midbrain, all remaining hyperintensities correspond to PVS.

In the datasets where the intensity method achieved good or reasonable results (midbrain and basal ganglia), Guided-backpropagation performed best. In the datasets where the intensity method failed (hippocampi and centrum semiovale), GP-Unet reached the best performance (similar to that of Grad-CAM in the centrum semiovale). More generally, gradients methods seem to work best when the target objects are also the most salient objects, while CAM methods work best when saliency alone is not discriminative enough. This observation can also be extended to the MNIST datasets, where saliency alone is not sufficient, and regression CAM methods (Gated Attention excluded) outperform regression gradient methods.

Recently Adebayo et al. (2018) showed that, for Guided-backpropagation, classification networks trained with random labels obtained similar attention maps as networks trained with the correct labels, hinting that attention maps method may focus more on salient objects in the image than the target object. In these experiments, attention maps computed with Grad and Grad-CAM obtained better results. Adebayo et al. warn of the evaluation of attention maps by only visual appeal, and advocate more rigorous forms of evaluation. This fits exactly with the purpose of the current article, in which we aimed to quantify the detection performance of attention maps in large real world datasets.

For the evaluation of the detection of PVS, images were annotated by a single rater. With the same resources, we could also have had multiple raters annotating fewer scans and use their consensus for the evaluation, which may reduce the risk of mislabeling. We preferred to evaluate the detection using more scans to better encompass the anatomical variability, and we quantified the performance of the single rater by computing her intra-rater agreement on a smaller set.

In our preliminary work on PVS detection in the basal ganglia using GP-Unet no residual (Dubost et al., 2017) we obtained slightly different results than what is presented in the current work. This reflects differences in the test data set, the annotations, method and postprocessing. Our previous annotations (Dubost et al., 2017) were done directly on the segmented and cropped basal ganglia, while the annotations of the current work were done on the full scan. The rater sometimes annotated lesions at the borders of the basal ganglia which are barely visible after preprocessing. In addition, the current work also includes scans without annotations (because the rater found no lesion), where there could have been errors in finding the slice evaluated by the rater. In the current work, Grad reaches better results than in Dubost et al. (2017), because it benefits from the more sophisticated postprocessing: the non-maximum suppression clears the noise in the attention maps.

Next to the methods presented in this paper, we experimented with the perturbation method with masks proposed by Petsiuk et al. (2018). For this method, masks are first sampled in a low dimensional space and resized to the size of the input image. It appeared that the size of this lower dimensional needs to be adapted to the size of the target object in the image. If the target objects are small, one may need to sample relatively large masks. In our experiments, we experimented with a range of values for the size of this low dimensional space, and did not manage to compute discriminative attention maps for PVS, that are small objects relatively to the image resolution.

The work presented in this article implies that pixel-level annotations may not be needed to train accurate models for detection problems. This is especially relevant in medical imaging, where annotation requires expert knowledge and high quality annotations are therefore difficult to obtain. Weakly supervised methods enable learning from large databases, such as UK biobank (Sudlow et al., 2015) or Framingham study (Maillard et al., 2016), with less annotation effort, and could also help to reduce the dependence on annotator biases. The global label may even be more reliable, because for some abnormalities raters can agree well on the presence or global burden of the abnormalities but poorly on their boundaries or spatial distribution.

The variety of challenges present in the brain datasets are well suited to the evaluation of weakly-supervised detection methods. Observations and results might generalize to the detection of other types of small objects, such as microinfarcts, microbleeds, or small white matter hyperintensities.

6 Conclusion

We proposed a new weakly supervised detection method, GP-Unet, that uses an encoder-decoder architecture optimized only with global labels such as the count of lesions in a brain region. The decoder part upsamples feature maps and enables the computation of attention maps at the resolution of the input image, which thus helps the detection of small objects. We also showed the advantage of using regression objectives over classification objectives for the optimization of weakly supervised detection methods, when the target object appears multiple times in the image. We compared the proposed method to four state-of-the-art methods on the detection of digits in MNIST-based datasets, and on the detection of enlarged perivascular spaces – a type of brain lesion – from 3D brain MRI. The best weakly supervised detection methods were Guided-backpropagation (Springenberg et al., 2015), and the proposed method GP-Unet. We noticed that methods based on the gradient of the output of the network, such as Guided-backpropagation, worked best in datasets where the target objects are also the most salient objects. In other datasets, methods using class activation maps, such as GP-Unet, worked best. The performance of the detection enlarged perivascular spaces using the weakly supervised methods was close to the intrarater agreement of an expert rater. The proposed method could consequently facilitate studies of enlarged perivascular and help advance research in their etiology and relationship with cerebrovascular diseases.

Appendix A. Results MNIST – Regression Objectives.

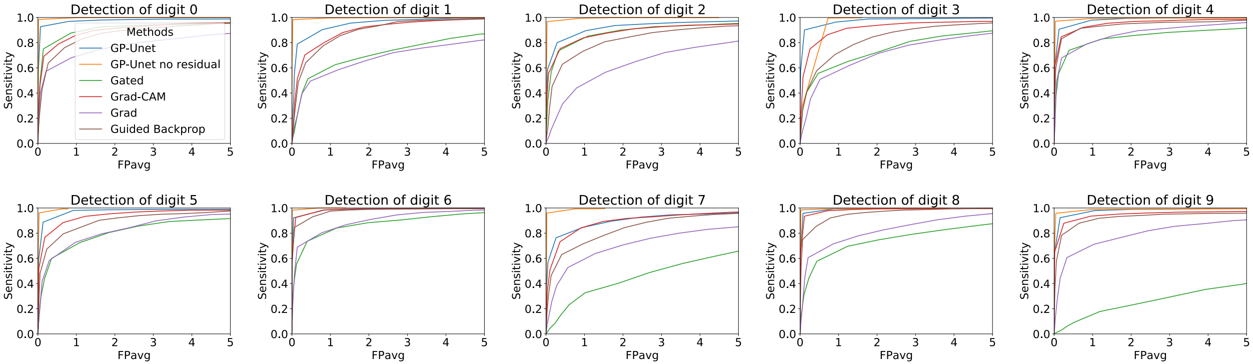


Figure 3.12: **FROC MNIST regression.** Each subplot corresponds to the detection of a different digit.

Table 3.5: **FAUCs MNIST regression.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no residual	Gated Attention	Grad-CAM	Grad	Guided-backprop
	(this paper)	Dubost et al. (2017)	Schlemper et al. (2018)	Selvaraju et al. (2017)	Simonyan et al. (2014)	Springenberg et al. (2015)
0	97.4 (96.4 - 98.4)	**99.7 (99.5 - 99.8)**	90.5 (88.9 - 92.0)	89.2 (87.4 - 90.9)	76.1 (73.8 - 78.3)	85.8 (83.8 - 87.5)
1	94.4 (93.5 - 95.1)	**99.5 (99.3 - 99.7)**	69.7 (67.4 - 72.0)	89.1 (87.8 - 90.4)	66.3 (64.0 - 68.5)	88.3 (86.9 - 89.5)
2	91.7 (90.5 - 92.8)	**99.3 (99.0 - 99.5)**	87.5 (85.8 - 89.3)	87.8 (85.8 - 89.6)	61.0 (58.5 - 63.6)	81.3 (79.0 - 83.3)
3	97.3 (96.6 - 97.9)	**99.6 (99.2 - 99.9)**	73.4 (70.9 - 75.6)	91.2 (89.7 - 92.6)	69.8 (67.4 - 72.0)	82.1 (80.0 - 84.1)
4	97.8 (97.2 - 98.3)	**99.0 (98.5 - 99.5)**	83.9 (81.8 - 86.0)	95.0 (93.8 - 96.0)	86.4 (84.8 - 87.9)	93.5 (92.1 - 94.6)
5	97.1 (96.3 - 97.8)	**98.9 (98.4 - 99.4)**	79.6 (77.4 - 81.8)	93.6 (92.4 - 94.7)	81.5 (79.7 - 83.3)	88.9 (87.2 - 90.5)
6	98.6 (98.2 - 99.0)	**99.9 (99.8 - 99.9)**	86.5 (84.4 - 88.5)	98.4 (97.8 - 98.8)	89.3 (87.9 - 90.6)	97.3 (96.6 - 97.9)
7	89.3 (87.6 - 91.0)	**99.1 (98.5 - 99.6)**	43.9 (41.2 - 46.5)	88.5 (86.9 - 90.0)	69.6 (67.3 - 71.9)	82.4 (80.7 - 84.0)
8	**98.8 (98.3 - 99.2)**	**99.5 (99.1 - 99.8)**	73.6 (70.8 - 76.2)	98.2 (97.7 - 98.7)	81.3 (79.1 - 83.3)	95.2 (94.5 - 96.0)
9	**97.6 (96.8 - 98.3)**	**98.6 (98.1 - 99.1)**	24.8 (22.6 - 27.1)	94.3 (93.0 - 95.5)	77.9 (75.8 - 80.0)	91.8 (90.4 - 93.2)
Average	96.0 +/- 3.0	99.3 +/- 0.4	71.3 +/- 20.0	92.5 +/- 3.7	75.9 +/- 8.6	88.7 +/- 5.4

Table 3.6: **Sensitivity MNIST regression.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no residual	Gated Attention	Grad-CAM	Grad	Guided-backprop
	(this paper)	Dubost et al. (2017)	Schlemper et al. (2018)	Selvaraju et al. (2017)	Simonyan et al. (2014)	Springenberg et al. (2015)
0	92.7 (91.4 - 93.9)	**98.4 (97.8 - 99.1)**	75.1 (73.1 - 77.0)	69.2 (67.2 - 71.3)	57.3 (55.3 - 59.2)	63.8 (61.7 - 65.8)
1	78.9 (77.3 - 80.4)	**98.3 (97.8 - 98.8)**	51.4 (49.4 - 53.4)	70.3 (68.6 - 72.1)	49.5 (47.7 - 51.2)	64.2 (62.5 - 65.8)
2	80.0 (78.2 - 81.8)	**96.7 (95.9 - 97.5)**	73.5 (71.5 - 75.5)	75.3 (73.4 - 77.2)	44.0 (41.9 - 46.1)	62.8 (60.9 - 64.7)
3	90.1 (88.8 - 91.5)	**97.9 (97.4 - 98.4)**	55.6 (53.5 - 57.7)	75.0 (73.1 - 76.9)	50.9 (48.9 - 52.9)	57.7 (55.5 - 59.9)
4	90.7 (89.3 - 92.1)	**97.0 (96.3 - 97.8)**	73.9 (71.8 - 76.0)	84.6 (83.0 - 86.3)	67.9 (66.0 - 69.7)	83.1 (81.4 - 84.8)
5	88.7 (87.2 - 90.2)	**96.1 (95.2 - 97.0)**	60.0 (57.8 - 62.2)	76.7 (74.6 - 78.7)	58.4 (56.2 - 60.5)	67.3 (65.2 - 69.4)
6	92.2 (91.0 - 93.5)	**98.3 (97.7 - 98.9)**	73.6 (71.5 - 75.6)	92.4 (91.1 - 93.7)	68.9 (67.1 - 70.7)	84.6 (82.9 - 86.2)
7	76.3 (74.6 - 78.1)	**95.9 (94.9 - 97.0)**	32.7 (30.7 - 34.6)	73.2 (71.4 - 75.0)	52.7 (50.7 - 54.6)	63.0 (61.2 - 64.8)
8	95.8 (95.0 - 96.5)	**98.5 (98.0 - 98.9)**	57.7 (55.5 - 59.9)	93.5 (92.5 - 94.4)	60.5 (58.7 - 62.4)	75.0 (73.3 - 76.7)
9	92.3 (91.1 - 93.5)	**95.8 (95.0 - 96.6)**	17.8 (16.2 - 19.3)	87.9 (86.5 - 89.3)	60.6 (58.7 - 62.5)	78.3 (76.6 - 80.0)
Average	87.8 +/- 6.4	97.3 +/- 1.0	57.1 +/- 18.2	79.8 +/- 8.6	57.1 +/- 7.5	70.0 +/- 9.0

Table 3.7: **FPavg MNIST regression.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no residual	Gated Attention	Grad-CAM	Grad	Guided-backprop
	(this paper)	Dubost et al. (2017)	Schlemper et al. (2018)	Selvaraju et al. (2017)	Simonyan et al. (2014)	Springenberg et al. (2015)
0	0.07 (0.05 - 0.08)	**0.02 (0.01 - 0.03)**	0.14 (0.11 - 0.17)	0.16 (0.13 - 0.18)	0.21 (0.18 - 0.24)	0.27 (0.23 - 0.31)
1	0.14 (0.11 - 0.16)	**0.02 (0.01 - 0.03)**	0.42 (0.36 - 0.47)	0.34 (0.30 - 0.38)	0.48 (0.42 - 0.53)	0.36 (0.31 - 0.40)
2	0.29 (0.25 - 0.32)	**0.03 (0.01 - 0.04)**	0.32 (0.28 - 0.36)	0.37 (0.33 - 0.41)	0.79 (0.73 - 0.86)	0.42 (0.38 - 0.47)
3	0.11 (0.09 - 0.14)	**0.00 (0.00 - 0.00)**	0.47 (0.42 - 0.52)	0.26 (0.22 - 0.29)	0.52 (0.46 - 0.58)	0.43 (0.38 - 0.48)
4	0.13 (0.10 - 0.15)	**0.03 (0.02 - 0.04)**	0.39 (0.34 - 0.43)	0.18 (0.15 - 0.21)	0.19 (0.16 - 0.22)	0.19 (0.16 - 0.22)
5	0.13 (0.11 - 0.16)	**0.03 (0.02 - 0.05)**	0.36 (0.31 - 0.40)	0.18 (0.15 - 0.21)	0.30 (0.25 - 0.34)	0.23 (0.20 - 0.26)
6	0.04 (0.03 - 0.06)	**0.01 (0.00 - 0.01)**	0.40 (0.35 - 0.44)	0.10 (0.08 - 0.12)	0.14 (0.11 - 0.17)	0.06 (0.05 - 0.08)
7	0.26 (0.23 - 0.29)	**0.02 (0.01 - 0.03)**	1.02 (0.93 - 1.11)	0.36 (0.32 - 0.41)	0.57 (0.51 - 0.62)	0.41 (0.36 - 0.45)
8	0.08 (0.06 - 0.10)	**0.02 (0.01 - 0.03)**	0.44 (0.38 - 0.49)	0.11 (0.09 - 0.14)	0.21 (0.17 - 0.25)	0.07 (0.05 - 0.09)
9	0.16 (0.13 - 0.18)	**0.04 (0.03 - 0.06)**	1.19 (1.07 - 1.30)	0.25 (0.22 - 0.28)	0.33 (0.29 - 0.38)	0.19 (0.16 - 0.23)
Average	0.14 +/- 0.08	0.02 +/- 0.01	0.51 +/- 0.31	0.23 +/- 0.10	0.37 +/- 0.20	0.26 +/- 0.13

Table 3.8: **FNavg MNIST regression.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no residual	Gated Attention	Grad-CAM	Grad	Guided-backprop
	(this paper)	Dubost et al. (2017)	Schlemper et al. (2018)	Selvaraju et al. (2017)	Simonyan et al. (2014)	Springenberg et al. (2015)
0	0.11 (0.08 - 0.14)	**0.02 (0.01 - 0.03)**	0.37 (0.32 - 0.43)	0.47 (0.41 - 0.53)	0.62 (0.55 - 0.69)	0.54 (0.47 - 0.61)
1	0.35 (0.30 - 0.40)	**0.03 (0.02 - 0.05)**	0.75 (0.67 - 0.82)	0.49 (0.43 - 0.56)	0.82 (0.73 - 0.91)	0.58 (0.52 - 0.64)
2	0.30 (0.25 - 0.34)	**0.05 (0.03 - 0.07)**	0.40 (0.34 - 0.45)	0.37 (0.32 - 0.43)	0.82 (0.73 - 0.90)	0.56 (0.50 - 0.63)
3	0.15 (0.12 - 0.18)	**0.04 (0.02 - 0.05)**	0.58 (0.52 - 0.65)	0.34 (0.30 - 0.39)	0.68 (0.61 - 0.76)	0.56 (0.49 - 0.62)
4	0.12 (0.10 - 0.15)	**0.05 (0.03 - 0.06)**	0.36 (0.31 - 0.41)	0.22 (0.18 - 0.25)	0.49 (0.43 - 0.56)	0.25 (0.21 - 0.29)
5	0.14 (0.11 - 0.16)	**0.05 (0.04 - 0.07)**	0.50 (0.44 - 0.56)	0.29 (0.25 - 0.34)	0.54 (0.48 - 0.60)	0.42 (0.37 - 0.47)
6	0.11 (0.08 - 0.13)	**0.02 (0.01 - 0.03)**	0.35 (0.31 - 0.40)	0.09 (0.07 - 0.12)	0.47 (0.41 - 0.52)	0.22 (0.18 - 0.26)
7	0.36 (0.31 - 0.41)	**0.04 (0.03 - 0.06)**	1.00 (0.90 - 1.10)	0.43 (0.37 - 0.49)	0.73 (0.65 - 0.81)	0.58 (0.51 - 0.65)
8	0.07 (0.05 - 0.09)	**0.02 (0.01 - 0.04)**	0.58 (0.50 - 0.65)	0.11 (0.08 - 0.13)	0.59 (0.52 - 0.66)	0.37 (0.32 - 0.43)
9	0.11 (0.09 - 0.14)	**0.07 (0.05 - 0.09)**	1.22 (1.09 - 1.34)	0.20 (0.16 - 0.24)	0.63 (0.55 - 0.70)	0.36 (0.31 - 0.41)
Average	0.18 +/- 0.10	0.04 +/- 0.02	0.61 +/- 0.28	0.30 +/- 0.14	0.64 +/- 0.12	0.44 +/- 0.13

Appendix B. Results MNIST – Classification Objectives.

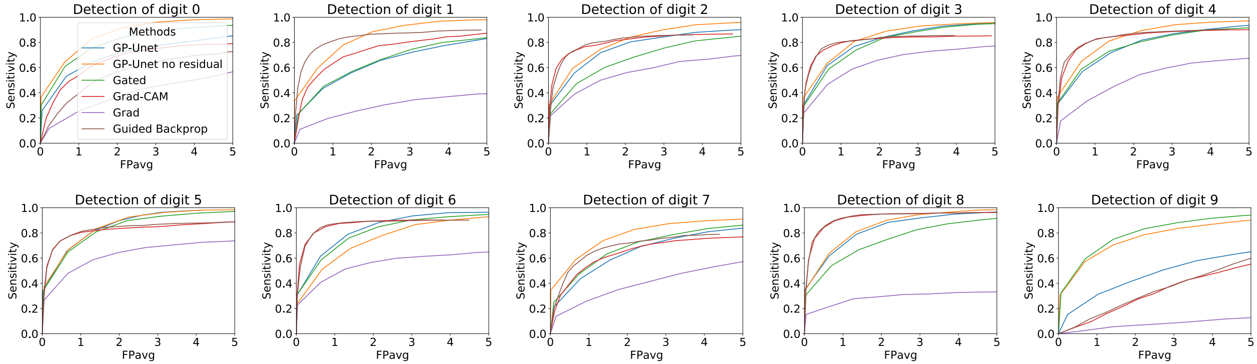


Figure 3.13: **FROC MNIST classification.** Each subplot corresponds to the detection of a different digit.

Table 3.9: **FAUCs MNIST classification.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet (this paper)	GP-Unet no residual Dubost et al. (2017)	Gated Attention Schlemper et al. (2018)	Grad-CAM Selvaraju et al. (2017)	Grad Simonyan et al. (2014)	Guided-backprop Springenberg et al. (2015)
0	69.3 (66.7 - 71.7)	**85.8 (84.6 - 87.1)**	79.5 (77.4 - 81.6)	65.2 (61.8 - 68.6)	37.5 (34.4 - 40.6)	54.1 (50.7 - 57.7)
1	62.6 (60.3 - 65.0)	**83.9 (82.5 - 85.3)**	63.8 (61.3 - 66.3)	72.3 (69.5 - 75.3)	28.6 (25.6 - 31.7)	**83.2 (80.3 - 86.1)**
2	75.5 (73.4 - 77.5)	**80.3 (78.7 - 81.9)**	65.9 (63.7 - 68.3)	**80.3 (76.9 - 83.6)**	54.9 (50.7 - 58.6)	**80.1 (76.5 - 83.5)**
3	80.6 (79.0 - 82.2)	**83.4 (81.7 - 84.9)**	78.9 (77.0 - 80.7)	**80.7 (77.0 - 84.3)**	63.7 (59.9 - 67.4)	**80.9 (77.2 - 84.5)**
4	77.1 (74.9 - 79.1)	**84.1 (82.6 - 85.7)**	77.1 (75.2 - 79.0)	**83.7 (80.7 - 86.5)**	51.3 (47.4 - 55.0)	**84.3 (81.1 - 87.3)**
5	**85.5 (83.9 - 86.8)**	**85.5 (84.3 - 86.7)**	**83.4 (81.9 - 84.9)**	**81.7 (78.5 - 84.7)**	62.0 (58.3 - 65.8)	**82.5 (79.0 - 85.8)**
6	82.8 (81.1 - 84.3)	74.1 (71.9 - 76.2)	80.0 (77.8 - 81.9)	**86.6 (83.7 - 89.3)**	54.6 (50.4 - 58.9)	**86.4 (83.6 - 89.2)**
7	63.7 (61.1 - 66.2)	**77.9 (75.9 - 79.9)**	67.7 (65.4 - 70.0)	63.5 (60.1 - 67.0)	38.4 (35.8 - 41.2)	68.4 (64.7 - 72.1)
8	82.7 (81.0 - 84.2)	84.7 (83.2 - 86.0)	73.2 (71.1 - 75.3)	**91.6 (89.4 - 93.6)**	28.6 (24.1 - 33.1)	**91.9 (89.9 - 93.7)**
9	44.3 (41.5 - 47.3)	74.3 (72.2 - 76.5)	**78.1 (76.0 - 80.0)**	30.4 (27.6 - 33.3)	7.2 (5.7 - 8.8)	31.7 (28.8 - 34.7)
Average	72.4 +/- 12.0	81.4 +/- 4.3	74.8 +/- 6.4	73.6 +/- 16.7	42.7 +/- 17.0	74.4 +/- 17.4

Table 3.10: **Sensitivity MNIST classification.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no residual	Gated Attention	Grad-CAM	Grad	Guided-backprop
	(this paper)	Dubost et al. (2017)	Schlemper et al. (2018)	Selvaraju et al. (2017)	Simonyan et al. (2014)	Springenberg et al. (2015)
	26.0 (24.7 - 27.4)	**36.7 (34.9 - 38.5)**	30.3 (28.7 - 31.9)	19.4 (18.0 - 20.8)	11.8 (10.6 - 12.9)	13.6 (12.3 - 14.9)
1	22.3 (21.0 - 23.6)	**33.5 (31.9 - 35.1)**	23.4 (22.1 - 24.7)	19.4 (18.2 - 20.6)	11.0 (9.8 - 12.2)	25.9 (24.6 - 27.3)
2	**30.1 (28.6 - 31.7)**	**30.4 (28.8 - 31.9)**	23.7 (22.3 - 25.1)	27.3 (25.8 - 28.8)	21.6 (20.0 - 23.3)	25.0 (23.5 - 26.5)
3	30.3 (28.8 - 31.7)	**35.1 (33.5 - 36.8)**	29.8 (28.4 - 31.2)	28.2 (26.7 - 29.6)	23.9 (22.3 - 25.6)	26.9 (25.5 - 28.3)
4	30.8 (29.2 - 32.4)	**36.5 (34.8 - 38.3)**	31.2 (29.6 - 32.8)	29.6 (28.1 - 31.1)	17.6 (15.9 - 19.3)	30.8 (29.2 - 32.3)
5	34.1 (32.5 - 35.8)	**37.4 (35.6 - 39.1)**	34.4 (32.8 - 36.1)	31.4 (29.8 - 33.0)	26.7 (25.0 - 28.4)	30.1 (28.5 - 31.6)
6	**30.9 (29.4 - 32.3)**	23.9 (22.7 - 25.0)	**31.0 (29.6 - 32.5)**	**31.3 (29.7 - 32.8)**	22.9 (21.1 - 24.7)	**30.8 (29.3 - 32.4)**
7	21.6 (20.3 - 22.8)	**34.6 (32.9 - 36.3)**	25.0 (23.5 - 26.5)	17.5 (16.2 - 18.8)	13.9 (12.7 - 15.0)	19.1 (17.9 - 20.3)
8	**34.7 (33.1 - 36.4)**	**34.0 (32.4 - 35.6)**	30.7 (29.2 - 32.2)	**33.5 (31.9 - 35.1)**	15.3 (13.3 - 17.3)	**33.7 (32.1 - 35.3)**
9	15.2 (13.9 - 16.6)	**30.7 (29.2 - 32.2)**	**32.4 (30.8 - 34.0)**	4.9 (4.1 - 5.7)	1.7 (1.1 - 2.2)	5.1 (4.3 - 6.0)
Average	27.6 +/- 5.9	33.3 +/- 3.9	29.2 +/- 3.6	24.2 +/- 8.4	16.6 +/- 7.1	24.1 +/- 8.5

Table 3.11: **FPavg MNIST classification.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet (this paper)	GP-Unet no residual Dubost et al. (2017)	Gated Attention Schlemper et al. (2018)	Grad-CAM Selvaraju et al. (2017)	Grad Simonyan et al. (2014)	Guided-backprop Springenberg et al. (2015)
0	0.05 (0.03 - 0.06)	**0.02 (0.01 - 0.03)**	**0.02 (0.01 - 0.04)**	0.15 (0.12 - 0.17)	0.23 (0.20 - 0.26)	0.21 (0.18 - 0.24)
1	0.06 (0.05 - 0.08)	**0.00 (0.00 - 0.01)**	0.11 (0.08 - 0.13)	0.10 (0.08 - 0.13)	0.14 (0.12 - 0.17)	0.04 (0.02 - 0.05)
2	0.04 (0.03 - 0.05)	**0.01 (0.00 - 0.02)**	0.06 (0.04 - 0.08)	**0.03 (0.01 - 0.04)**	0.05 (0.03 - 0.06)	0.05 (0.03 - 0.07)
3	0.02 (0.01 - 0.04)	**0.00 (0.00 - 0.01)**	0.04 (0.03 - 0.05)	0.02 (0.01 - 0.03)	0.03 (0.02 - 0.04)	0.03 (0.02 - 0.04)
4	**0.02 (0.01 - 0.03)**	**0.02 (0.01 - 0.03)**	**0.01 (0.00 - 0.02)**	0.04 (0.03 - 0.06)	0.10 (0.08 - 0.13)	0.03 (0.02 - 0.05)
5	**0.02 (0.01 - 0.03)**	0.05 (0.03 - 0.06)	**0.03 (0.02 - 0.04)**	0.04 (0.03 - 0.05)	0.05 (0.03 - 0.06)	0.05 (0.03 - 0.06)
6	**0.03 (0.01 - 0.04)**	**0.03 (0.02 - 0.04)**	**0.02 (0.01 - 0.03)**	**0.02 (0.01 - 0.03)**	0.03 (0.02 - 0.05)	0.03 (0.02 - 0.04)
7	0.10 (0.08 - 0.12)	**0.01 (0.00 - 0.02)**	0.09 (0.07 - 0.11)	0.11 (0.08 - 0.13)	0.15 (0.12 - 0.18)	0.08 (0.06 - 0.10)
8	**0.01 (0.00 - 0.01)**	**0.01 (0.00 - 0.02)**	0.04 (0.02 - 0.05)	0.04 (0.02 - 0.05)	0.04 (0.03 - 0.06)	0.03 (0.02 - 0.04)
9	0.24 (0.20 - 0.27)	**0.04 (0.03 - 0.05)**	0.07 (0.05 - 0.08)	0.42 (0.38 - 0.45)	0.48 (0.44 - 0.51)	0.42 (0.38 - 0.45)
Average	0.06 +/- 0.06	0.02 +/- 0.01	0.05 +/- 0.03	0.10 +/- 0.11	0.13 +/- 0.13	0.10 +/- 0.12

Table 3.12: **FNavg MNIST classification.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no residual	Gated Attention	Grad-CAM	Grad	Guided-backprop
	(this paper)	Dubost et al. (2017)	Schlemper et al. (2018)	Selvaraju et al. (2017)	Simonyan et al. (2014)	Springenberg et al. (2015)
0	1.13 (1.01 - 1.25)	**1.07 (0.95 - 1.19)**	1.11 (0.99 - 1.23)	1.22 (1.09 - 1.35)	1.31 (1.17 - 1.44)	1.28 (1.15 - 1.41)
1	1.34 (1.20 - 1.48)	**1.24 (1.11 - 1.38)**	1.32 (1.18 - 1.46)	1.36 (1.22 - 1.50)	1.52 (1.36 - 1.68)	1.29 (1.15 - 1.43)
2	**1.15 (1.02 - 1.28)**	**1.14 (1.02 - 1.27)**	1.20 (1.07 - 1.32)	1.17 (1.04 - 1.30)	1.27 (1.13 - 1.42)	1.19 (1.06 - 1.32)
3	1.05 (0.93 - 1.16)	**1.02 (0.90 - 1.14)**	1.05 (0.93 - 1.16)	1.06 (0.95 - 1.18)	1.14 (1.02 - 1.27)	1.08 (0.96 - 1.19)
4	1.09 (0.97 - 1.21)	**1.05 (0.93 - 1.17)**	1.08 (0.96 - 1.21)	1.09 (0.97 - 1.22)	1.27 (1.13 - 1.42)	1.09 (0.97 - 1.21)
5	0.99 (0.88 - 1.10)	**0.97 (0.86 - 1.08)**	0.98 (0.87 - 1.10)	1.00 (0.89 - 1.12)	1.09 (0.96 - 1.21)	1.01 (0.90 - 1.12)
6	**1.11 (0.99 - 1.23)**	1.16 (1.04 - 1.28)	**1.11 (0.99 - 1.23)**	**1.11 (0.99 - 1.23)**	1.26 (1.12 - 1.40)	**1.12 (1.00 - 1.24)**
7	1.26 (1.13 - 1.40)	**1.17 (1.04 - 1.31)**	1.24 (1.11 - 1.38)	1.32 (1.18 - 1.45)	1.37 (1.23 - 1.51)	1.29 (1.16 - 1.43)
8	**1.06 (0.94 - 1.18)**	**1.06 (0.94 - 1.19)**	1.09 (0.96 - 1.21)	**1.07 (0.94 - 1.19)**	1.34 (1.18 - 1.49)	**1.06 (0.94 - 1.19)**
9	1.32 (1.18 - 1.46)	**1.15 (1.02 - 1.29)**	**1.15 (1.01 - 1.28)**	1.46 (1.31 - 1.61)	1.52 (1.36 - 1.68)	1.46 (1.31 - 1.61)
Average	1.15 +/- 0.11	1.10 +/- 0.08	1.13 +/- 0.09	1.19 +/- 0.14	1.31 +/- 0.13	1.19 +/- 0.13

Appendix C. Results MNIST – GP-Unet architectures.

Table 3.13: **FAUCs MNIST architecture.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no skip	GP-Unet pax pooling
0	**97.4 (96.3 - 98.3)**	**98.0 (97.5 - 98.5)**	89.2 (87.2 - 91.3)
1	**94.4 (93.6 - 95.2)**	86.9 (85.6 - 88.2)	80.0 (77.2 - 82.6)
2	**91.7 (90.4 - 92.9)**	**92.9 (91.4 - 94.3)**	80.3 (77.5 - 83.0)
3	**97.3 (96.6 - 97.9)**	**97.9 (97.4 - 98.5)**	88.6 (86.4 - 90.8)
4	**97.8 (97.2 - 98.3)**	**96.2 (95.2 - 97.0)**	82.9 (80.2 - 85.5)
5	**97.1 (96.4 - 97.8)**	**97.6 (97.1 - 98.1)**	86.1 (83.6 - 88.6)
6	**98.6 (98.1 - 99.1)**	**98.1 (97.5 - 98.7)**	87.6 (85.2 - 89.9)
7	**89.3 (87.6 - 90.9)**	0.8 (0.5 - 1.2)	56.3 (52.8 - 59.8)
8	**98.8 (98.2 - 99.2)**	**98.8 (98.4 - 99.1)**	87.7 (85.3 - 89.9)
9	**97.6 (96.8 - 98.2)**	**96.9 (96.1 - 97.6)**	78.4 (75.5 - 81.1)
Average	96.0 +/- 3.0	86.4 +/- 28.7	81.7 +/- 9.3

Table 3.14: **Sensitivity MNIST architecture.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no skip	GP-Unet pax pooling
0	**92.7 (91.4 - 93.9)**	**90.9 (89.6 - 92.3)**	81.8 (80.1 - 83.5)
1	**78.9 (77.3 - 80.4)**	63.3 (61.5 - 65.0)	71.9 (70.1 - 73.8)
2	**80.0 (78.2 - 81.8)**	**83.4 (81.7 - 85.1)**	75.8 (74.0 - 77.7)
3	90.1 (88.8 - 91.5)	**91.8 (90.7 - 93.0)**	82.8 (81.1 - 84.5)
4	**90.7 (89.3 - 92.1)**	87.8 (86.3 - 89.2)	76.1 (74.1 - 78.1)
5	**88.7 (87.2 - 90.2)**	**89.4 (87.9 - 90.8)**	74.7 (72.7 - 76.8)
6	**92.2 (91.0 - 93.5)**	**91.9 (90.5 - 93.3)**	81.4 (79.5 - 83.4)
7	**76.3 (74.6 - 78.1)**	0.6 (0.3 - 0.9)	55.2 (52.9 - 57.5)
8	**95.8 (95.0 - 96.5)**	**94.5 (93.7 - 95.4)**	83.7 (82.1 - 85.3)
9	**92.3 (91.1 - 93.5)**	90.1 (88.8 - 91.5)	70.8 (68.8 - 72.9)
Average	87.8 +/- 6.4	78.4 +/- 27.3	75.4 +/- 8.0

Table 3.15: **FPavg MNIST architecture.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no skip	GP-Unet pax pooling
0	**0.07 (0.05 - 0.08)**	**0.07 (0.05 - 0.09)**	**0.08 (0.06 - 0.10)**
1	**0.14 (0.11 - 0.16)**	0.30 (0.26 - 0.35)	**0.16 (0.13 - 0.19)**
2	0.29 (0.25 - 0.32)	0.22 (0.18 - 0.25)	**0.15 (0.12 - 0.17)**
3	0.11 (0.09 - 0.14)	**0.06 (0.05 - 0.08)**	**0.09 (0.07 - 0.12)**
4	0.13 (0.10 - 0.15)	0.10 (0.08 - 0.13)	**0.06 (0.04 - 0.07)**
5	0.13 (0.11 - 0.16)	**0.10 (0.08 - 0.12)**	**0.09 (0.06 - 0.11)**
6	**0.04 (0.03 - 0.06)**	0.08 (0.06 - 0.10)	0.11 (0.09 - 0.14)
7	**0.26 (0.23 - 0.29)**	1.99 (1.98 - 2.00)	**0.21 (0.17 - 0.24)**
8	0.08 (0.06 - 0.10)	0.06 (0.05 - 0.08)	**0.03 (0.02 - 0.04)**
9	0.16 (0.13 - 0.18)	0.19 (0.16 - 0.22)	**0.10 (0.08 - 0.13)**
Average	0.14 +/- 0.08	0.32 +/- 0.56	0.11 +/- 0.05

Table 3.16: **FNavg MNIST architecture.** Each row corresponds to the detection of a different digit. 95 percent confidence interval is indicated in brackets. The average and standard deviation of the performance of each method across all digits is given in the last row. Best performance are indicated in bold.

	GP-Unet	GP-Unet no skip	GP-Unet pax pooling
0	**0.11 (0.08 - 0.14)**	**0.13 (0.10 - 0.16)**	0.30 (0.25 - 0.35)
1	**0.35 (0.30 - 0.40)**	0.60 (0.53 - 0.67)	0.49 (0.42 - 0.56)
2	0.30 (0.25 - 0.34)	**0.25 (0.21 - 0.29)**	0.39 (0.34 - 0.45)
3	**0.15 (0.12 - 0.18)**	**0.13 (0.10 - 0.16)**	0.26 (0.22 - 0.31)
4	**0.12 (0.10 - 0.15)**	0.18 (0.14 - 0.21)	0.37 (0.31 - 0.42)
5	**0.14 (0.11 - 0.16)**	**0.13 (0.11 - 0.16)**	0.36 (0.30 - 0.41)
6	**0.11 (0.08 - 0.13)**	**0.10 (0.08 - 0.12)**	0.26 (0.22 - 0.30)
7	**0.36 (0.31 - 0.41)**	1.58 (1.42 - 1.74)	0.71 (0.62 - 0.80)
8	**0.07 (0.05 - 0.09)**	**0.09 (0.07 - 0.11)**	0.26 (0.21 - 0.31)
9	**0.11 (0.09 - 0.14)**	0.15 (0.12 - 0.18)	0.47 (0.41 - 0.54)
Average	0.18 +/- 0.10	0.33 +/- 0.44	0.39 +/- 0.13

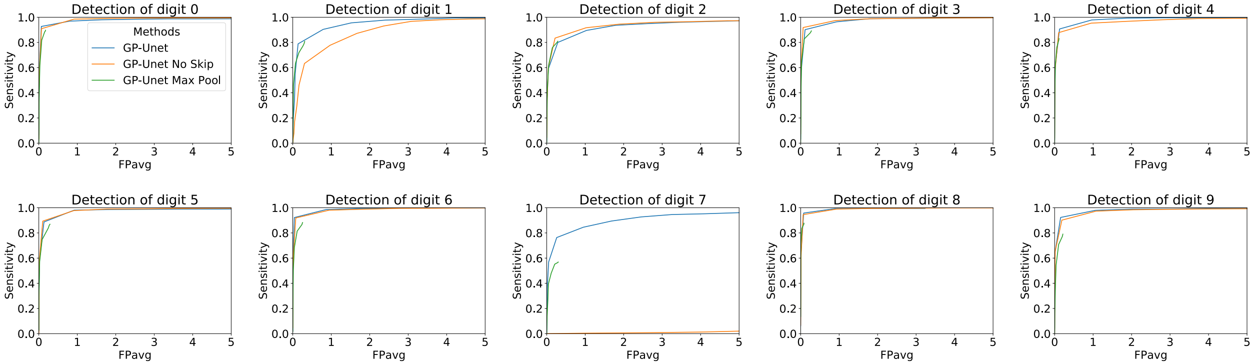


Figure 3.14: **FROC MNIST architecture.** Each subplot corresponds to the detection of a different digit. GP-Unet is the standard GP-Unet architecture. In GP-Unet No Skip, blockwise skip connections are removed. In GP-Unet Max Pool, the global average pooling is replaced by global max pooling.

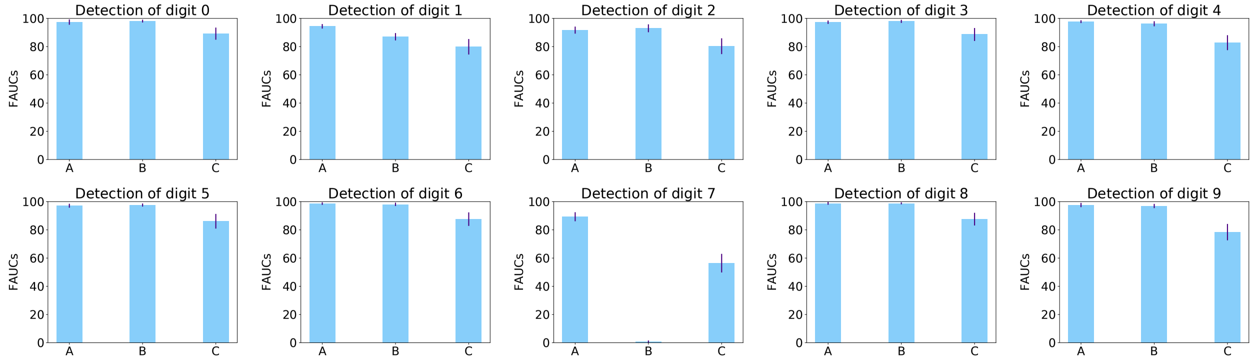


Figure 3.15: **FAUCs on the MNIST dataset for different architectures of the proposed method: GP-Unet.** Each subplot corresponds to the detection of a different digit. FAUCs are displayed with confidence intervals computed by bootstrapping the test set. A: GP-Unet; B: GP-Unet without blockwise skip connections; C: GP-Unet with Global Max Pooling. For digit 7, GP-Unet without blockwise skip connections convergence to a very high value. We tried repeating the experiments with different random initializations of the weights and let the optimization run longer, but we achieved the same results. This supports the argument that adding blockwise skip connections eases the optimization.

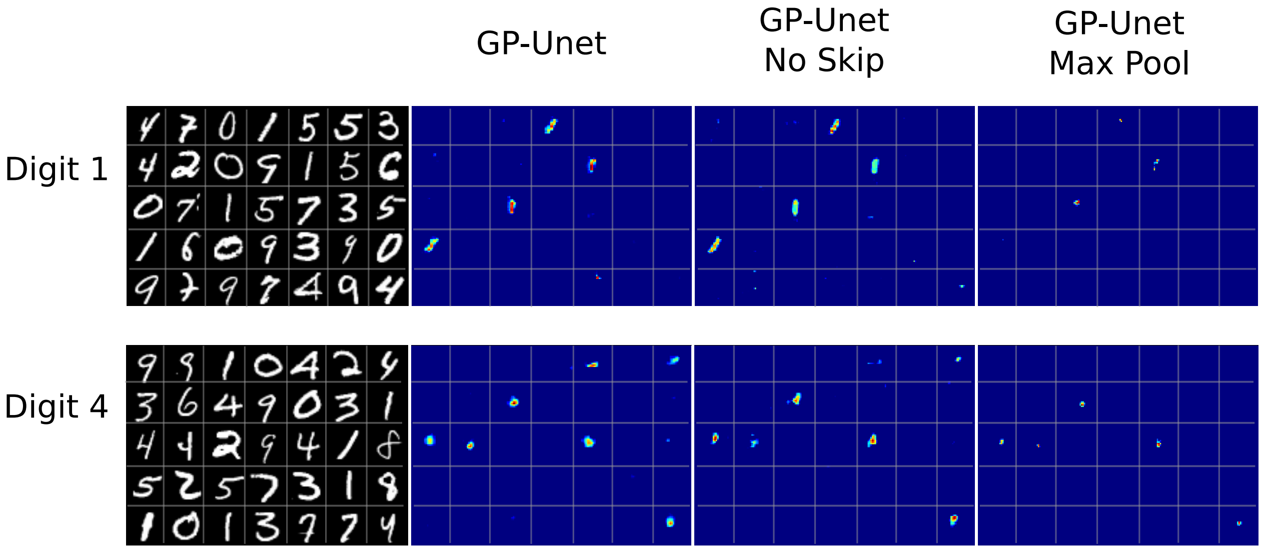


Figure 3.16: **Attention Maps for the different architectures of GP-Unet.** In the first row, the digits 1 had to be detected. In the second row, digits 4 had to be detected. GP-Unet is the standard GP-Unet architecture. In GP-Unet No Skip, blockwise skip connections are removed. In GP-Unet Max Pool, the global average pooling is replaced by global max pooling.

Acknowledgements

This research was funded by The Netherlands Organisation for Health Research and Development (ZonMw) Project 104003005, with additional support of Netherlands Organisation for Scientific Research (NWO), project NWO-EW VIDI 639.022.010 and project NWO-TTW Perspectief Programme P15-26. This work was partly carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Chapter 4

Automated Lesion Detection by Regressing Intensity-Based Distance with a Neural Network

Abstract

Localization of focal vascular lesions on brain MRI is an important component of research on the etiology of neurological disorders. However, manual annotation of lesions can be challenging, time-consuming and subject to observer bias. Automated detection methods often need voxel-wise annotations for training. We propose a novel approach for automated lesion detection that can be trained on scans only annotated with a dot per lesion instead of a full segmentation. From the dot annotations and their corresponding intensity images we compute various distance maps (DMs), indicating the distance to a lesion based on spatial distance, intensity distance, or both. We train a fully convolutional neural network (FCN) to predict these DMs for unseen intensity images. The local

optima in the predicted DMs are expected to correspond to lesion locations. We show the potential of this approach to detect enlarged perivascular spaces in white matter on a large brain MRI dataset with an independent test set of 1000 scans. Our method matches the intra-rater performance of the expert rater that was computed on an independent set. We compare the different types of distance maps, showing that incorporating intensity information in the distance maps used to train an FCN greatly improves performance.

1 Introduction

Obtaining the location of focal vascular lesions on brain scans, such as white matter hyperintensities, lacunes, enlarged perivascular spaces or microbleeds is extremely useful for studying the association of these lesions with neurological disorders. However the manual annotation of these lesions can be challenging, time-consuming and subject to observer bias due to the difficulty of distinguishing a specific type of lesion from other similarly appearing structures. An automated method for detecting lesions could improve reliability, generalization and speed of lesion detection, which could greatly advance neuropathology research.

Various promising automated methods have been proposed to detect lesions. Deep learning methods often provide the best accuracy, but depend on expensive manual annotations for training like voxel-wise segmentations (Brosch et al., 2016; Ghafoorian et al., 2017) or bounding boxes (Dou et al., 2016) marking the lesions. This hinders applicability of these techniques in practice.

Annotating by placing a single dot per lesion instead is considerably more time-efficient, allowing to collect larger annotated datasets for training and evaluation. In this chapter we therefore propose a novel method for lesion detection that requires only dot annotations. Dot annotations have been effectively used to train convolutional neural networks (CNNs) for other applications, such as cell detection in histology images (Xie et al., 2018c), lacune detection in placental ultrasound (Qi et al., 2018) and landmark detection in retinal images (Meyer et al., 2018a). An approach that has shown great promise is regression of a distance map (DM) that is computed from these dot annotations (Meyer et al., 2018a; Qi et al., 2018; Xie et al., 2018c). Contrary to many other deep learning detection methods that use a two-stage approach (Dou et al., 2016), this approach directly outputs predicted detections and is optimized in an end-to-end fashion.

We use a similar approach for detecting lesions based on dot annotations. Previous distance regression approaches for detection (Meyer et al., 2018a; Xie et al., 2018c) have used Euclidean distance. This is especially suited for the detection of circular objects such as cells. Brain lesions on the other hand often have a morphology that is complex and discriminative (Boespflug et al., 2017).

In this chapter we investigate the effect of including intensity information in DMs for lesion detection. Intensity distance incorporates local image context enabling the DM to capture complicated morphologies. Voxels surrounding dot annotations which have similar intensity values (inside the lesions) will have a lower value in the DM than dissimilar voxels (outside the lesions). This could encourage the CNN to learn the characteristic morphology of the lesions and propose more accurate detections than when trained on a Euclidean distance map (EDM) that does not make this distinction. We compare Euclidean distance, intensity distance, and geodesic distance that combines both Euclidean and intensity distances. For geodesic distance the image is seen as a curved surface defined by the spatial coordinates and one intensity coordinate, where the shortest path on the surface is the geodesic distance (Toivanen, 1996).

In this chapter we show that including image intensity information in the DM improves optimization of a CNN for detecting lesions in brain MRI. We compute DMs from the dot annotations and their corresponding intensity images. Subsequently we train a fully convolutional neural network (FCN) to predict these DMs for unseen intensity images. The local minimal distances in the predicted DMs correspond to the proposed detection candidates.

We show the potential of regressing intensity-based DMs for the detection of enlarged perivascular spaces (PVS). PVS burden has been associated with cerebral small vessel disease (Charidimou et al., 2013). As PVS follow the course of the vessel they surround, they appear as elongated structures on 3D brain MRI scans. Several methods have been proposed to detect PVS. The majority of the proposed algorithms is however evaluated on a relatively small sets (less than 30 images) due to the need for voxel-wise annotations for testing (and training) (Boespflug et al., 2017; Lian et al., 2018b). We train and validate on a set of 1202 MRI scans and test on a separate set of 1000 images. As the centrum semiovale (CSO) is seen as the most difficult brain region for PVS detection and most clinically relevant, we focused on this brain region (Ballerini et al., 2018).

2 Method

We train an FCN to regress a DM for a given intensity image. Our approach requires MRI scans with dot annotations for training. The local optima in the predicted DMs are expected to correspond to lesion locations. We compare geodesic distance maps (GDMs), EDMs and intensity distance maps (IDMs).

2.1 Distance Transform

To compute DMs we use a distance transform, that requires a definition of the foreground – in our case the set of dot annotations Φ – and a gray-scale image $G(x)$ in the case of intensity and geodesic distances, with x the position in the image. The distance map $DM(x)$ is defined by

$$DM(x) = min(\Lambda(\gamma), \gamma \in \Psi(x, \Phi)) \tag{4.1}$$

with $\Psi(x, \Phi)$ the set of possible paths γ between a position x in the image and the set of dot annotations Φ. The length $\Lambda(\gamma)$ of the path γ is

$$\Lambda(\gamma) = \sum_{i=1}^{n-1} d(x_i, x_{i+1}) \tag{4.2}$$

with n the number of voxels in the path γ between a position x and a dot annotation $x_{dot} \in \Phi$ and d the distance measure. The geodesic distance d_G in a 2D gray-scale image between voxel x_i and the next voxel in the path x_{i+1}, with intensities $G(x_i)$ and $G(x_{i+1})$ respectively, is defined by Toivanen (1996) as

$$d_G(x_i, x_{i+1}) = \sqrt{d_I\left(x_i, x_{i+1}\right)^2 + d_E\left(x_i, x_{i+1}\right)^2} \tag{4.3}$$

with the intensity distance $d_I(x_i, x_{i+1}) = G(x_i) - G(x_{i+1})$ and the Euclidean distance $d_E(x_i, x_{i+1})$ which is 1 for $x_{i+1} \in N_4(x_i)$ (voxels connected horizontally and vertically) and $\sqrt{2}$ for $x_{i+1} \in N_8(x_i) \setminus N_4(x_i)$ (voxels connected diagonally). EDMs are consequently computed by setting $d_I = 0$ in equation 4.3, while IDMs are computed by setting $d_E = 0$. We approximate these DMs using the optimization algorithm *iterative raster scan* described by Toivanen (1996). This approach is for computing DMs in 2D, though it can easily be extended to 3D.[1]

The resulting $DM(x)$ is normalized by dividing by the maximum distance in the $DM(x)$ and inverted as this is convenient for implementation. Furthermore, we add a parameter p to influence how steeply the distance decays. The final map $M_p(x)$ is calculated using

$$M_p(x) = \Big(1 - \frac{DM(x)}{max\big(DM(x)\big)}\Big)^p \tag{4.4}$$

2.2 Fully Convolutional Neural Network

We use an architecture similar to a shallow U-Net for our FCN shown in Figure 4.1, which was shown to work well for regressing the number of perivascular spaces in the basal ganglia (Dubost et al., 2017; Ronneberger et al., 2015). For optimization we use mean square error loss $MSE = \frac{1}{N}\sum\limits_{x}\Big(\widehat{M}_p(x) - M_p(x)\Big)^2$, with $\widehat{M}(x)$ the predicted map and N the number of voxels in $M_p(x)$.

Non-maximum suppression is applied to the predicted distance map to detect local optima. We use a 5×5 maximum filter with a connectivity of 8. By thresholding the local optima the proposed detections are acquired.

[1] Our code for computing 2D as well as 3D distance maps is available at https://github.com/kimvwijnen/geodesic_distance_transform

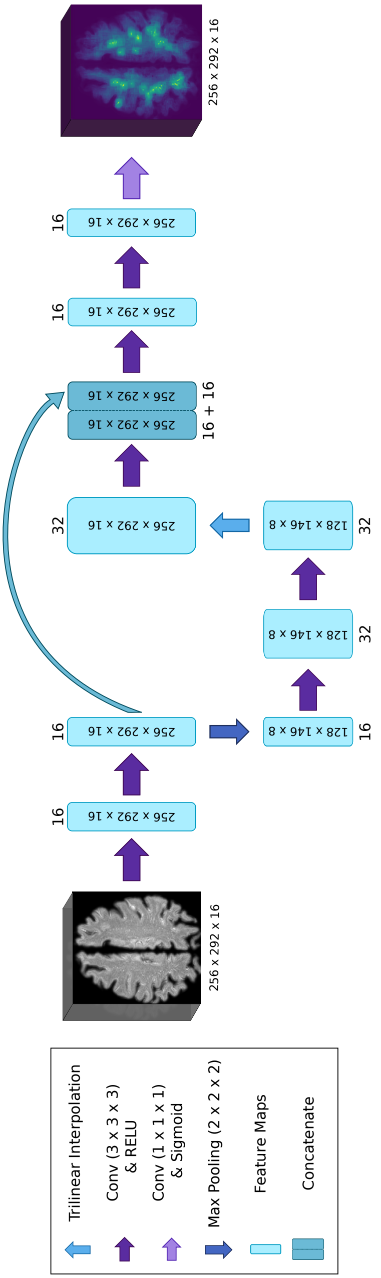


Figure 4.1: Network architecture, on the left the input preprocessed brain scan is shown and the output predicted distance map is shown on the right

3 Experiments

3.1 Data

Our data set consists of 2202 T2-weighted MRI scans from the Rotterdam Scan Study. All scans were from different individuals and were acquired on a 1.5 T MRI scanner. The images have a size of $512 \times 512 \times 192$ with a voxel resolution of $0.49 \times 0.49 \times 0.8\text{mm}^3$. Further details on the image acquisition of this data are discussed by Ikram et al. (2015).

The number of PVS in the axial slice 1 cm above the lateral ventricles is highly correlated with the total number of PVS in the CSO (Adams et al., 2015). The rater selected this specific slice and annotated it with dots indicating PVS between 1 - 3 mm in diameter in line with the guidelines described by Adams et al. (2013). The intra-rater performance was evaluated on a separate set of 40 MRI scans (see Table 4.1 and Figure 4.3).

3.2 Preprocessing

Images are preprocessed as proposed in chapter 1. We segment the CSO with the FreeSurfer multi-atlas segmentation algorithm (Desikan et al., 2006a) producing a binary mask that we smooth with a Gaussian kernel. The image are multiplied with the smoothed mask and cropped to a fixed size containing only the slices close to the annotated slice. The resulting images are normalized to the range [0,1] by dividing by the maximum intensity in the image.

Annotated dots were not always inside the PVS. To solve this problem, we shift the dots to the highest intensity value within the same connected component and within 3 voxels distance. The shifted dots were only used to compute the distance maps for the training and validation set. For evaluation of the detection performance, the original annotated dots were used.

3.3 Experimental Setup

Random sampling was used to split the 2202 scans into a set of 1202 for development of the method (1000 for training and 202 for validation) and a separate set of 1000 for testing. As only one slice per scan was annotated, DMs were computed in 2D and the loss was only evaluated for this slice. Non-maximum suppression and evaluation of detection performance was also only done on the slice that was annotated.

Weights for the convolutional layers were initialized by random sampling from a truncated normal distribution with zero mean and unit variance. For optimization we use Adadelta and a batch of one due to memory limitations. We use on-the-fly augmentation for the training set. For every image a random rotation around the depth direction with a maximum of 20° in both directions is applied combined with random flipping in horizontal and in vertical direction. Methods were implemented in Python and Keras with Tensorflow as backend.

3.4 Detection Performance

The candidate detections of each method are compared to the expert annotations using the hungarian algorithm to find a one-to-one mapping between these sets. Only detections within a 6 voxel radius of the annotations were counted as true positive. We use 6 voxels as this is the maximum PVS diameter (corresponds to 3 mm (Adams et al., 2013)).

The detection performance is mainly evaluated with the Free-Response Operating Characteristic (FROC) curve and its area under the curve (FAUC) until

10 FP_{avg}, which is approximately twice the FP_{avg} of the rater. The FAUC is calculated as the percentage of the highest possible area. We used bootstrapping to quantify the uncertainty, resulting in a mean FAUC and confidence interval based on 1000 sampled sets. Bootstrapping was performed by random sampling with replacement from the test set.

3.5 Evaluation Approach

We ran experiments varying the decay parameter p (see Figure 4.2). For higher values of p the FCN did not train, we expect because of label imbalance. Based on the FAUC on the validation set we set p to 5 for geodesic distance, to 6 for intensity distance and 9 for Euclidean distance. During training, the model parameters were chosen as the ones minimizing the FAUC computed on the validation set. Only the best model per distance type ($GDM_5(x)$, $EDM_9(x)$, $IDM_6(x)$) was tested on the test set of 1000 scans.

The operating point on the FROC was chosen per model as the threshold with a sensitivity on the validation set closest to the average intra-rater sensitivity. For $GDM_5(x)$ the threshold was chosen at 0.525, for $EDM_9(x)$ at 0.500 and for $IDM_6(x)$) at 0.495. This threshold was used as the detection threshold during evaluation on the test set.

3.6 Results

Figure 4.3 shows the FROC curves computed on the test set and examples of the output of the FCNs. Table 4.1 shows the corresponding FAUCs, the sensitivity and FP_{avg} of the methods on the test set at the chosen thresholds (based on the validation set) and the average intra-rater performance.

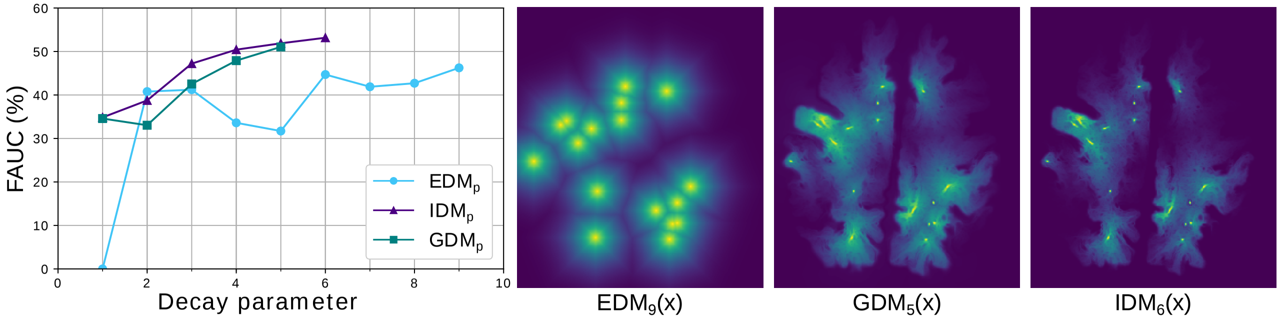


Figure 4.2: Influence of decay parameter p on detection performance on the validation set and the chosen distance maps

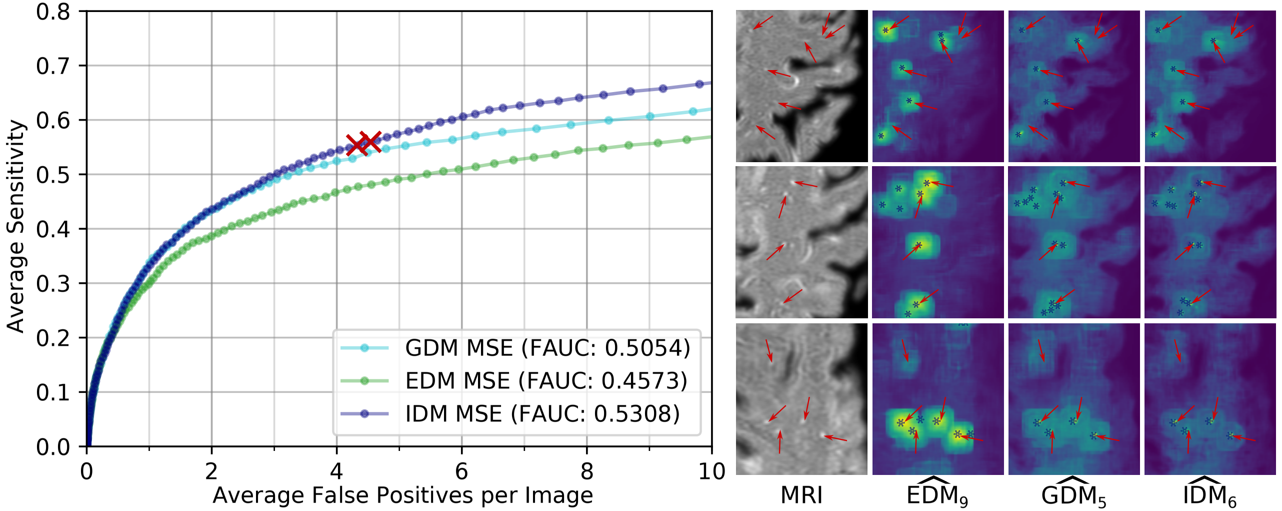


Figure 4.3: FROC curves and crops of the output of the FCNs and their proposed detections (intra-rater performance indicated with red crosses, annotations with red arrows and predictions with blue stars)

4 Discussion and Conclusion

Our experiments indicate that incorporating image intensity information in a distance map used to train an FCN substantially improves performance of PVS detection. Results show that using GDMs and IDMs both result in a similar detection performance, with IDMs sometimes reaching higher performance than GDMs. This indicates that intensity difference is the most discriminative information, and that Euclidean distance could even be ignored. Using higher values of the decay parameter also increases the PVS detection performance, and stabilizes the optimization.

The FCN trained using IDMs reaches a sensitivity and FP_{avg} similar to the intra-rater performance computed on a smaller independent set (Figure 4.3).

We expect our method could perform well for detecting other types of focal vascular lesions in the brain. Using intensity information in the computation of DMs could help the detection lesions that either have a complex morphology, or can have substantial variation in their size, such as microbleeds, white matter hyperintensities or lacunes. Additionally, in this work we evaluate the intensity-based distance maps only for their performance in detecting PVS. However, we observe that the PVS detections in the output maps of the FCNs trained on intensity-based distance maps (Figure 4.3) seem to approximate the PVS shape quite well. We therefore expect our approach might also work well for segmentation.

Table 4.1: PVS detection performance on the test set for the detection methods and the average intra-rater performance on a smaller independent set

	FAUC	FP_{avg}	Sensitivity
$EDM_9(x)$	45.761 (± 0.052)	7.49	53.63
$GDM_5(x)$	50.575 (± 0.050)	5.10	55.26
$IDM_6(x)$	53.078 (± 0.051)	4.35	55.35
Avg intra-rater	-	4.43	55.66

Acknowledgments

This research is part of the research project Deep Learning for Medical Image Analysis (DLMedIA) with project number P15-26, funded by the Dutch Technology Foundation STW (part of the Netherlands Organisation for Scientific Research (NWO), which is partly funded by the Ministry of Economic Affairs), and with co-financing by Quantib. This research was also funded by the Netherlands Organisation for Health Research and Development (ZonMw), Project 104003005. Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative and on a Quadro P6000 donated by the NVIDIA Corporation.

Part D

Automated Quantification of Enlarged Perivascular Spaces

Chapter 5

Enlarged Perivascular Spaces in Brain MRI: Automated Quantification in four Regions

Abstract

Enlarged perivascular spaces (PVS) are structural brain changes visible in MRI, are common in aging, and are considered a reflection of cerebral small vessel disease. As such, assessing the burden of PVS has promise as a brain imaging marker. Visual and manual scoring of PVS is a tedious and observer-dependent task. Automated methods would advance research into the etiology of PVS, could aid to assess what a ”normal” burden is in aging, and could evaluate the potential of PVS as a biomarker of cerebral small vessel disease. In this work, we propose and evaluate an automated method to quantify PVS in the midbrain, hippocampi, basal ganglia and centrum semiovale. We also compare associations between (earlier established) determinants of PVS and visual PVS scores versus the automated PVS scores, to verify whether automated PVS scores could replace visual scoring of PVS in epidemiological and clinical studies. Our approach is a deep learning algorithm based on convolutional neural network regression. We trained and validated our method on T2-contrast MR images acquired from 2115 subjects participating in a population-based study. These scans were visually scored by an expert rater, who counted the number of PVS in each brain region. Agreement between visual and automated scores was found to be excellent for all four regions, with intraclass correlation coefficients (ICCs) between 0.75 and 0.88. These values were higher than the inter-observer agreement of visual scoring (ICCs between 0.62 and 0.80). Scan-rescan reproducibility was high (ICC 0.82 – 0.93). The association between 20 determinants of PVS, including aging, and the automated scores were similar to those between the same 20 determinants of PVS and visual scores. We conclude that this method may replace visual scoring and facilitate large epidemiological and clinical studies of PVS.

1 Introduction

This chapter proposes and evaluates an algorithm for the automated quantification of enlarged perivascular spaces (PVS) in four brain regions. Perivascular spaces are fluid-filled areas surrounding cerebral arteries or veins. These spaces tend to enlarge locally in aging subjects (Wardlaw et al., 2013). Enlarged perivascular spaces can be identified as hyperintensities in T2-contrast MRI, as illustrated in Figure 5.1. Though initially considered a strictly normal phenomenon, the presence of PVS is increasingly thought to reflect the presence of cerebral small vessel disease and to function as a potential biomarker for various brain diseases such as dementia (Mills et al., 2007), stroke (Selvarajah et al., 2009), multiple sclerosis (Achiron and Faibel, 2002), and Parkinson (Zijlmans et al., 2004).

The progressive enlargement of PVS, their widespread occurrence in the brain, and presence of mimics with similar appearance on MRI make the manual annotation of individual PVS challenging and time consuming (see Figure 5.2). Instead, current studies largely rely on visual scoring systems. Two types of scoring systems have been proposed in the literature: expert raters either count the number of PVS within a region of interest (Adams et al., 2013, 2015) or categorize the PVS burden using a scale (e.g. Potter scores (Potter et al., 2015a) and Patankar scores (Patankar et al., 2005)). Automated quantification of PVS would be preferred as it is more objective and faster than visual scoring. Furthermore, it would hold great potential to study burden of PVS as a continuous rather than a categorical measure, enabling to better disentangle ”normal” structural brain changes in aging from a pathological load of PVS.

In Chapter 1, we proposed a regression convolutional network to quantify PVS in the basal ganglia. In the present work, we extend this method to other brain regions most clinically relevant for PVS quantification, namely the midbrain, hippocampi and centrum semiovale (Adams et al., 2013, 2015), and we provide a more elaborate evaluation.

Similar to the method described in Chapter 1, the input in our method is a T2-contrast brain scan, and the output is an automated PVS score. The aim of the method is to reproduce the visual scores of an expert rater, considered here as the reference standard. Our method uses a 3D convolutional neural network inspired by ResNet (He et al., 2016) and optimized with a mean squared error (MSE) loss function to minimize the difference between visual scores and predicted scores in a set of training images.

In all four brain regions, we compare the agreement between our automated PVS scores and the visual PVS score of the expert rater, with the level of inter-observer agreement. We assess scan-rescan reproducibility. Finally we check in a subset of 1485 scans whether the associations between determinants of PVS and the automated scores are similar to those between the same determinants and visual scores. The determinants of PVS investigated here include demographics, cardiovascular risk factors, ApoE genotypes, and MRI markers.

1.1 Related Work

Other researchers have published automated PVS quantification methods involving the use of the visual scores as ground truths.

Ballerini et al. (2018) proposed to enhance PVS in the centrum semiovale using multiscale vessel enhancement filtering (Frangi et al., 1998). The parameters of these filters are optimized with ordered logit models, using PVS category scores (Potter et al., 2015a; Patankar et al., 2005) as ground truth. To evaluate their methods, the authors compute correlations between the visual ratings and their segmentation-derived PVS count and PVS volume in two different datasets. This method has only been evaluated in the centrum semiovale. Results were mixed with correlations ranging from 0.47 to 0.74 in different datasets.

Gonzalez-Castro et al. (2017) addressed PVS quantification in the basal ganglia as a binary classification problem, where the objective is to discriminate between scans with few ($\leqslant$10) or many ($>$10) PVS. Their method uses support vector machines and bag-of-words descriptors. The agreement between their classifier and a human observer is similar to the inter-observer agreement. The authors also show associations between determinants of PVS (age, Fazekas scale, and presence of lacunar infarcts) and the binary score of the classifier. Our work extends this by proposing a continuous score indicating the number of PVS instead of a binary score, leading to a finer quantification. We evaluate our method in four brain regions, and investigate associations with a wider range of determinants.

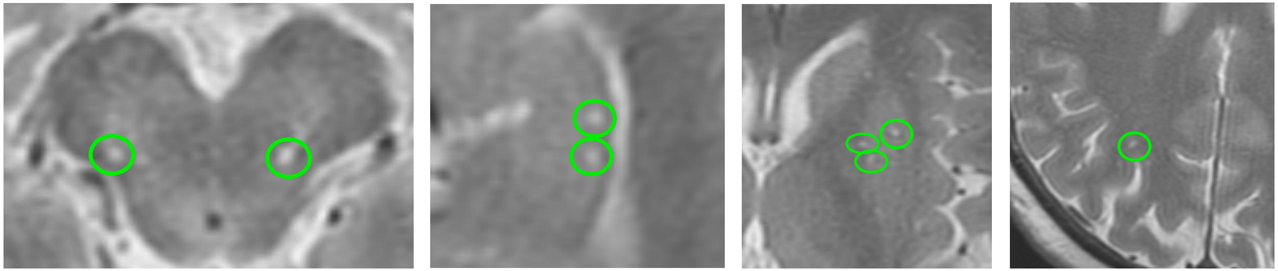


Figure 5.1: **Examples of enlarged perivascular spaces in different brain regions.** T2-contrast MRI images in the axial view. PVS are circled in green. From left to right: midbrain, hippocampus, basal ganglia and centrum semiovale. On these images the PVS are relatively easy to detect for an expert rater, contrary to Figure 5.2.

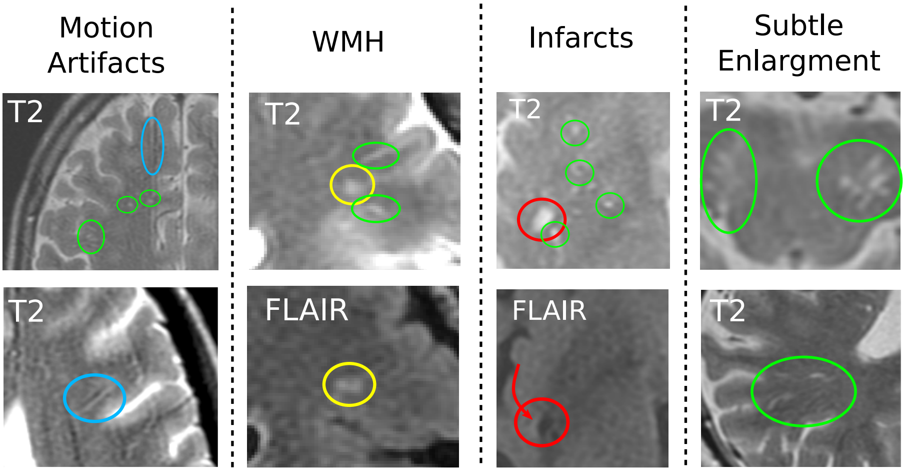


Figure 5.2: **Examples of enlarged perivascular spaces and their mimics in different brain regions.** All images are in the axial view. PVS are circled in green, white matter hyperintensities (WMH) in yellow, lacunar infarcts in red and motion artifacts in blue. In the first column, motion artifact could be mistaken for an elongated PVS in the centrum semiovale. In the second column, the WMH could be mistaken for PVS. However on the FLAIR-weighted scan WMH are hyperintense, while PVS are hypointense and less visible (bottom image). In the third column, the lacunar infarct in the basal ganglia could be mistaken for a group of several PVS, which individual borders could not be seen because of partial volume effect (this lesion would unlikely be mistaken for a single PVS because of its irregular shape). The FLAIR-weighted scan shows a hyperintense rim (red arrow) around the lesion, indicating the presence of a lacunar infarct. In the last column, the scans present several PVS, some of which are at the limit of being considered as enlarged. According to the visual scoring guidelines presented by Adams et al. (2013), to be considering enlarged, perivascular spaces should have a diameter larger than 1mm. For many small perivascular spaces in these images, this is difficult to evaluate.

2 Methods and Materials

The objective of our method is to automatically predict the PVS visual scores. Our framework consists of two steps. We first extract the region of interest (ROI) (Section 2.2) and then apply a regression convolutional neural network (CNN) (Section 2.3) to compute the PVS score. The CNN is trained on an independent set of visually scored scans (N=400 or N=1600).

2.1 Data

In our experiments we used brain MRI scans from the Rotterdam Study. The Rotterdam Study is a prospective population study investigating - among others - neurological diseases in the middle aged and elderly, applying brain MRI in all participants (Ikram et al., 2017). In our experiments, we use 2115 scans of 2115 subjects, acquired between 2005 and 2011.

In addition, we used 60 other scans for which 30 study participants were scanned twice within a short period (19 ± 11 days). The 60 scans of this reproducibility set are not part of the 2115 scans mentioned above and were not visually scored for PVS.

The Medical Ethics Committee of Erasmus MC according to the Population Study Act: Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands has approved the Rotterdam study. All participants provided written informed consent to participate in the study and for information to be obtained from their physicians.

2.1.1 MRI scan protocol

Brain MRI was performed on a 1.5-tesla MRI scanner (GE-Healthcare, Milwaukee, WI, USA) with an eight-channel head coil to obtain: T1-weighted (T1), T2-contrast (T2), fluid-attenuated inversion recovery (FLAIR) and T2*-weighted gradient-recalled-echo sequences.

To compute the automated PVS scores, we only used the T2 scans. These scans were acquired according to the following protocol: 12,300 ms repetition time, 17.3 ms echo time, 16.86 KHz bandwidth, 90-180° flip angle, 1.6 mm slice thickness, 25 cm^2 field of view, 416×256 matrix size. The images are reconstructed to a $512 \times 512 \times 192$ matrix. The voxel resolution is $0.49 \times 0.49 \times 0.8\text{mm}^3$.

More details of the imaging protocol have been described elsewhere (Ikram et al., 2015).

2.1.2 Visual PVS scores

Visual PVS scores have been created, for each region, according to a standard procedure proposed in the international consortium UNIVRSE (Adams et al., 2015). PVS ratings are defined as linear, ovoid or round shaped hyperintensities on T2 scans and considered to be enlarged when $\geq$ 1mm and $<$ 3mm. For this visual scoring, a trained observer counts the number of PVS in the midbrain, hippocampi, basal ganglia and centrum semiovale. For the midbrain and hippocampi, the PVS are counted in the whole volume. In the basal ganglia and centrum semiovale, PVS are counted in a single anatomically defined slice. For the basal ganglia, this is the slice showing the anterior commissure. For the centrum semiovale it is the slice 1cm above the upmost part of the lateral ventricles. The number of PVS in these slices correlates well with the number of

PVS in the whole volume of the regions (Adams et al., 2013).

The inter-observer and intra-observer agreements of this scoring have previously been computed in the Rotterdam Study in every region (Adams et al., 2013). Inter-observer intraclass correlation coefficients (ICCs) have been computed with 105 MRI scans, and intra-observer ICCs with 85 scans (Table 5.2). The images in our dataset (2115 scans) were visual scored by a single expert rater (Dr. H. Adams).

2.1.3 Potential determinants of PVS

From the 2115 participants, we randomly selected 400 participants to optimize the parameters of our algorithm, and used the remaining 1715 participants to investigate associations between 20 determinants of PVS and automated and visual PVS scores. From these 1715 participants, we excluded participants without informed consent to access medical records and hospital discharge letters (n=8), participants who already suffered stroke (n=98) or were diagnosed with dementia (n=32) or had incomplete information for stroke or dementia (n=1) at time of MRI scan (de Bruijn et al., 2015; Wieberdink et al., 2012). We also excluded scans for which the brain region segmentation algorithm (FreeSurfer, Desikan et al. (2006a)) failed for one or more regions (n=91). Excluding these resulted in a set of 1485 participants, from which the highest number of missing values was 25 for cholesterol, HDL cholesterol and glucose. Table 5.1 lists the characteristics of the study population.

2.1.3.1 Assessments of determinants Education was obtained from self-reported history and scaled in number of years according to the UNESCO classification [1]. Smoking behaviour was assessed during home interviews and categorized as ever- and non-smokers. Blood pressure measurements were averaged over two readings with a random-zero sphygmomanometer at the right upper arm, in sitting position and a resting period of 5 minutes. Data on serum glucose, total serum cholesterol, serum high-density lipoprotein (HDL) cholesterol were obtained using an automated enzymatic procedure (Boehringer Mannheim System). Diabetes mellitus was defined as a fasting glucose level of $\geq$ 7.0 mmol/L, or the use of antidiabetic medication. Body mass index was calculated by dividing weight (in kilograms) by the height squared (in meters). ApoE genotyping on coded genomic DNA samples was performed for the $\epsilon2$ and $\epsilon4$ alleles of Apolipoprotein E (ApoE-$\epsilon2$ and ApoE-$\epsilon4$) carrier status, with a one-stage polymerase chain reaction and TaqMan assay (Wenham et al., 1991). Participants who were classified ApoE - $\epsilon2\epsilon4$ counted both as $\epsilon2$ and $\epsilon4$ carriers. The majority of samples (81.1%) were genotyped with the Illumina 610K and 660K chips, the remaining (18.9%) were imputed to the Haplotype Reference Consortium reference panel (version 1.0) with Minimac 3.

2.1.3.2 Assessment of MRI markers Several focal and volumetric measures of subclinical brain damage were assessed. Cortical infarcts were defined as lesions involving cortical gray matter with tissue loss and lacunar infarcts as subcortical lesions $\geq$ 3mm and $<$ 15mm on FLAIR, T1, and T2 sequences. The presence of cortical and lacunar infarcts was visually rated by trained research physicians (Ikram et al., 2017). White matter hyperintensities (WMH) were measured quantitatively using a validated automated segmentation method (de Boer et al., 2010). This method was also used to segment the brain into gray

[1] United Nations Educational SaCOU. International Standard Classification of Education (ISCED) 1976. Available from: http://unesdoc.unesco.org/images/0002/000209/020992eb.pdf.

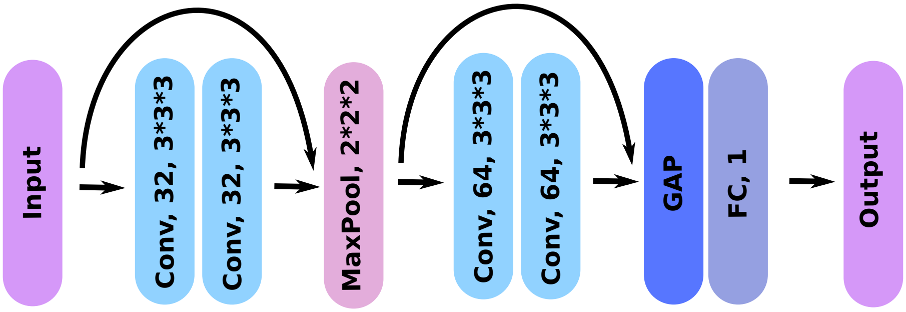


Figure 5.3: **Architecture of the neural network.** The input is a 3D scan cropped around the region of interest, and the output is the automated PVS score. 'Conv' stands for convolutional layer, and is followed by the number of filters and the filter size; 'MaxPool' stands for max pooling layer; 'GAP' for global average pooling, 'FC' for fully connected layer; and the curved arrows represent skip connections with concatenation of feature maps.

matter, WMH and cerebrospinal fluid. Total brain volume was defined as the sum of gray and white matter. And intracranial volume was defined as sum of gray and white matter, and cerebrospinal fluid.

2.2 Preprocessing

The first step of our method is to extract the target brain region from the scan and mask the surrounding structures. This preprocessing step is almost identical for all four regions.

We first apply the FreeSurfer multi-atlas segmentation algorithm (Desikan et al., 2006a) to obtain a binary mask for each region: midbrain, hippocampi, basal ganglia and centrum semiovale. All parameters are left as default, except for the skull stripping preflooding height threshold which is set to 10. These masks are then dilated (4 consecutive morphological binary dilations with a cube connectivity equal to one, i.e., 6-connected in 3D), with the exception of the mask of the midbrain, which is eroded (two consecutive morphological binary erosions with a square connectivity equal to one). These morphological operations can correct segmentation errors and are especially important for the basal ganglia and hippocampi, as PVS can often be located on the border of these regions. On the contrary, for the midbrain, PVS are almost always located in the center and dilating the mask can make the optimization of the model more difficult.

For each region, the borders of the masks are smoothed with a Gaussian kernel of standard deviation $\sigma = 2$ voxel units, and multiplied pixel-wise with the image intensities. These masked images are then cropped around the center of mass of the mask to reduce the image size and memory requirements. The size in voxels of these cropped images for midbrain, hippocampi, basal ganglia and centrum semiovale are 88x88x11, 168x128x84, 168x128x84 and 250x290x14 respectively. The image values are then rescaled between zero and one to ease the learning process. The cropped volume of the centrum semiovale is relatively small in the craniocaudal direction (z-axis). Contrary to the other three brain regions, the complete volume of the centrum semiovale could not be fit in the memory of our graphics processing unit (GPU). Therefore, as input to our algorithm we kept only the slices surrounding the slice visually scored by the expert rater. We automatically identified this slice by segmenting the lateral ventricles with FreeSurfer, and selecting the slice 1 cm above, as defined by (Adams et al., 2013).

In the left column, Figure 5.4 shows one example of the preprocessed images for each region.

2.3 3D Convolutional Regression Network

Once the images are preprocessed, they are given as input to a convolutional neural network (CNN) similar to the one proposed in Chapter 1 but with skip connections between layers. This network computes the automated PVS scores using a combination of learned filters.

We train a different network for each region. There are two reasons for this. PVS can have a different shape depending on their location in the brain. For instance, in the hippocampi, the shape of PVS is more round, while in the centrum semiovale, PVS are more elongated. Differentiating from mimics is also region specific. For instance, motion artifacts affect mostly the centrum semiovale and have a much lower influence in the midbrain, and lacunar infarcts are often located in the basal ganglia.

Our CNN architecture (Figure 5.3) is similar to that of a small ResNet (He et al., 2016) adapted for regression in 3D images. Our CNN has two 3D 3x3x3 convolutional layers, followed by a 2x2x2 max-pooling layer, again two 3D 3x3x3 convolutional layers, a global average pooling layer, and a fully connected layer, combining the contribution of the different features into a single score. The output of the network is hence a scalar and spans $\mathbb{R}$. The first two convolutional layers have 32 filters each, and the last two convolutional layers have 64 filters each. The convolutions are zero-padded, and are followed by a ReLU activation. We use skip connections between the input and output of two successive convolutional layers, to allow the network to skip unnecessary operations and adapt its complexity to the tasks, which can ease the learning process (He et al., 2016). For instance, we expect the quantification of PVS to be simpler in the midbrain than in the centrum semiovale. When using skip connections, we concatenate the features maps, instead of summing them as proposed by He et al. (2016). There is little evidence that using either one or the other strongly impacts the performance. However, the concatenation is easier to implement as it does not require to have the same number of feature maps. In total, our model has less than 200 000 parameters.

For the regularization, we use on-the-fly data augmentation (translation, rotation and flipping), and when training with smaller sets, we used dropout (30%) after each convolutional layer and after the global pooling layer. See section 3.1 for details.

To train the network, we minimize the MSE loss function between the outputs of the network and the ground truth labels indicating the number of PVS in the given brain region.

In Chapter 1, we quantified PVS in the basal ganglia also with regression CNN, but with a different architecture. There are three differences with the CNN we proposed in this chapter. Firstly, the proposed network is simpler and lighter. Experiments on the parameters of the network in Chapter 1, indeed suggested that simpler models performed equally good with enough training data. In our experiments, the training of deeper models was also much longer with small training set (400 scans), especially for the centrum semiovale and hippocampi. The second change is the introduction of skip connections between blocks. The third and last change is the use of global pooling instead of two fully connected layers of 2000 neurons. Using global pooling does not harm the performance and saves large amounts of GPU memory. This change was also proposed by He et al. (2016) over the architecture proposed by Simonyan and Zisserman (2015a), the preceding state-of-the-art neural network on the imagenet challenge (Deng et al., 2009).

2.4 Model Training

During training, a validation set is used to stop the optimization of network before over-fitting happens.

As mentioned in section 2.3, a separate model is trained for each region. The training of such models can be unpredictably long for the hippocampi and centrum semiovale. To speed up the training, we first train the models in the basal ganglia, as the convergence is faster there. Then we fine-tune the networks with the target region only (hippocampi or centrum semiovale). The training in the midbrain converges quickly and no pre-training is needed. We chose to pre-train with the basal ganglia and not in the midbrain, as PVS in the basal ganglia are more similar to PVS in the hippocampi and centrum semiovale.

2.5 Statistical Analyses

To evaluate associations between determinants of PVS and PVS scores, we used zero-inflated negative binomial regression models with the PVS score as outcome, as in the study of Adams et al. (2014). We used the ’glmmADMB’ package for generalized linear mixed models in R. The models were corrected for age and sex (except for the associations of age, sex respectively) and additionally for intracranial volume when computing associations with volumetric measures (white matter, gray matter, and cerebrospinal fluid). To account for the skewed distribution of WMH, we log transformed the WMH volumes. Continuous determinants were normalized by computing z-scores. Bonferroni correction was used, therefore associations with a p-value below $0.05/(20$ determinants $\times 4$ brain regions$) = 6.25 \times 10^{-4}$ were considered significant.

Table 5.1: **Characteristics of the Study Population, N=1485.**

Age, years, mean (SD)	74.61 (7.32)
Men, n (%)	679 (46)
Education, years, mean (SD)	12.06 (3.76)
Ever-smoker, n (%)	1001 (67)
Systolic blood pressure, mmHg, mean (SD)	150.25 (21.05)
Diastolic blood pressure, mmHg, mean (SD)	82.70 (11.07)
Diabetes mellitus, n (%)	219 (15)
Glucose, mmol/L, mean (SD)	5.71 (1.15)
Total Cholesterol, mmol/L, mean (SD)	5.48 (1.02)
HDL Cholesterol, mmol/L, mean (SD)	1.45 (0.39)
Body mass index, kg/m^2, mean (SD)	27.31 (3.87)
ApoE ϵ2 allele carrier, n (%)	241 (17)
ApoE ϵ4 allele carrier, n (%)	385 (26)
Cortical Infarct on MRI, n (%)	61 (4)
Lacunar Infarct on MRI, n (%)	192 (13)
White matter hyperintensity volume, mL, median [interquartile range]	6.72 [3.84 - 13.11]
Intracranial volume, mL, mean (SD)	1144.26 (117.89)
White matter volume, mL, mean (SD)	390.33 (58.47)
Gray matter volume, mL, mean (SD)	518.84 (53.77)
Cerebrospinal fluid volume, mL, mean (SD)	232.82 (53.83)

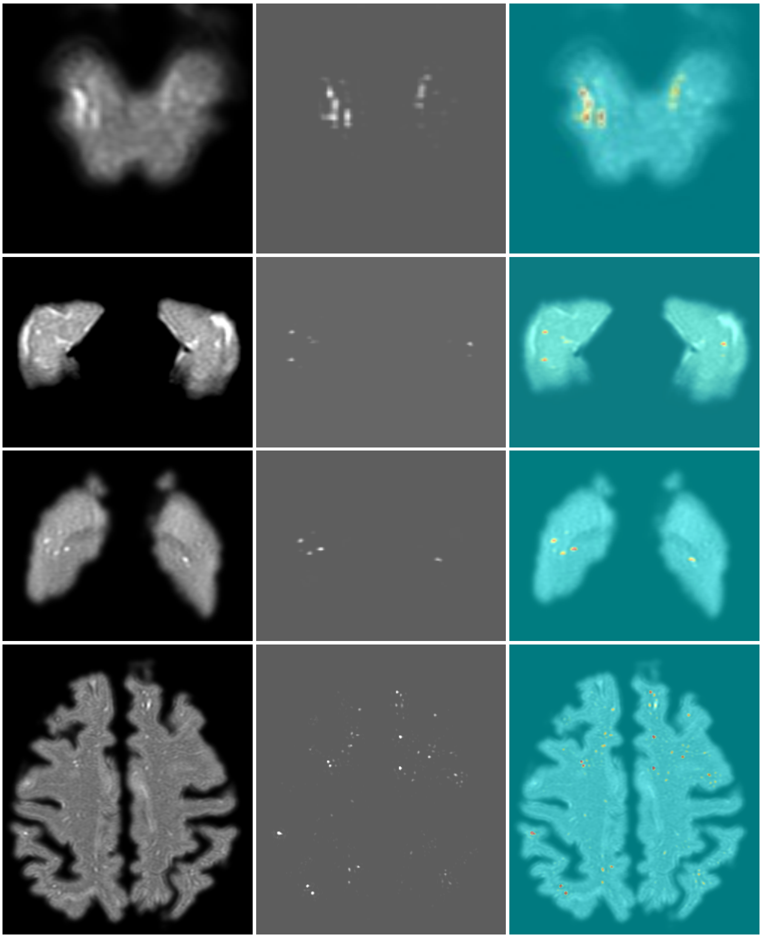


Figure 5.4: **Attention maps of the neural network.** From left to right: preprocessed input image, attention map, overlay of the input image and the attention map. From top to bottom: midbrain, hippocampi, basal ganglia, and centrum semiovale. In the overlay, the heatmaps reflect the contribution of pixels to the prediction of the networks: red pixels contributed the most, while blue pixel did not contribute. One can notice that many slightly enlarged perivascular spaces appear in orange. The network detected these, but they influenced its prediction less than the larger PVS.

3 Results

We evaluate the performance of the proposed model with three series of experiments. First, we inspect attention maps of the model, revealing that the model indeed focuses on PVS. Second, we measure the agreement between automated and visual scores, and show that this agreement is at least at the level of the human inter-observer agreement for each region. Then we verify the scan-rescan reproducibility of the automated PVS scores. Finally, we show that the associations between 20 determinants of PVS and the automated scores are similar to associations between the same determinants and visual scores.

3.1 Experimental Settings

We initialize the weights of the CNN by sampling from a Gaussian distribution, use Adadelta (Zeiler, 2012) for optimization, and augment the training data with randomly transformed samples. The transformation parameters for augmentation are uniformly drawn from an interval of 0.2 radians for rotation, 2 pixels for translation and flipping in the x and y direction. The network is trained per sample (mini-batches of a single 3D image). We implemented our algorithms in Python in Keras (Chollet et al., 2015) with Tensorflow as backend, and ran the experiments on a Nvidia GeForce GTX 1070 GPU and Nvidia Tesla K40 [2]. The average training time is one day. We stop the training after the validation loss converged to a stable value, or before over-fitting happens. Once the CNN is trained, the automatic PVS scoring, given the segmented region of interest, takes on average 287 ms per region.

[2] from Cartesius GPU cluster, a national cluster

3.2 Attention Maps

As first qualitative evaluation we check whether the neural networks learned to identify the structures of interest (PVS), or detected some other features that are correlated to the PVS. We use attention maps computed via ”guided backpropagation” (Springenberg et al., 2015).

Figure 5.4 shows examples of these attention maps for each of the four regions. We notice that the neural networks focus on the PVS, even though they are trained using global, image-wise labels only.

3.3 Agreement between Automated and Visual Scores

In this section, we evaluate the proposed automated scores by comparing with expert visual scores. We optimized the parameters of the CNN on a set of 1600 scans (1200 for training and 400 for validation). We also optimized the same model using only a subset of 400 scans (320 training and 80 validation), where we used dropout after each convolution to avoid over-fitting. We evaluated both models on an independent set: the remaining 515 scans. The results are reported in Table 5.2. Figure 5.5 shows Bland Altman plots for each region. Note that on the Bland Altman plots, the discrete nature of the distribution of the points, especially visible for the midbrain and hippocampi, is a consequence of the visual PVS scores being integer numbers.

When trained on 1600 scans, the ICC between the automated and visual scores were higher than the inter-observer agreement previously reported for each region. On the Bland Altman plots, one can notice that the largest errors usually occur for scans with many PVS, and for which there are only few training examples. Also, even for expert raters the rating becomes more difficult and variable for scans with many PVS. This is due to the continuous nature of the enlargement of perivascular spaces: keeping a consistent threshold of enlargement becomes more challenging.

3.4 Reproducibility

The reproducibility of the automated PVS scores is evaluated on a reproducibility set of 30 participants scanned twice (see Section 2.1). The ICC of the automated PVS scores between the first and second sets of scans is 0.82 for the midbrain, 0.93 for the hippocampi, 0.92 for the basal ganglia, and 0.87 for the centrum semiovale. Except for the centrum semiovale, all values are higher than the intra-rater agreement computed on another subset of the same dataset and reported by Adams et al. (2013) (Table 5.2).

3.5 Associations with determinants of PVS

We investigate associations between 20 potential determinants of PVS (characteristics in Table 5.1) and the automated PVS scores, and compare them with the associations between the same determinants and the visual PVS scores. The neural networks are first optimized using 400 scans for each region (we reuse the second model from the previous section), and then applied to the remaining 1715 independent scans to produce the automated scores. We investigate associations on this set of 1715 scans. After excluding participants as described in section 2.1, this resulted in 1485 stroke-free and non-demented participants with available brain imaging.

Figure 5.6 shows forest plots for each determinant, and a sorted list of all p-values can be found in supplementary materials. Overall, association patterns are very similar for visual and automated scores.

We found that white matter hyperintensity volume is associated with both visual and automated PVS scores in the basal ganglia and in the hippocampi. Age is associated with both visual and automated PVS scores in the basal ganglia. The presence of lacunar infarcts is also associated with both visual and automated PVS scores in the basal ganglia. And finally, intracranial volume is associated with both visual and automated PVS scores in the centrum semiovale. In all cases, determinants that are significantly associated with visual PVS scores, also show significant association with the automated PVS scores, and in almost the same order of p-values.

As the automated method takes as input the MRI scans, and is only optimized using global labels (the number of PVS), in the scans other information than PVS might be used to compute the automated PVS scores. This is an unwanted behavior. We did not notice any bias of the automated method towards more significant associations with imaging markers. For instance, for both visual and automated PVS scores, 9 of the 20 most significant associations were between imaging markers and PVS scores. However, computing the p-value of the difference of z-scores of the associations showed a significant difference for gray matter and PVS scores in the basal ganglia. In Figure 5.6, we notice the same trend for the association between intracranial volume and PVS scores in the basal ganglia. There was also a significant difference (though with a higher p-value) for associations between intracranial volume and PVS scores in the hippocampi.

Computing the p-value of the difference of z-scores of the associations revealed a last significant difference: the association between age and the automated PVS scores in the midbrain (odds-ratio 1.008 [1.002-1.0013]) was significantly stronger than the association between age and the visual scores in the midbrain (odds-ratio 0.999 [0.992-1.006]).

Table 5.2: **Agreement between automated and visual PVS scores for each brain region.** The metric reported is the intraclass correlation coefficient (ICC), computed on an independent set of 515 scans. These ICCs are compared to the inter-observer and intra-observer agreements reported by Adams et al. (2013). Note that the inter-observer and intra-observer agreements were computed on a different subset of the same dataset (Section 2.1).

Region	Intra-observer Agreement	Inter-observer Agreement	Trained on 1600 scans	Trained on 400 scans
Midbrain	0.82	0.75	0.75	0.74
Hippocampi	0.85	0.82	0.88	0.74
Basal Ganglia	0.80	0.62	0.82	0.73
Centrum Semiovale	0.88	0.80	0.86	0.80

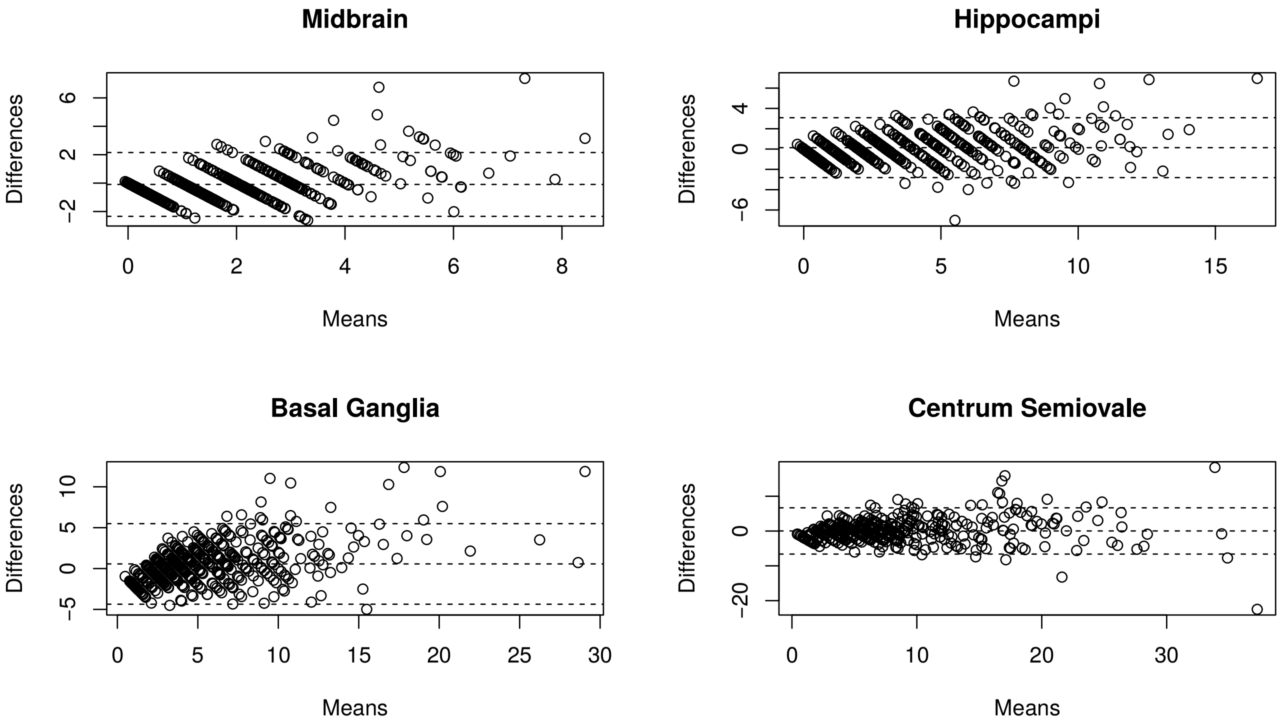


Figure 5.5: **Bland-Altman plots between the automated and visual PVS scores in the four regions.** The algorithms were optimized with 1600 scans, and evaluated on 515 scans. For the differences, the automated scores were subtracted from the visual scores.

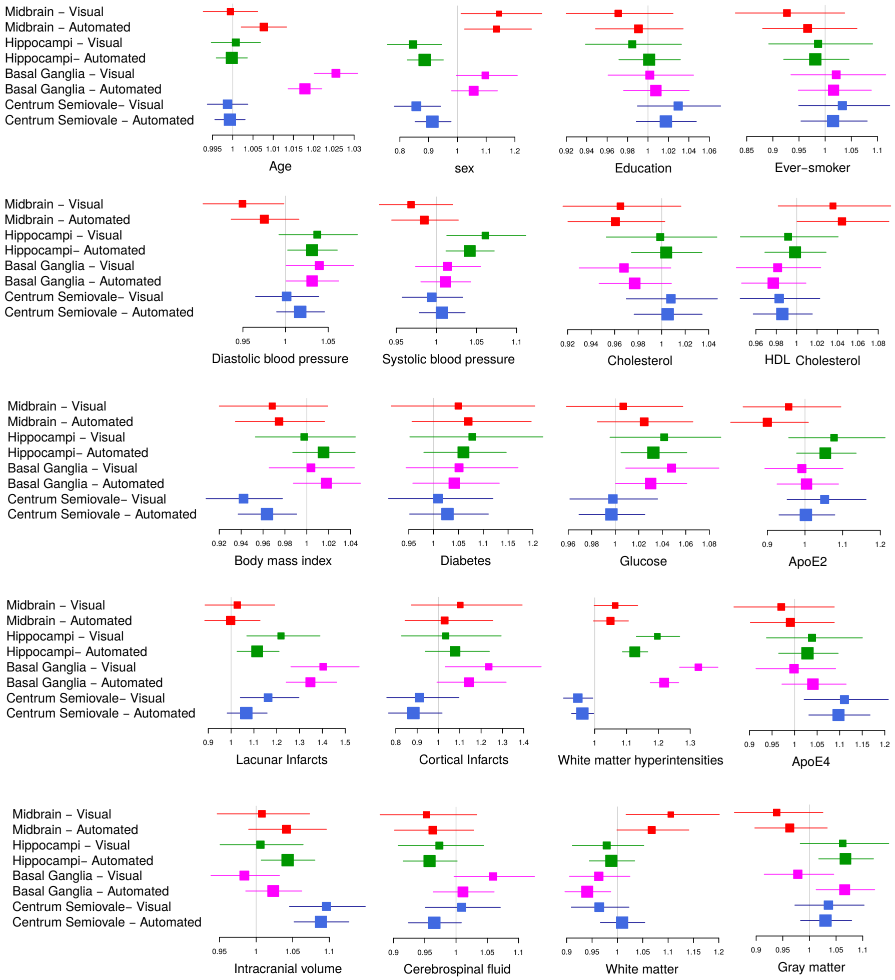


Figure 5.6: **Associations between determinants of PVS and PVS scores.** Odds ratio with 95% confidence intervals (non Bonferroni corrected). Characteristics of the study population are given in Table 5.1. The size of the colored boxes is inversely proportional to the size of the confidence intervals of the odds ratio.

4 Discussion

The algorithm developed in this work computes automated scores to quantify enlarged perivascular spaces (PVS) in the midbrain, hippocampi, basal ganglia and centrum semiovale - the four brain regions currently deemed most clinically relevant for PVS quantification. We demonstrated the performance of our algorithm using a set of 2115 MRI scans that were visually scored by an expert rater. For all four regions, the intraclass correlation coefficient between the automated scores and the visual scores was found to be higher than the inter-observer agreement, which was previously computed on a smaller subset of the same study population (Adams et al., 2013). Scan-rescan reproducibility was high (ICC 0.82 - 0.93). We also demonstrated the application of our automated scores by verifying the associations between determinants of PVS and our automated scores in a test set of 1485 scans, and comparing these associations to the visual scores. Based on these results, we believe that our automated scores could ultimately replace visual scores in future research projects studying the etiology and clinical relevance of PVS.

Automated PVS scores have two major advantages over visual scores: they are more objective (because the algorithm is deterministic), and can be computed more quickly. While a trained expert rater needs several minutes to score a scan, the computation of the automated PVS score on modern hardware (GPU) lasts less than a second. This makes our automated approach suited to be used in large scale studies, investigating for instance the etiology of PVS, their distribution in brain aging, their implications, and their potential as a biomarker for early diagnosis of cerebral small vessel disease. In addition, our method could be extended to fully quantify PVS by assessing their volume with the attention maps produced by the neural networks (Figure 5.4). These attention maps indeed provide a voxel-wise probability of PVS presence, which can for instance be summed over a region of interest, to yield a total volume or burden of PVS.

Gonzalez-Castro et al. (2017) computed automatic binary scores of PVS burden in the basal ganglia and investigated associations with determinants of PVS. They found significant associations with higher age, Fazekas WMH scale, and the presence of lacunar infarcts, while there was no significant associations with brain atrophy, hypertension, or stroke subtype. In the current study, we found the same significant associations (age, WMH, and presence of lacunar infarcts) for the basal ganglia.

As mentioned in section 3.5, other imaging information than PVS should not be used by the algorithm to compute the automated PVS scores. Apart from the associations between gray matter volume and PVS scores in the basal ganglia, we did not notice any strong trend of our method towards a stronger association with imaging markers. This difference of association in the basal ganglia most probably results from the automated PVS scores being computed across the complete volume of the basal ganglia, while visual PVS scores are rated in a single slice (Section 2.1). The consequences of this difference have been thoroughly investigated in Chapter 1, and seems to favor the automated PVS scores, as they are less sensitive to perturbations, such as missed PVS.

There is increasing evidence that ageing affects PVS, and putative mechanisms are dysfunction of the blood-brain barrier, or impaired perivascular drainage (Brown et al., 2018). Higher age was previously shown to be associated with higher visual PVS scores in the four regions investigated in this chapter: midbrain, hippocampi, basal ganglia, and centrum semiovale (Adams et al., 2014). The study by Adams et al. (2014) has been carried out in a significantly larger population study (3146 participants against 1485 for our study). In the current study, age was only associated with visual PVS scores in the basal ganglia. Higher age was also associated with higher automated PVS scores in the basal ganglia. Previous studies with visual PVS scoring have shown similar associations with age and basal ganglia PVS (Gutierrez et al., 2013; Martinez-Ramirez et al., 2013; Potter et al., 2015b). In the current study, in comparison with visual PVS scores, the automated PVS scores showed a significantly higher association power in the midbrain, which may suggest that they better capture the burden of PVS than visual scores. We did not find significant associations between age and PVS in the hippocampi or in the centrum semiovale (neither with visual PVS scores, nor with automated PVS scores). Similarly, in a recent study on a 7T scanner by Bouvy et al. (2016), no association was found between age and PVS in centrum semiovale. While Adams et al. (2014) found the weakest association between age per decade and PVS to be in the hippocampi (odds ratio of 1.07 [1.02-1.12]), they also found the strongest association between age per decade and PVS to be in the centrum semiovale with an odds ratio of 1.24 [1.19-1.30]. While there seems still to be controversy in the detailed relationship between age and PVS, automated PVS scores could possibly be more powerful to better disentangle possible mechanisms of PVS which effect brain health in ageing.

The main limitation of this work is that, contrary to the UNIVRSE rating system (Adams et al., 2015), the method was evaluated using MRI scans acquired on a single scanner, precluding the assessment of performance on different datasets. However, we believe this method can easily be applied to other datasets by only fine-tuning the CNN parameters on a few scans (Yosinski et al., 2014). Besides the performance of the algorithm should also be evaluated in multi-center or multi-scanner data.

5 Conclusion

We present a regression method to automatically quantify the number of enlarged perivascular spaces in the midbrain, hippocampi, basal ganglia, and centrum semiovale. The automated scores are more objective than visual scores and less time consuming. We validated our approach on 1485 brain MRI scans, demonstrated that the automated PVS show good agreement with visual PVS scores, and showed that the automated PVS scores are associated with several determinants of PVS, in a similar fashion to the PVS visual scores. We believe that this method could replace visual scoring of PVS in epidemiological and clinical studies, and therefore advance research into the etiology of PVS and its potential as a risk indicator of small vessel disease.

Acknowledgment

This work was partly carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. This research was funded by The Netherlands Organisation for Health Research and Development (ZonMw) Project 104003005, with additional support of Netherlands Organisation for Scientific Research (NWO), project NWO-EW VIDI 639.022.010 and project NWO-TTW Perspectief Programme P15-26.

Chapter 6

Automated Quantification of Enlarged Perivascular Spaces in Clinical Brain MRI across Sites

Abstract

Enlarged perivascular spaces (PVS) are structural brain changes visible in MRI, and are a marker of cerebral small vessel disease. Most studies use time-consuming and subjective visual scoring to assess these structures. Recently, automated methods to quantify enlarged perivascular spaces have been proposed. Most of these methods have been evaluated only in high resolution scans acquired in controlled research settings. We evaluate and compare two recently published automated methods for the quantification of enlarged perivascular spaces in 76 clinical scans acquired from 9 different scanners. Both methods are neural networks trained on high resolution research scans and are applied without fine-tuning the networks' parameters. By adapting

the preprocessing of clinical scans, regions of interest similar to those computed from research scans can be processed. The first method estimates only the number of PVS, while the second method estimates simultaneously also a high resolution attention map that can be used to detect and segment PVS. The Pearson correlations between visual and automated scores of enlarged perivascular spaces were higher with the second method. With this method, in the centrum semiovale, the correlation was similar to the inter-rater agreement, and also similar to the performance in high resolution research scans. Results were slightly lower than the inter-rater agreement for the hippocampi, and noticeably lower in the basal ganglia. By computing attention maps, we show that the neural networks focus on the enlarged perivascular spaces. Assessing the burden of said structures in the centrum semiovale with the automated scores reached a satisfying performance, could be implemented in the clinic and, e.g., help predict the bleeding risk related to cerebral amyloid angiopathy.

1 Introduction

Enlarged perivascular spaces (PVS) are structural brain changes visible on MRI. They can be identified as thin hyperintense tubular structures on T2-weighted MRI scans. PVS are increasingly thought to reflect the presence of cerebral small vessel disease, which represents a leading cause of cognitive decline and functional loss in elderly patients. In most studies, enlarged perivascular spaces are quantified using visual scores that either classify the burden of PVS in several categories (Potter et al., 2015a), or count PVS (Adams et al., 2013). These quantification methods are tedious and observer-dependent. Several methods have been proposed to automatically quantify PVS burden in Chapter 5 and by other researchers (Boespflug et al., 2017; Sudre et al., 2018; Zhang et al., 2016). None of these methods have been evaluated in clinical scans, which present multiple challenges for the quantification of PVS. While in research studies, the scanning is highly standardized (same machine, same protocol, same scanning parameters, same investigators, etc.) to yield comparable results, this is not the case in clinical routine. The lower resolution of clinical scans also results in the computation of less accurate shape features, the most discriminative feature for the detection of PVS. Moreover, other MRI markers related to cerebral small vessel disease – such as white matter hyperintensities – are more prevalent in clinical scans than in population studies (as in Chapter 5 and Boespflug et al. (2017)) and could be confused with PVS because of their similar appearance.

In most studies, PVS are quantified separately in one or several clinically and epidemiologically relevant brain regions: midbrain, hippocampi, thalamus, basal ganglia, and centrum semiovale. In PVS research, the centrum semiovale is the most studied region, as PVS burden there has been most strongly associated to potential determinants of PVS and outcomes thereof. The centrum semiovale is also often the region with highest inter-observer agreement in the visual scoring of PVS (Adams et al., 2013). In this study, we quantified PVS in the hippocampi, basal ganglia, and centrum semiovale.

Zhang et al. (2016) automatically quantified PVS on 7T MRI scans. Boespflug et al. (2017) proposed an automated quantification method combining image intensities and morphologic features from several MRI sequences. They evaluated their method in the centrum semiovale in research scans. Sudre et al. (2018) proposed to use recurrent neural networks to detect PVS and lacunar infarcts in 16 subjects of a longitudinal study investigating the relationship between cardiovascular risk factors and brain health. van Wijnen et al. (2019) regressed intensity distance maps of PVS in the centrum semiovale using neural networks. In Chapter 5, we proposed to quantify PVS burden in four brain regions – midbrain, hippocampi, basal ganglia, and centrum semiovale – with neural network regressors trained with image level labels: the count of PVS in the target brain region. In research scans, the authors showed that they could reach a correlation between visual scores and automated scores similar to that of the inter-observer agreement in each region. They also found that associations between 20 determinants of PVS and visual PVS scores, and between the same determinants and automated PVS scores, were similar. We also proposed to use a more advanced model (GP-Unet) for weakly supervised detection of enlarged perivascular spaces in Chapter 3. This method estimates simultaneously the number of PVS and a high resolution attention map that can be used to detect and segment PVS. We decided to study the methods presented in Chapters 3 and 5 as the validation experiments with associations with clinical variables already brought them one step ahead of other methods for the application to clinical practice.

In this chapter, we applied and compared the two methods corresponding to Chapters 3 and 5 on 76 clincial MRI scans with a varying, low resolution acquired in the clinical routine of a hospital using nine different scanners and different protocols, while using models' weights learned from high-resolution population study MRI scans acquired at another hospital in a highly controlled and standardized setting using a single scanner and protocol. The networks were not fine-tuned to the clinical data. For preprocessing, we used FSL packages instead of FreeSurfer parcellations as in Chapters 3 and 5 to segment the regions of interest. Finally, we show examples of attention maps of GP-Unet.

2 Datasets

2.0.0.1 Training data. The training data consists of 1600 T2-weighted MRI scans from 1600 elderly participants in a population study: the Rotterdam Study (Ikram et al., 2017). Scans were acquired on a single 1.5T GE scanner, in a highly controlled and standardized setting. The scan resolution was 0.5x0.5x0.8 mm^3. PVS were visually scored by a single rater in all scans in the hippocampi, basal ganglia and centrum semiovale, following the guidelines of Adams et al. (2013).

2.0.0.2 Evaluation data. The MRI data used for evaluation were gathered retrospectively from the Picture Archiving and Communication System (PACS) of University Hospital Magdeburg. MRI scans with visible signs of cerebral small vessel disease were selected. All selected patients had cerebral microangiopathy, and were diagnosed with at least one of the following: ischemic (i.e. lacunar) stroke or transient ischemic attack, spontaneous intracerebral hemorrhage, dementia (i.e. Alzheimer's disease or vascular dementia), and epileptic seizures. Initially, 100 acquisitions from 100 different patients were collected. 24 Scans were excluded from the experiments either because FSL segmentation of the brain structures failed or because scans could not be rated visually, e.g. due to insufficient image quality caused by motion artifacts or presence of other

Table 6.1: Characteristics of the clinical dataset (minimum, maximum, mean and standard deviation)

	min	max	mean	std
Patient age (years)	35	89	71.39	9.32
In-plane (axial) resolution (mm^2)	0.39	0.68	0.45	0.04
Resolution in z (mm)	3.30	7	4.94	0.89
Spacing between slices (mm)	0.60	6.60	4.73	1.04

pathologies such as extremely large lesions. This leaves a total of 76 scans for the study. Since the acquisitions have been obtained during the clinical routine, they present a considerable variance with respect to various image properties such as artifacts or image resolution. T1-weighted and T2-weighted MRI scans have been acquired with 9 different scanners. Two of these scanners, a 3T and a 1.5T from Philips, make up 66 of the 76 images. In total, there are three 3T-, four 1.5T- and two 1T-scanners. Three of them were Siemens (two 3T, one 1.5T), the rest were Philips machines. The time frame in which the data was acquired is almost 15 years and ranges from August 2004 until March 2019. The majority of the scans (43) has been acquired within the last 5 years of this period. The number of male and female patients was 46 and 30, respectively. Table 6.1 provides additional information about the data set. PVS were scored visually in the hippocampi, basal ganglia and centrum semiovale following the guidelines of Adams et al. (2013). Two raters scored PVS, the inter-rater agreement is reported in Table 6.2.

3 Methods

The target brain regions (hippocampi, basal ganglia, and centrum semiovale) are first segmented, masked and cropped. The result is then processed by trained convolutional neural networks that predict the count of PVS in each region. The neural networks were trained with high resolution MRI scans of a population study, but were used to predict PVS count in routine clinical scans of a hospital. The study was approved by the local ethics committee (No 28/16).

3.1 Preprocessing

To match the resolution of scans in the training set, all clinical scans were linearly interpolated to a resolution of 0.5x0.5x0.8 mm^3.

In Chapters 3 and 5, we used FreeSurfer parcellations to segment brain regions. FreeSurfer brain parcellation lasts usually several hours, which may prevent its use in clinical routine. In this study, we used instead FIRST and FAST algorithms from the FSL package (Smith et al., 2004) to segment brain regions from the T1 sequence in a matter of minutes. FIRST could compute segmentation of the basal ganglia and hippocampi. FAST was used to segment the white matter for the centrum semiovale region. In Chapters 3 and 5, we also evaluated our methods in the midbrain. As midbrain segmentation is not implemented in FSL, this region was excluded from the study. The T1 sequence was then rigidly registered to the T2 sequence using FSL FLIRT, and the segmentation labels were propagated from the T1 space to the T2 space.

Following the guidelines of Adams et al. (2013) for visual scoring of PVS, in Chapters 3 and 5, we quantified PVS in the centrum semiovale in the neighborhood of the slice located 1 cm above the top of the lateral ventricles. As

FSL does not compute ventricle segmentation, we used instead the segmentation of the basal ganglia as approximation, and selected the slice 1 cm above the top of the caudate nucleus.

The following preprocessing steps were computed exactly as described in Chapter 5. Namely, the segmentation masks were dilated, convolved with a gaussian kernel to smooth the border of the mask, and multiplied pixelwise with the T2 intensities. The masked regions were then cropped, normalized between 0 and 1 using the minimum and maximum intensity values in the masked region, and given as input to the neural networks.

3.2 Neural Networks

The preprocessed images were given as input to two different types of neural networks proposed for automated PVS quantification: (1) a neural network with four convolutional layers and a max-pooling layer which outputs the number of PVS in a region (Chapter 5) and that we call *CNN*, and (2) *GP-Unet*, a similar neural network proposed (Chapter 3, in which the downsampling path is followed by an upsampling path to enable weakly supervised detection of PVS. Networks of both methods were trained with only image-level labels.

Attention maps of GP-Unet were computed to visualize the focus of the networks using a linear combination of the feature maps of the last convolutional layer, as described in Chapter 3.

4 Results and Discussion

Table 6.2 shows the Pearson correlation, and Table 6.3 the mean absolute error, between visual and automated PVS scores for each region and for each method, and the corresponding inter-rater agreement. Scatter-plots are shown in Figure 6.1. Attention maps of GP-Unet are displayed for each region in Figure 6.2.

There was no noticeable difference in the computation of the regions of interest when using FSL masks instead FreeSurfer masks, but the interpolation to 0.5x0.5x0.8 mm^3 was needed to reuse the networks optimized on high resolution scans. The visual PVS scores were highly correlated to the automated PVS scores of GP-Unet in the centrum semiovale (0.78 Pearson correlation), were moderately correlated in the hippocampi (0.52), and a lower correlation in the basal ganglia (0.28). Attention maps of GP-Unet (Figure 6.2) show that, as expected, the method focuses on perivascular spaces.

While on research scans, CNN and GP-Unet reached a similar performance in all regions, our experiments on clinical scans show that the correlation between visual PVS scores and automated PVS scores of GP-Unet was significantly higher than that of visual PVS scores and automated scores of CNN in the centrum semiovale (Williams' test, p-value < 0.0001) and in the hippocampi (p-value $<$ 0.05). Contrary to CNN, GP-Unet combines features of different scales via skip connections, which may have assisted the computation of discriminative shape features, and improved the detection of single PVS, as opposed to detecting – or missing because of their too large size – a cluster of PVS without being able to individually count them.

The correlation in the basal ganglia (0.31 for GP-Unet) is lower than in the other regions and is notably lower than the inter-rater agreement (0.56). Attention maps (Figure 6.2) show that the network only detects the largest PVS in the basal ganglia, and misses less enlarged PVS. The scatter-plots (Figure 6.1) seem to confirm this observation: in the basal ganglia, the networks underestimate the number of PVS, and predict similarly low numbers of PVS for all scans.

Table 6.2 shows lower inter-rater agreement for the basal ganglia than for the other regions. This might be a consequence of PVS being visually rated only in a single slice in this region (Adams et al., 2013). The low resolution of clinical scans in z direction might cause a large variability in the selection of this slice, which might negatively influence the reproducibility of the visual rating. The automated methods quantify PVS in the complete volume of the basal ganglia, which was previously shown to be more reproducible than the visual PVS scores (Chapter 1). Interestingly, the automated PVS scores of both methods – CNN and GP-Unet – are highly correlated in the basal ganglia (0.73 Pearson correlation). The correlation between their scores was higher in the basal ganglia than in other regions.

Results in the centrum semiovale (0.78 Pearson correlation) are similar to the inter-rater agreement (0.75). This is also close to the inter-rater agreement (0.80 intraclass correlation coefficient) as reported in earlier studies in high resolution research scans (Adams et al., 2013). Demonstrated quantification of PVS burden in the centrum semiovale could aid in the better stratification of cerebral small vessel disease subtypes, i.e. hypertensive arteriopathy and cerebral amyloid angiopathy, especially in large and hospital-based cohorts. This would presumably have important therapeutic and prognostic implications in terms of prescribing oral anticoagulants and preventing intracerebral hemorrhage. This is of particular importance in cerebral amyloid angiopathy, that has not only been related to severe PVS burden in the centrum semiovale (Charidimou et al., 2017), but also to a significantly higher risk for intracerebral bleeding in face of oral

Table 6.2: **Correlation between visual and automated PVS scores.** Pearson correlations between the first rater and GP-Unet, CNN, and the second rater for each region. Correlations were all significant (p-value < 0.01). Significant correlations after Bonferroni correction are in bold.

	GP-Unet	CNN	R2
Centrum Semiovale	**0.78**	**0.52**	**0.75**
Basal Ganglia	0.31	0.25	**0.56**
Hippocampi	**0.51**	0.33	**0.64**

Table 6.3: **Mean absolute errors between visual and automated PVS scores.** Mean absolute error between the first rater and GP-Unet, CNN, and the second rater for each region.

	GP-Unet	CNN	R2
Centrum Semiovale	5.58	6.39	4.67
Basal Ganglia	5.67	5.49	3.78
Hippocampi	2.58	3.0	2.08

anticoagulant treatment (Wilson et al., 2018).

In future work, the results in the basal ganglia and the hippocampi may be improved by fine-tuning the neural networks using the clinical dataset, and by adding data augmentation during training with research scans to imitate the resolution of clinical scans and contrast variations between different scan protocols or scanners. The results presented are already promising considering the large differences between training and test sets.

The complete computation of the automated PVS scores lasts only a few minutes on CPU. Most of the computation time is spent on FSL brain structures segmentation and registration from the T1-weighted scans to the T2-weighted scans. After this preprocessing, the computation of the automated PVS scores took only about 6 seconds per brain region on CPU. This low computation time can facilitate the implementation of such a method in clinical practice.

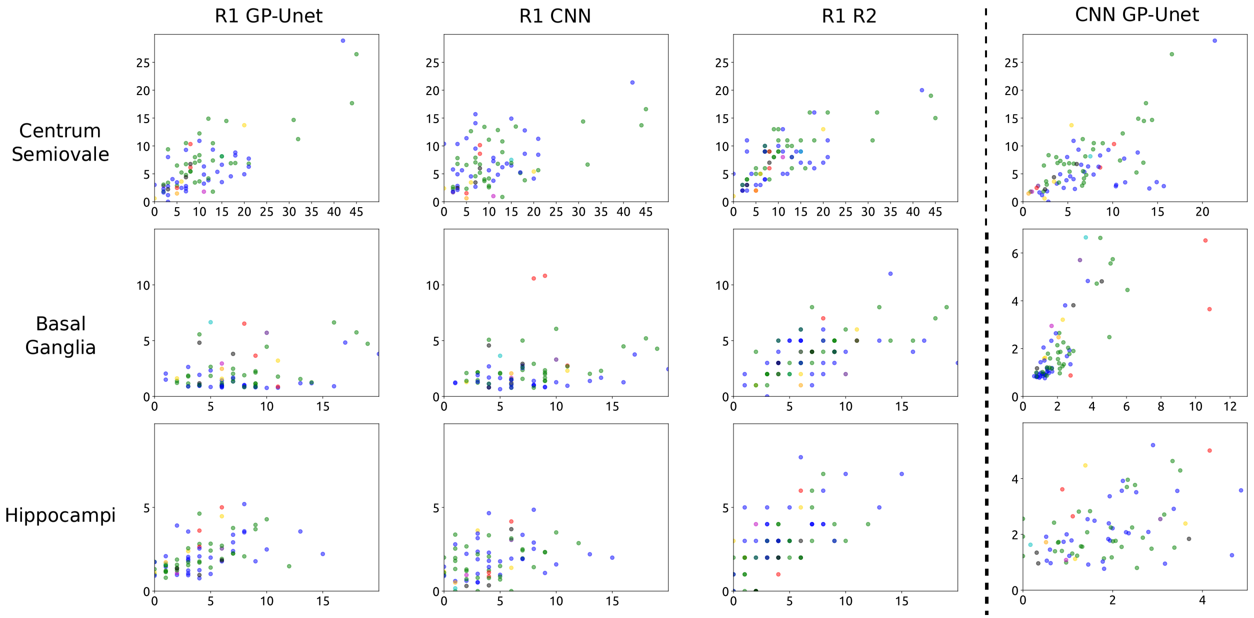


Figure 6.1: **Comparison between visual and automated PVS scores.** The different colors represent different scanners. The visual PVS scores of the first rater (R1), on the x-axis, are compared with the predictions of GP-Unet, CNN, and with the visual scores of the second rater (R2), on the y-axis. In the right column we plotted the automated PVS scores of GP-Unet versus those of CNN.

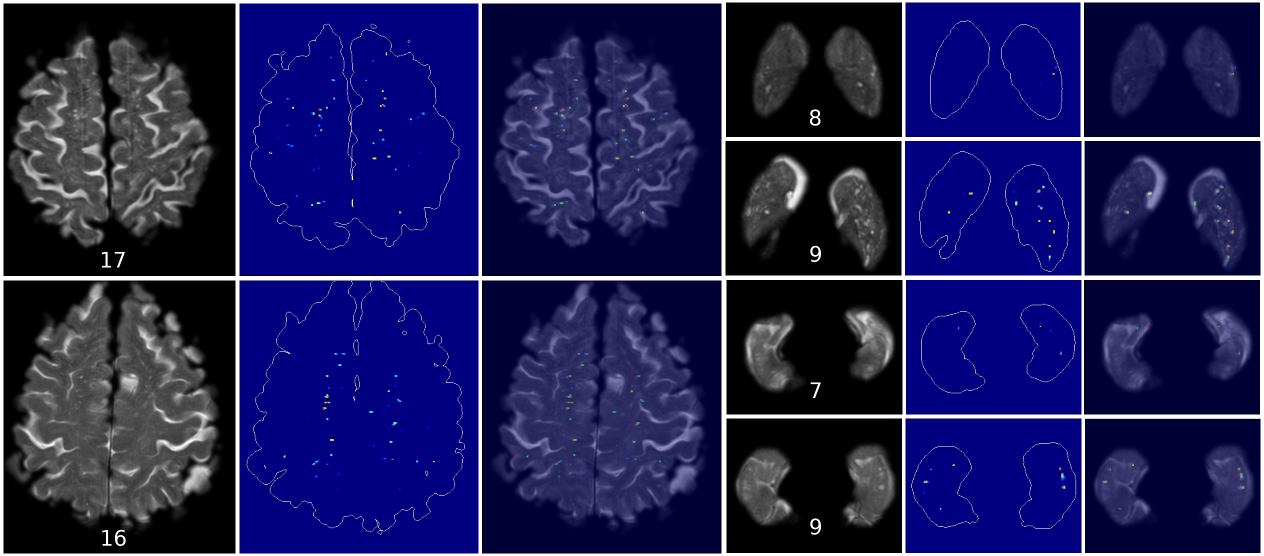


Figure 6.2: **Attention Maps of GP-Unet in an axial view.** Attention maps for the centrum semiovale are displayed on the left, for the basal ganglia on the top right, and for hippocampi, on the bottom right. Visual scores are indicated below each region. For each selected image, from left to right, we show the original image, the attention map with drawn contours of the region, and the overlay of both. The colormaps of the attention maps were manually adjusted for each image. Highlighted structures are considered as PVS by the networks. The redder a structure is, the higher is its weight in the computation of the automated PVS scores by the network. For the centrum semiovale, we selected two images that correspond to an average agreement between automated and visual score (human rater R1). For the basal ganglia and hippocampi, we selected one image with poor agreement (top), and another image with good agreement (bottom).

5 Conclusion

We showed that PVS burden could be automatically quantified in the centrum semiovale in clinical scans, with an agreement with visual scores that was similar to the inter-observer agreement. Automated PVS scores were computed with a neural network that was trained high-quality research scans and with only global labels of PVS burden. These results could contribute to bringing automated PVS quantification to the clinic and guide the administration of anti-coagulant drugs.

Acknowledgements

This work received funding from the Netherlands Organisation for Health Research and Development (ZonMw - Project 104003005) and the federal state of Saxony-Anhalt, Germany (Project I 88).

Part E

Neural networks for other applications in neuroimaging research

Chapter 7

Event-Based Modeling with High-Dimensional Imaging Biomarkers for Estimating Spatial Progression of Dementia

Abstract

Event-based models (EBM) are a class of disease progression models that can be used to estimate temporal ordering of neuropathological changes from cross-sectional data. Current EBMs only handle scalar biomarkers, such as regional volumes, as inputs. However, regional aggregates are a crude summary of the underlying high-resolution images, potentially limiting the accuracy of EBM. Therefore, we propose a novel method that exploits high-dimensional voxel-wise imaging biomarkers: n-dimensional discriminative EBM (nDEBM).

nDEBM is based on an insight that mixture modeling, which is a key element of conventional EBMs, can be replaced by a more scalable semi-supervised support vector machine (SVM) approach. This SVM is used to estimate the degree of abnormality of each region which is then used to obtain subject-specific disease progression patterns. These patterns are in turn used for estimating the mean ordering by fitting a generalized Mallows model. In order to validate the biomarker ordering obtained using nDEBM, we also present a framework for Simulation of Imaging Biomarkers' Temporal Evolution (SImBioTE) that mimics neurodegeneration in brain regions. SImBioTE trains variational auto-encoders (VAE) in different brain regions independently to simulate images at varying stages of disease progression. We also validate nDEBM clinically using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). In both experiments, nDEBM using high-dimensional features gave better performance than state-of-the-art EBM methods using regional volume biomarkers. This suggests that nDEBM is a promising approach for disease progression modeling.

1 Introduction

In 2015, approximately 46.8 million people were estimated to be living with dementia, and by 2050 this number is expected to have increased to 131.5 million (Prince et al., 2015). Dementia is characterized by a cascade of neuropathological changes which are quantified using several imaging and non-imaging biomarkers. Understanding how the different biomarkers progress from normal to abnormal state after disease onset enables precise estimation of disease severity in an objective and quantitative way. This can help in identifying individuals at risk of developing dementia as well as monitor the effectiveness of preventive and supportive therapies.

Event-based models (EBM) are a class of disease progression models that estimate the order in which biomarkers become abnormal during disease progression using cross-sectional data (Fonteijn et al., 2012; Venkatraghavan et al., 2019; Young et al., 2014; Huang and Alexander, 2012). It was reported in a recent paper on discriminative EBM (DEBM) (Venkatraghavan et al., 2019) that the EBMs are very sensitive to the quality of biomarkers used for building the model. Hence, to infer the neuropathological changes that occur during dementia accurately, good quality biomarkers are important.

An essential step in an EBM involves mixture modeling to obtain biomarker distributions in normal and abnormal classes (Fonteijn et al., 2012; Venkatraghavan et al., 2019). This restricts the current EBMs to only handle scalar biomarkers. In case of imaging biomarkers, regional volumes from structural MRIs are often used (Venkatraghavan et al., 2019; Oxtoby and Alexander, 2017; Young et al., 2018, 2014; Fonteijn et al., 2012). However, regional volumes are a crude summary of the high-dimensional information available from structural MRI, resulting in suboptimal EBM performance, as shall be demonstrated later in this chapter. Therefore, we propose a novel method that exploits voxel-wise imaging biomarkers: n-dimensional discriminative EBM (nDEBM).

Estimating the accuracy of ordering obtained by EBMs is not feasible as ground-truth ordering is not known for a disease. In order to validate the proposed method and compare its accuracy with that of existing state-of-the-art EBM methods, we also present a framework for Simulation of Imaging Biomarkers' Temporal Evolution (SImBioTE). SImBioTE uses variational auto-encoders (VAE) to simulate neurodegeneration in brain regions. These regions are represented by a vector in the latent space of the VAE. Synthetic brain regions were created by sampling latent representations corresponding to target degrees of abnormality which were determined by a ground-truth ordering of disease progression. The generated synthetic brain regions were used as inputs for nDEBM, and the regional aggregates were used as inputs for state-of-the-art EBMs to evaluate the accuracies.

2 nDEBM

In Section 2.1, a brief introduction to the current DEBM (Venkatraghavan et al., 2019) model is given. Section 2.2, presents a novel framework to use semi-supervised SVMs in DEBM for estimating posterior probabilities of abnormality for high-dimensional biomarkers. In Section 2.3, we use these posterior probabilities to estimate severity of disease progression in an individual.

2.1 DEBM

In a cross-sectional dementia dataset (X) of M subjects (consisting of cognitively normal (CN) and patients with dementia (DE)), let X_j denote a measurement of biomarkers for subject $j \in [1, M]$, consisting of N scalar biomarker values $x_{j,i}$. As dementia is characterized by a cascade of neuropathological changes that occurs over several years, even CN subjects can show some abnormal biomarker values. On the other hand, in DE subjects, a proportion of biomarkers may still have normal values, especialy in patients at an early disease stage. This leads to label noise in the data and hence clinical labels cannot directly be propagated to individual biomarkers. The DEBM model introduced by Venkatraghavan et al. (2019), similar to previously proposed EBMs (Fonteijn et al., 2012; Huang and Alexander, 2012; Young et al., 2014), fits a Gaussian mixture model (GMM) to construct the normal and abnormal distributions. These are used to compute pre-event and post-event likelihoods $p(x_{j,i}|\neg E_i)$ and $p(x_{j,i}|E_i)$ respectively, where an event E_i is defined as the corresponding biomarker becoming abnormal. The mixing parameters are used as prior probabilities to convert these likelihoods to posterior probabilities $p(\neg E_i|x_{j,i})$ and $p(E_i|x_{j,i})$.

$p(E_i|x_{j,i})\forall i$ are used to estimate the subject-specific orderings s_j. s_j is established such that:

$$s_j \ni p(E_{s_j(1)}|x_{j,s_j(1)}) > p(E_{s_j(2)}|x_{j,s_j(2)}) > ... > p(E_{s_j(N)}|x_{j,s_j(N)}) \tag{7.1}$$

Finally, DEBM computes the central event ordering S from the subject-specific estimates s_j. To describe the distribution of s_j, a generalized Mallows model is used. The central ordering is defined as the ordering that minimizes the sum of distances to all subject-specific orderings s_j, with probabilistic Kendall's Tau being the distance measure.

2.2 n-Dimensional Biomarker Progression

It was reported by Venkatraghavan et al. (2019) that the accuracy of EBMs depends on the quality of biomarkers used to build the model. Greater separability of individual biomarkers results in estimation of more accurate event ordering. We hypothesize that high-dimensional imaging biomarkers can increase the separability between the normal and abnormal groups, thus improving the accuracy when used as inputs to EBMs. The use of GMM in EBMs however restricts it to using only scalar or low-dimensional biomarkers as GMMs do not scale well to high-dimensional features. SVMs do scale well to high-dimensional features, but a supervised soft-margin SVM cannot be used because of the large amounts of label noise (up to one third of the elderly CN population could be in pre-symptomatic stages of DE (Schott et al., 2010)). In this section, we present a way in which scalable semi-supervised SVM classifiers can be used within the DEBM framework with high-dimensional inputs.

Let $X_{j,i}$ denote the high-dimensional imaging biomarker for brain region i. Since the clinical diagnosis of the subject cannot be propagated to each region, the labels cannot be trusted while training a classifier. If we were to train a classifier trusting these labels, independently on each biomarker ($X_{\forall j,i}$), we hypothesize that labels of the data close to the decision boundary or on either side of it cannot be completely trusted for that biomarker. For identifying the labels that cannot be trusted for a biomarker, we propose to train a linear classifier assuming equal class-priors. Fitting a non-linear classifier risks over-fitting to the wrongly-labeled data whereas class-priors derived from labeled data could be misleading as some of the labels might be wrong, for that biomarker.

For biomarker $X_{\forall j,i}$, subjects whose labels are preserved are considered as labeled data ($X_{\mathbb{L},i}$). Subjects whose labels have been rejected, along with any prodromal subjects in the dataset are considered as unlabeled data ($X_{\mathbb{U},i}$). Semi-supervised classifiers can be used in this context for obtaining the decision boundary for each biomarker.

To identify the subjects for whom labels can be trusted when considering $X_{\forall j,i}$, we first train a linear SVM $(f_{0;i})$ based on CN and DE subjects. After rejecting labels that cannot be trusted (with distance $d_{0;i} < |d_t|$ from the decision boundary), we use semi-supervised learning with EM (Nigam et al., 2000) using linear SVM with subject-specific costs (Brefeld et al., 2003) $(f_{1;i}, ..., f_{k+1;i})$ to iteratively refine the decision boundary. The algorithm for this semi-supervised classification is given below:

[H] Semi-Supervised SVM Learning with Subject-specific weights [1] $i \in \{1...N\}$ Train $f_{0;i}$ with $X_{\forall j \in \{CN,DE\},i}$ as inputs $d_{0;\forall j,i} \leftarrow$ prediction of $X_{\forall j,i}$ using $f_{0;i}$ $j \in \{1...M\}$ $d_{0;j,i} > |d_t|$: $X_{\mathbb{L},i} \leftarrow X_{j,i}$: $X_{\mathbb{U},i} \leftarrow X_{j,i}$ Estimate $\hat{p_0}(E_i|X_{\mathbb{U},i})$ from $d_{0;\mathbb{U},i}$ (using Platt scaling (Platt et al., 1999)). Train $f_{1;i}$ using $X_{\forall j,i}$ using $|\hat{p_0}(E_i|X_{\mathbb{U},i})$ - $\hat{p_0}(\neg E_i|X_{\mathbb{U},i})|$ as weights of $X_{\mathbb{U},i}$. Estimate $\hat{p_1}(E_i|X_{\mathbb{U},i})$ from $d_{1;\mathbb{U},i}$ $k \leftarrow 1$ $||\hat{p}_k(E_i|X_{\mathbb{U},i}) - \hat{p}_{k-1}(E_i|X_{\mathbb{U},i})||^2 < \epsilon$ Train $f_{k+1;i}$ using

$X_{\forall j,i} \ni |\hat{p_k}(E_i|X_{\mathbb{U},i})$ - $\hat{p_k}(\neg E_i|X_{\mathbb{U},i})|$ are weights of $X_{\mathbb{U},i}$. Estimate $\hat{p}_{k+1}(E_i|X_{\mathbb{U},i})$ from $d_{k+1;\mathbb{U},i}$. $k \leftarrow k+1$ Estimate $\hat{p}_{k+1}(E_i|X_{\forall j,i})$ from $d_{k+1;\forall j,i}$ $p(E_i|X_{j,i}) \leftarrow \hat{p}_{k+1}(E_i|X_{j,i})$

d_t was chosen such that such that 5% of correctly classified data closest to decision boundary are treated as unlabeled. The weights for $X_{\mathbb{U},i}$ in the above algorithm is motivated based on the work of Brefeld and Scheffer (2004). It is done because unlabeled data close to the decision boundary are not the ideal support vectors. The samples which are farther away from the decision boundary of the previous iteration can be trusted more as support vectors for the next iteration of training.

2.3 Patient Staging

Patient staging refers to the process of positioning individuals on a disease progression timeline characterized by the obtained event ordering. Patient stage (Υ_j) is computed as an expectation of event-centers (λ_n) with respect to $p(n, S, X_j)$, where n denotes the possible discrete stages in the timeline characterized by N biomarker events. Event-centers are the positions of the biomarker events on a normalized disease progression timeline $[0, 1]$, that capture relative distances between events.

$$\Upsilon_j = \frac{\sum_{n=1}^{N} \lambda_n p(n, S, X_j)}{\sum_{n=1}^{N} p(n, S, X_j)} \tag{7.2}$$

$p(k, S, X_j)$ can be expressed in-terms of posterior probabilities of events obtained from semi-supervised SVM as:

$$p\left(n, S, X_j\right) \propto \prod_{i=1}^{n} p\left(E_{S(i)}|X_{j,S(i)}\right) \times \prod_{i=n+1}^{N} p\left(\neg E_{S(i)}|X_{j,S(i)}\right) \tag{7.3}$$

3 SImBioTE: A Validation Framework

For validating classical EBMs and nDEBM in a unified framework, we extend the framework developed by Young et al. (2015) for simulating datasets consisting of scalar biomarkers, to be capable of generating datasets with realistic voxel-wise imaging biomarkers. It was built on the assumption that the trajectory of biomarker progression follows a sigmoid. Using a similar assumption, we consider the degree of abnormality in different regions $(a_{j,i})$ follows a sigmoidal trajectory.

$$a_{j,i}(\Psi) = \frac{1}{1 + \exp(-\rho_i(\Psi - \xi_{j,i}))} + \epsilon \tag{7.4}$$

Ψ denotes disease stage of a subject which we take to be a random variable distributed uniformly throughout the disease timeline. ϵ is the equivalent of measurement noise, which represents randomness in the measurement of abnormality. ρ_i signifies the rate of progression of a biomarker, which we take to be equal for all subjects for all biomarkers. It was shown by Venkatraghavan et al. (2019) that the performance of EBMs is similar for equal $\rho_i \forall i$ and unequal ρ_i. $\xi_{j,i}$ denotes the disease stage at which the biomarker becomes abnormal.

After randomly choosing degrees of abnormalities for different regions, we use a variational autoencoder (VAE) (Kingma and Welling, 2014) for each region i, to generate 3D images of these brain regions at a target degree of abnormality $a_{j,i}(\Psi)$. VAEs are neural networks consisting of two main components: an encoder E which projects input images into a lower dimensional space $\mathbb{R}^K$ called the latent space, and a decoder D which generates images from their hidden representation in the latent space $Z \in \mathbb{R}^K$. Once the VAE has been trained using a large dementia dataset, a latent representation $Z_{j,i;t}$ corresponding to the target degree of abnormality $a_{j,i}(\Psi)$ can be sampled in the latent space. The decoder D then generates a 3D image $D(Z_{j,i;t})$ corresponding to $a_{j,i}(\Psi)$. Below we describe the VAE used in this work, and the sampling strategy in the latent space.

3.1 Implementation of the Convolutional Variational Autoencoder

Figure 7.1 summarizes the architecture of our VAE. We use a ReLU activation after each convolutional layer, except after the last 1*1*1 convolutional layer. We implemented the loss function as proposed by Kingma and Welling (2014), with mean-square-error (MSE) and Kullback-Leibler divergence. We optimized the network with Adadelta (Zeiler, 2012).

3.2 Sampling Strategy in the Latent Space

To navigate in the latent space $\mathbb{R}_i^K$ of region i, we use Euclidean geometry. We first build a scale vector U_i in the latent space to describe the range of the disease from CN to DE. In order to generate a point $Z_{j,i;t} \in \mathbb{R}_i^K$ at the target degree of abnormality $a_{j,i}(\Psi)$, we first randomly sample a point $Z_{j,i;s} \in \mathbb{R}_i^K$, and translate it along the direction of the scale vector U_i until we reach the target abnormality $a_{j,i}(\Psi)$.

3.2.1 Scale Vector from Cognitively Normal to Dementia.

To build the scale vector U_i, we first compute the latent representations of all the images of region i in the training dataset by projecting these images in the latent space $\mathbb{R}_i^K$ using the encoder E. Then we use the binary labels – CN and DE – of each subject j to compute the means $\mu_{i;CN} \in \mathbb{R}_i^K$ and $\mu_{i;DE} \in \mathbb{R}_i^K$, and standard deviations $\sigma_{i;CN} \in \mathbb{R}_i^K$ and $\sigma_{i;DE} \in \mathbb{R}_i^K$ for each of the two categories respectively.

This is followed by computing the vector joining the two mean points as $u_i = \mu_{i;DE} - \mu_{i;CN}$. The idea is to create a vector U_i spanning the range of the disease progression, from CN to DE. However, u_i joins only the means, if we want to capture the whole distribution, we need to lengthen this vector by a multiple of the standard deviations, on both sides: for instance by $3\sigma_{i;CN}$ in the CN side, and $3\sigma_{i;DE}$ on the DE side. To do so, we compute the scalar projections of the standard deviations as $\sigma_{i;CNp} = |\sigma_{i;CN}.\widehat{u_i}|$ and $\sigma_{i;DEp} = |\sigma_{i;DE}.\widehat{u_i}|$, where $\widehat{u_i} = u_i/||u_i||_2$. Now we can compute the new origin point (CN) as $O = \mu_{i;CN} - 3\sigma_{i;CNp}\widehat{u_i}$, and the new end point (DE) as $M = \mu_{i;DE} + 3\sigma_{i;DEp}\widehat{u_i}$. Finally, we can compute $U_i = M - O$. Note that $\widehat{U_i} = U_i/||U_i||_2 = \widehat{u_i}$.

3.2.2 Navigation for generation

We first randomly sample a point $Z_{j,i;s}$ using the mean and standard deviation of the latent representations of all subjects j for region i. The degree of abnormality $a_{j,i;s}$ of this randomly sampled point $Z_{j,i;s}$ can be computed as $a_{j,i;s} = OZ_{j,i;s}.\widehat{U_i}/||U_i||_2$. To reach the target point $Z_{j,i;t}$, we need to translate the randomly sampled point $Z_{j,i;s}$. This now can be done by computing $Z_{j,i;t} = Z_{j,i;s} + (a_{j,i;t} - a_{j,i;s})U_i$. To generate the corresponding brain region we can now use the decoder and compute $D(Z_{j,i;t})$.

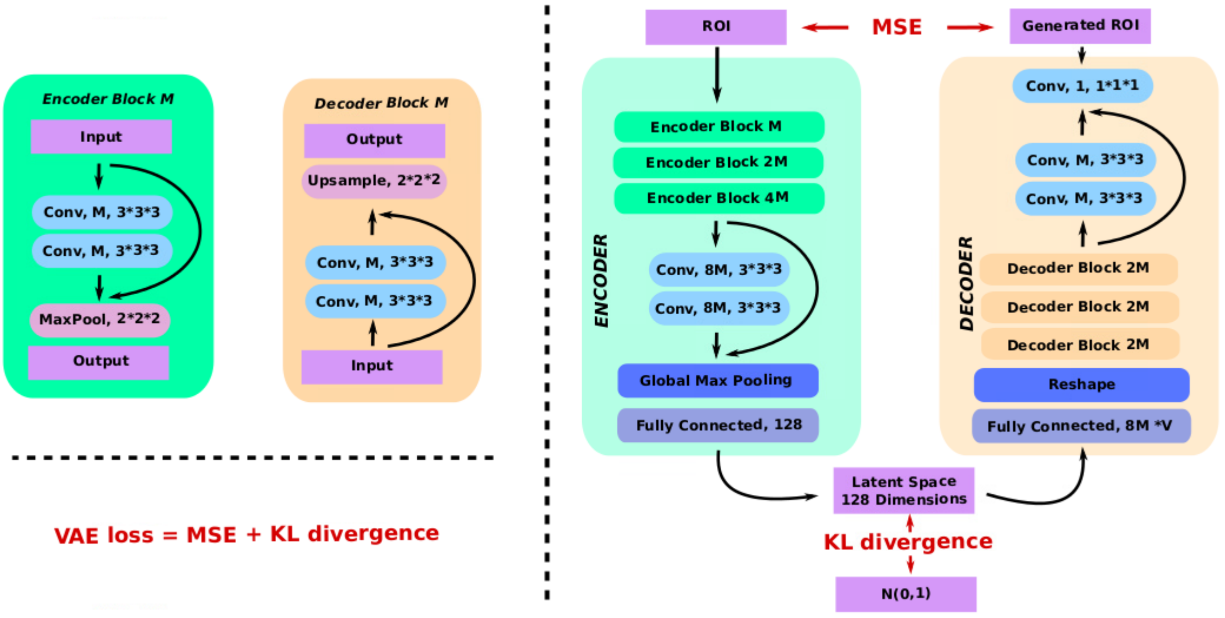


Figure 7.1: Architecture of the Variational Autoencoder.

4 Experiments and Results

This section describes the experiments performed to validate the proposed nDEBM algorithm and also compare it with classical EBM (Fonteijn et al., 2012) and DEBM (Venkatraghavan et al., 2019) algorithms.

4.1 ADNI Data

We considered 1737 ADNI subjects (417 CN, 106 with significant memory concern (SMC), 872 with mild cognitive impairment (MCI) and 342 AD subjects) who had a 1.5T structural MRI (T1w) scan at baseline. This was followed by multi-atlas brain extraction using the method described by Bron et al. (2014). Gray matter (GM) volumes of segmented regions were regressed on age, sex and intra-cranial volume (ICV) and the effects of these factors were subsequently corrected for. Student's t-test between CN and AD was performed on these confounding factor corrected GM volumes and 15 regions with smallest p-values were retained. They were subsequently used as inputs for DEBM and EBM (Fonteijn et al., 2012) models. The optimization routine proposed by Venkatraghavan et al. (2019) was used to train the GMM in these two models.

The T1w images were registered to a common template space based on the method used by Bron et al. (2014). Probabilistic tissue segmentations were obtained for white matter (WM), GM, and cerebrospinal fluid on the T1w image using the unified tissue segmentation method (Ashburner and Friston, 2005). The voxel-wise GM density maps were computed based on the Jacobian of the local deformation map and the probabilistic GM volume. The GM density maps from the corresponding 15 regions were used as inputs for nDEBM.

4.1.1 Model Validation

Since the groundtruth ordering is not known in a clinical setting, validation of these models was done based on the resulting patient stages for classifying AD subjects from CN as well as for classifying MCI non-converters (MCI-nc) from converters (MCI-c)[1]. We performed 10-fold cross-validation with 10 repetitions. The training set was used to train the three models. The disease timeline created during training was used to stage the patients in the test-set.

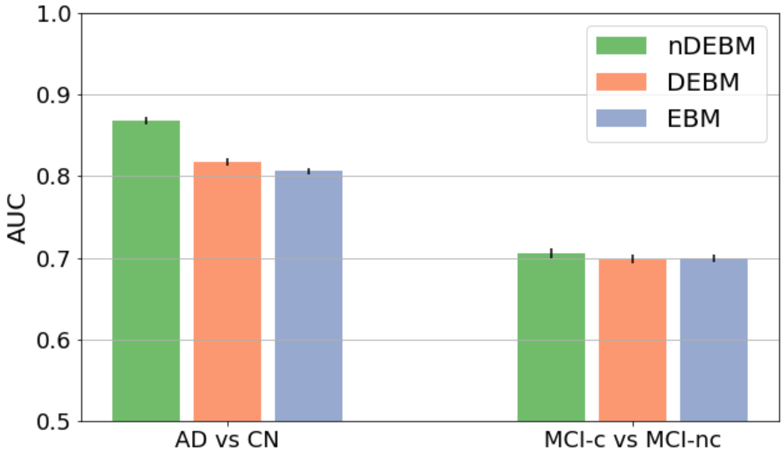


Figure 7.2: AUC measures when patient stages of nDEBM, DEBM and EBM were used for classifying AD vs CN (left) and MCI-c vs MCI-nc (right). The error bar represents the standard deviation in 10 random repetitions.

Figure 7.2 shows the results of 10 random repetitions of 10-fold cross-validation on ADNI dataset. The error-bar shows the standard deviation

[1] MCI converters are subjects who convert to AD within 3 years of baseline measurement

of the AUCs when the patient stages obtained from nDEBM, DEBM and EBM were used to classify AD vs CN and MCI-c vs MCI-nc.

4.1.2 Uncertainty in Estimation

Variation of the positions of the biomarker events on a normalized disease progression timeline (event-centers) estimated by nDEBM and DEBM was studied by creating 100 bootstrapped samples of the data and applying nDEBM on those samples [2].

Figure 7.3 shows event-centers estimated by nDEBM and DEBM along with the uncertainty in their estimations. The biomarkers are ordered along the y-axis based on the event-ordering obtained by nDEBM.

4.2 Simulation Data

In our experiments, $\xi_{j,i}\ \forall j$ are random variables with $\mathbb{N}(\mu_{\xi_i}, \Sigma_{\xi_i})$. μ_{ξ_i} were equally spaced for different i. The value of Σ_{ξ_i} was set to be $\Delta\xi$ where $\Delta\xi$ is the difference in μ_{ξ_i} of adjacent events. ρ_i was considered to be equal for all biomarkers. Ψ of the simulated subjects were distributed uniformly throughout the disease timeline.

We first trained 15 VAEs (one per selected region) on the GM density maps of the ADNI dataset. Then we generated - as detailed in Section 3 - images for these 15 regions and for 1737 artificial subjects according to pre-computed degrees of abnormality as defined in Equation 7.4. These degrees of abnormality are different for each region and each subject. We repeated this process 10 times, with different random simulations. The voxel-wise GM density maps of regions were used for obtaining the ordering using nDEBM. The GM volume of

[2] EBM was left out of this experiment as the concept of event-centers was not introduced for EBM.

the simulated regions (computed by integrating the GM density map over the region of interest) were used as biomarkers for DEBM and EBM.

SimBioTE results depicting Lateral occipitotemporal gyrus atrophy in simulated images is shown in Figure 7.4. The images thus generated were used for validating different EBM methods.

The errors made by different EBM methods on SImBioTE data are shown in Figure 7.5. The estimated ordering and the ground-truth orderings were compared using Kendall's Tau distance.

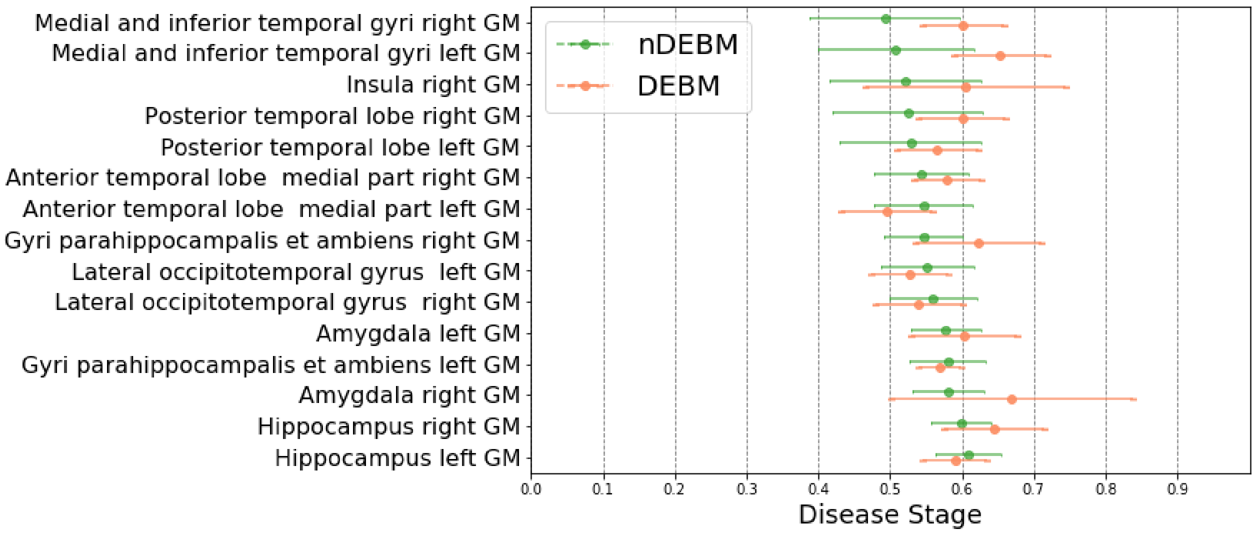


Figure 7.3: Variation of event-centers estimated by nDEBM and DEBM in 100 bootstrapped samples of the ADNI data. The error bar represents the standard deviation of the respective event-centers.

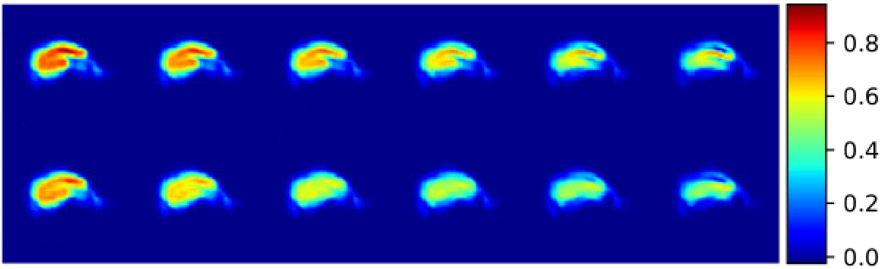


Figure 7.4: An example of lateral occipitotemporal gyrus (right) atrophy as simulated by SImBioTE. The interpolation spans the full range U_i, as described in section 3. Left is normal (CN) and right is abnormal (DE). The two rows shows disease progression in two different simulated subjects.

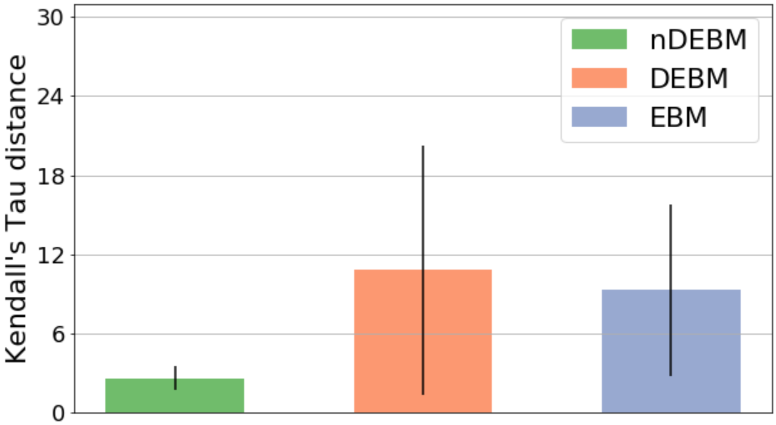


Figure 7.5: Inaccuracies, as measured by Kendall's Tau distance from groundtruth, of nDEBM, DEBM and EBM. The error bar represents the standard deviation of the errors made in 10 repetitions of simulations.

5 Discussions

We proposed a novel method (nDEBM) that exploits high-dimensional voxel-wise imaging biomarkers for event-based modeling using semi-supervised SVM. This was validated based on ADNI dataset, where the spatial spread of structural abnormality was estimated based on a cross-sectional dataset. However this is an indirect validation of the orderings based on accuracy of the estimated patient stages, since the ground-truth ordering for clinical data is unknown.

To unambiguously validate the orderings obtained, we also proposed a new simulation framework (SImBioTE) to simulate voxel-wise imaging biomarkers based on training VAEs on different regions. It is known that GM tissue is lost in AD progression. Therefore the voxel-wise GM density maps will become darker as the disease progresses, as can be observed in Figure 7.4. It was also observed in Figure 7.4 that simulated regions for different subjects shows considerable variations. This shows that the simulation framework is capable of generating datasets with realistic atrophy and with good inter-subject variability. This, in combination with the scalar biomarkers' simulation framework, results in images where the disease progression in different regions can be controlled. However, a more thorough validation of the simulation framework by comparing the atrophy patterns of the simulated data with that of real-life longitudinal data is needed to understand the effect of different model parameters. Possible extensions of SImBioTE includes simulating whole brain images from these independent regions, which can be used to validate wider range of disease progression models.

The datasets simulated by SImBioTE were used for inputs for different EBMs. It was observed in Figure 7.5 that the orderings obtained by nDEBM are much closer to the ground-truth as compared to DEBM and EBM. It was also observed in Figure 7.2 that the patient stages obtained by nDEBM delineates AD and CN subjects much better than the ones obtained by DEBM and EBM. The AUCs of classifying MCI-c vs MCI-nc are also marginally better for nDEBM as compared to the other two methods. These experiments serve as a validation for our initial hypothesis that increasing the dimensionality of the inputs helps in better delineation of normal and abnormal regions, which increases the accuracy of the resulting ordering. It can hence be concluded that the voxel-wise data helps nDEBM in estimating the disease progression more accurately than regional volumes. However, the choice of hyper-parameters in nDEBM (for e.g. d_t, SVM slack parameters) was done ad-hoc. The effect they have on the accuracy of the resulting ordering needs to be studied through more rigorous validation experiments.

The difference in event orderings obtained by nDEBM and DEBM as observed in Figure 7.3 suggests that the two types of inputs can lead to very different results. Hence, computing regional aggregates, such as volumes, and using that as inputs for EBMs as done by Venkatraghavan et al. (2019); Oxtoby and Alexander (2017); Young et al. (2018, 2014); Fonteijn et al. (2012) is not an optimal choice for estimating the spatial progression of disease.

6 Conclusion

We hypothesized that high-dimensional imaging biomarkers would result in better delineation of normal and abnormal regions thus leading to more accurate event-based models. We hence proposed a novel method (nDEBM) that exploits high-dimensional voxel-wise imaging biomarkers based on semi-supervised SVM to estimate temporal ordering of neuropathological changes in the brain structure using cross-sectional data. We also proposed a simulation framework (SImBioTE) using variational auto-encoders that mimics neurodegeneration in brain regions to validate nDEBM. Furthermore, we applied nDEBM framework to a set of 1737 subjects from ADNI dataset for clinically validating the method. In both experiments, nDEBM using high-dimensional features gave better performance than state-of-the-art EBM methods using regional volume biomarkers. This served as a validation for our initial hypothesis. nDEBM thus presents a new paradigm for estimating spatial progression of dementia.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 666992. E.E. Bron is supported by the Hartstichting (PPP Allowance, 2018B011). F. Dubost is supported by The Netherlands Organisation for Health Research and Development (ZonMw) Project 104003005.

Chapter 8

Multi-atlas image registration of clinical data with automated quality assessment using ventricle segmentation

Abstract

Registration is a core component of many imaging pipelines. In case of clinical scans, with lower resolution and sometimes substantial motion artifacts, registration can produce poor results. Visual assessment of registration quality in large clinical datasets is inefficient. In this work, we propose to automatically assess the quality of registration to an atlas in clinical FLAIR MRI scans of the brain. The method consists of automatically segmenting the ventricles of a given scan using a neural network, and comparing the segmentation to the atlas'

ventricles propagated to image space. We used the proposed method to improve clinical image registration to a general atlas by computing multiple registrations - one directly to the general atlas and others via different age-specific atlases - and then selecting the registration that yielded the highest ventricle overlap. Finally, as an example application of the complete pipeline, a voxelwise map of white matter hyperintensity burden was computed using only the scans with registration quality above a predefined threshold. Methods were evaluated in a single-site dataset of more than 1000 scans, as well as a multi-center dataset comprising 142 clinical scans from 12 sites. The automated ventricle segmentation reached a Dice coefficient with manual annotations of 0.89 in the single-site dataset, and 0.83 in the multi-center dataset. Registration via age-specific atlases could improve ventricle overlap compared to a direct registration to the general atlas (Dice similarity coefficient increase up to 0.15). Experiments also showed that selecting scans with the registration quality assessment method could improve the quality of average maps of white matter hyperintensity burden, instead of using all scans for the computation of the white matter hyperintensity map. In this work, we demonstrated the utility of an automated tool for assessing image registration quality in clinical scans. This image quality assessment step could ultimately assist in the translation of automated neuroimaging pipelines to the clinic.

1 Introduction

Image registration has proven a fundamental part of many processing pipelines in the biomedical imaging field, establishing spatial correspondence between images and enabling subsequent group or cohort analyses. However, when using clinical, low resolution brain data, image registration can be challenging. E.g. in acute ischemic stroke populations, high-resolution image acquisition in the acute disease state is not possible due to clinical time constraints. Nonetheless, such clinical cohorts offer great amounts of untapped information due to the large number of samples available, often in the range of thousands of patients (Giese et al., 2017; Courand et al., 2019), which can be utilized to unveil spatial patterns of disease burden (Bilello et al., 2016; Schirmer et al., 2019b). Importantly, as clinical images have more variability than scans acquired primarily for research, they necessitate quality control steps after registration to ensure that no gross errors occurred in the process. Quantifying the registration quality, utilizing only intensity-based metrics such as mutual information or cross-correlation, is often not enough, and in practice registration quality is assessed using manual ventricle segmentations to evaluate the overlap between the patient data and the registration target, i.e. brain template or atlas (Ou et al., 2014; Dalca et al., 2016; Ganzetti et al., 2018).

Considerable work has been conducted to generate appropriate brain templates for image registration, using data from healthy young adults (Dickie et al., 2017) or age appropriate cohorts from the general population (Schirmer et al., 2019b). These templates can consequently be used for segmentation of brain structures, but often yields unsatisfactory results in clinical scans. For instance, outlining of the ventricles in such clinical scans is often done manually, or semi-automatically (Hussain et al., 2013; Xia et al., 2004). Manually outlining the ventricles is a time intensive step, and hinders quality assessment in large scale cohorts. Deep learning techniques have been developed to automatically segment structures in clinical quality scans, using for instance U-Net architectures (Schirmer et al., 2019a; Nikolov et al., 2018; Guerrero et al., 2018). Given enough training data, these techniques can reliably generate accurate, fully automated masks of the structures of interest. The use of a U-Net architecture has been proposed to generate automated segmentations of the lateral ventricles alone (Ghafoorian et al., 2018), and recently of the complete ventricular system (Atlason et al., 2019; Shao et al., 2019), showing promising results, which can be utilized in automated assessment of image registration quality.

Automated registration quality assessment methods can also be used to improve the registration results in atlas selection methods. Multi-atlas segmentation has for instance become an increasingly popular segmentation method in neuroimaging pipelines (Iglesias and Sabuncu, 2015). One of its simplest implementations is to register several atlases pairwise to an image, propagate the labels of the atlases in image space, and choose the final label for each voxel using majority voting. Probabilistic label fusion strategies have also been proposed, such as Wang et al. (2013) who proposed to exploit the intensity similarity between atlases and the target image in the neighborhood of each voxel. Robinson et al. (2019) recently proposed a method to perform automated quality control of segmentations of cardiovascular data from the UK biobank. The authors registered a set of annotated images to a test image with unknown ground truth. The labels were then warped using the deformation field from image registration, and the overlap between the warped labels and the predicted segmentation was used to estimated the segmentation performance. In other words, the segmentation of the image with unknown ground truth is compared to that of a multi-atlas segmentation, where smaller difference between segmentations are assumed to reflect higher segmentation quality. Instead of using the same set of atlases for multi-atlas segmentation, a most appropriate subset of atlases can also be selected. Recently, Antonelli et al. (2019) proposed for instance to select subsets of atlases for each target image using a genetic selection algorithm, and evaluated their method in cardiac and prostate data. To decrease the computation time of multi-atlas segmentation, Dewey et al. (2017) proposed to add an intermediary registration step to a template constructed from the set of the considered atlases, using for instance multivariate template construction algorithm. Creating robust registration methods to map clinical scans to atlases is key to the field of lesion-symptom mapping. For example, Biesbroek et al. (2013) studied lesion-symptom mapping with brain lesions, such as white matter hyperintensities and lacunes, in relation to cognition.

In this work, we developed a ventricle segmentation deep learning algorithm based on a 3D U-Net-like architecture to segment the complete ventricular system in each subject's fluid-attenuated inversion recovery (FLAIR) sequence and validated it in a multi-center, clinical dataset comprising 12 sites. The ventricle segmentation was then used to assess registration quality by comparing it – using the Dice similarity coefficient – to the ventricles of the atlas propagated to the target image space. Over all brain regions, due to its very discriminative image intensity values and its relatively large size, the ventricular system presents a feature of the brain that is robust to variations in scanners and FLAIR protocols, making it a prime candidate for using its segmentation to assess registration quality. This automated registration quality assessment method can be used not only to flag or discard erroneous registrations, but also to select the best registration. As an example, we proposed to use this automated registration quality assessment method to improve registration quality by designing a multi-atlas registration (MAR) framework. Instead of directly registering images to a single template (general atlas), each image was additionally registered to five different atlases corresponding to different age categories, which in turn have been registered to the general atlas. The best atlas was then selected using the automated registration quality assessment method, and used as a transitional registration step before warping the subject image to the common space. Contrary to the above-mentioned multi-atlas segmentation methods, the purpose of the proposed MAR method was to improve the results of registration to the common space, and not to improve the results of segmentation of brain regions in the target image. Finally, we used the proposed MAR framework to create voxelwise maps of white matter hyperintensity (WMH) burden in a set of acute ischemic stroke patients, where Dice coefficient thresholds were used to control the quality of registration. In summary, our main contributions are an algorithm for the segmentation of the complete ventricular system in clinical scans, the evaluation of ventricle overlap as registration quality metric, and a multi-atlas registration framework to improve registration of images to a common space.

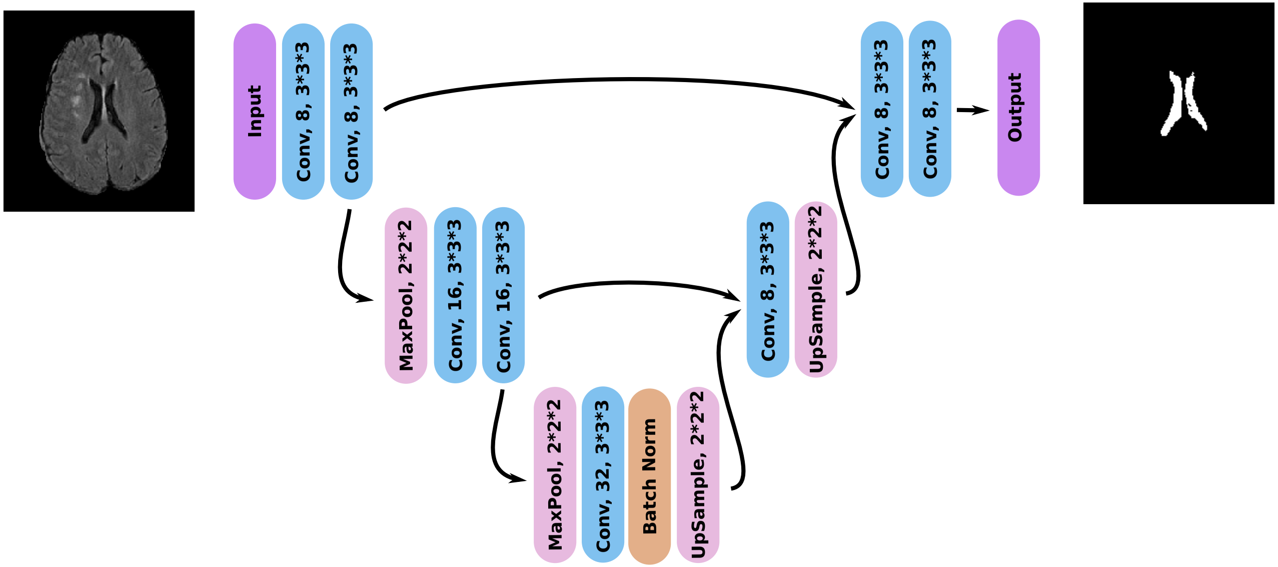


Figure 8.1: **Architecture of the deep learning ventricle segmentation algorithm.** The architecture is similar to that of a shallow 3D U-Net (Ronneberger et al., 2015) with only 104 feature maps to allow the processing of the full 3D images.

2 Material and Methods

2.1 Data

2.1.1 Onsite clinical data

We utilized data of the Genes Affecting Stroke Risk and Outcomes Study (GASROS) study (Zhang et al., 2015). Patients ($>$ 18 years old) presenting to the Massachusetts General Hospital Emergency Department (ED) between 2003 and 2011 with symptoms of acute ischemic stroke, were eligible for enrollment. Magnetic resonance images were acquired within 48 hours of admission and only patients with confirmed acute diffusion-weighted imaging lesions on brain MRI scans were included. 1132 patients underwent the standard acute ischemic stroke protocol on a 1.5T Signa scanner (GE Medical Systems), including T2-weighted FLAIR imaging (TR 5000ms, minimum TE of 62 to 116ms, TI 2200ms, FOV 220-240mm). For each patient, WMH were segmented using MRIcro software (University of Nottingham School of Psychology, Nottingham, UK; www.mricro.com), based on a previously published semi-automated method with high inter-rater reliability (Chen et al., 2006). Ventricles were manually segmented by a single rater in a subset of 300 patients' FLAIR images using 3D Slicer (Fedorov et al., 2012). Of the 300 scans, 100 were chosen to uniformly sample the age range in the GASROS cohort, 100 were chosen to span the range of WMH disease burden, and the remaining 100 were randomly selected. This set was used for network training and validation of the automated ventricle segmentation method. In addition, a test set of 100 patients were selected to approximately represent the range of ventricular volume in the patient population. Scans were selected with a semi-automated method that estimates ventricular volume using nonlinear registration to an atlas. The semi-automated method involved a quality control step to ensure that the range was uniformly sampled. These 100 scans were then segmented by a second rater.

2.1.2 Multi-center clinical data

The MRI-GENetics Interface Exploration (MRI-GENIE) study is a large-scale, international, hospital-based collaborative study of acute ischemic stroke patients (Giese et al., 2017), including FLAIR data from 12 sites (7 European, 5 US based), acquired as part of each hospital's clinical acute ischemic stroke protocol. For each acquisition site, 12 patients were selected (Schirmer et al., 2019a) and underwent manual ventricle segmentations. Two of the patients displayed substantial motion artifacts, and were excluded from our analysis, forming a total set of N=142 scans with manual brain and ventricle segmentation. This set was used as an additional test set for the evaluation of the ventricle segmentation algorithm and the proposed MAR framework.

2.1.3 ADNI data

Part of the data used in the preparation of this chapter were also obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

2.1.4 Brain atlases

Using 130 healthy controls from ADNI3 dataset (Jack et al., 2008) (Field strength 3T; 3D FLAIR; TE 119; TR 4800; TI 1650; 1.2x1x1mm3; see Appendix B for list of subject IDs), we created five FLAIR atlases, each corresponding to a different age category: under 70 years old (N=6 subjects), between 70 and 75 (N=22), between 75 and 80 (N=31), between 80 and 85 (N=39), and above 85 (N=32). The atlases were created using ANTs multivariate template construction algorithm with default parameters (Avants et al., 2011). Similarly, a general atlas was created by averaging the five age-specific atlases, also using using ANTs multivariate template construction algorithm with default parameters (Avants et al., 2011). All atlases were manually skull stripped and registered to MNI space. The resulting image resolution was 1mm3 and the image size 182x218x182 voxels. Ventricles were manually segmented in the general atlas. Each of the five age-specific atlases was diffeomorphically registered to the general atlas, to allow the propagation of the ventricle segmentation to age-specific atlases, and to warp the images to the general atlas space in the MAR framework. To assess which atlases were most similar to the general atlas, we computed the mean squared intensity difference between the age-specific atlases and the general atlas.

2.2 Automated ventricle segmentation

Image intensities were rescaled so that the 1st percentile of intensity values (without masking) is equal to 0 and the 99th percentile is equal to 1. The full 3D images were passed as input to a deep learning model. Prior to ventricle segmentation, each FLAIR image underwent brain extraction using a dedicated U-Net based deep learning method (Schirmer et al., 2019a) developed and

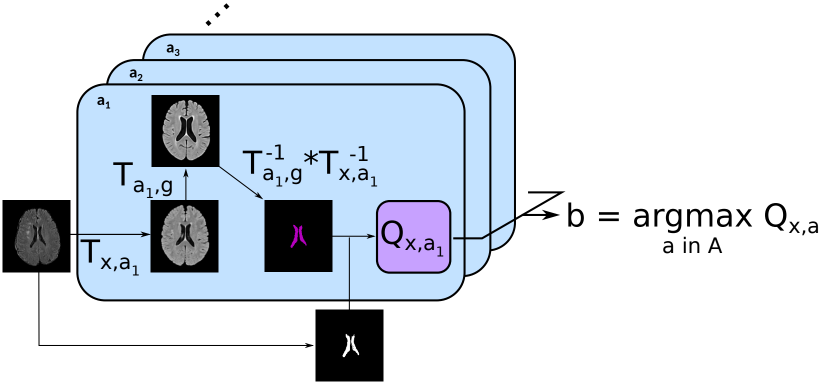


Figure 8.2: **Principle of the proposed MAR framework.** For each subject, the input image was first registered to each of the atlases $a \in A$, which had been previously registered to the general atlas. The ventricles segmented on the general atlas V_g are then propagated first to each atlas a, and then to the subject's image space. The propagated ventricles $V_{x,a,g}$ were subsequently compared to V_{CNN}, the subject's ventricles segmented using the proposed automatic algorithm. Finally, the atlas maximizing the registration quality was selected for the intermediary registration step.

validated in clinical scans. The resulting brain mask was also given as input to the model. While test data had varying voxel dimensions, training data consisted only of images with image size of 256x256 voxels in axial (inplane) direction, and less than 32 voxels in through plane direction. All images were then padded in z to have 32 slices. During inference, we resized images to 256x256x32 voxels using linear interpolation, predicted the corresponding ventricle maps, and resized these maps to the original image resolution.

We used a 3D U-Net-like architecture (Figure 8.1), based on two up-/down-sampling layers. Each convolution layer had a kernel size of 3x3x3 with ReLu activations, and we utilized 2x2x2 Max-Pooling for downsampling. To accelerate convergence without overloading the GPU memory, we added a

Batch Normalization layer (Ioffe and Szegedy, 2015) after the features maps with the lowest resolution (5th convolution layer). Additionally, to improve generalization, we added a Dropout layer (Srivastava et al., 2014) before the last convolution. The parameters of the network were optimized with the Adadelta optimizer (Zeiler, 2012). To improve generalisation, we also trained the algorithm with online data augmentation using random translations < 50 voxels, 3D rotations of maximum 0.2 radian and flipping according to the coronal plane. The intensity of the ventricles and of the sulci were also separately randomized for data augmentation. To artificially increase the intensity of the ventricles, we used the annotations and randomly added to the ventricles intensities a maximum of 2μ, with μ the mean intensity of the FLAIR scans after percentile normalization. To artificially modify the intensity of the sulci, we randomly added between -2μ and 2μ to regions of the images with an intensity value lower than 0.25 after percentile normalization. The algorithm was implemented using the publicly available Keras 2.2.0 library (Chollet et al., 2015) with TensorFlow 1.10 as backend (Abadi et al., 2016).

The network's outputs were binarized at a threshold of 0.5. To improve the segmentation, in the ventricle binary maps, we removed small connected components with a volume smaller than a manually determined threshold of 5 voxels.

2.3 Registration quality assessment

All pairwise registrations from image to atlas were performed using ANTs SyN nonlinear diffeomorphic registration algorithm with default parameters (Avants et al., 2011). Inverse registrations were computed to allow the propagation of atlases' ventricle segmentations to image space. The quality of the registration

$T_{x,a}$ of an image x to an atlas a can be assessed by measuring the overlap between the ventricles segmented by the CNN in image space (V_{CNN}) and the ventricles of the atlas a (V_a) propagated to image space $V_{x,a} = T_{x,a}^{-1}(V_a)$. We denote this registration quality metric as $Q_{x,a} = D(V_{CNN}, V_{x,a})$, where D is the Dice similarity coefficient.

Other more conventional metrics – that measure e.g. image similarity – could be used instead to assess registration quality. We assessed this based on the cross-correlation (CC), i.e. the registration metric itself (ANTS SyN (Avants et al., 2008; Sarvaiya et al., 2009; de Groot et al., 2013)) between the registered image x and each atlas a such that $Q_{x,a} = T_{x,a}(x) \star a$, where $\star$ denotes the cross-correlation operation. Prior to the computation of the cross-correlation, images were rescaled in $[0, 1]$ using their minimum and maximum intensity values.

2.4 Multi-Atlas Registration

Each scan was registered pairwise to each atlas in $A = a_1, ..., a_5, g$, where a_i are the age-specific atlases and the g is the general atlas. For a given scan, the best atlas b was then selected based on the registration quality metric Q, so that

$$b =_{a \in A} Q_{x,a}, \tag{8.1}$$

with, for the ventricle overlap quality metric, $Q_{x,a} = D(V_{CNN}, V_{x,a,g})$, where $V_{x,a,g} = T_{x,a}^{-1} T_{a,g}^{-1}(V_g)$. If the best atlas was not the general atlas, the scan uses the intermediate registration target b and is then warped to the general atlas using the deformation field of the registration of the intermediary atlas to the general atlas (Figure 8.2).

3 Experiments

3.1 Ventricle Segmentation

The ventricle segmentation algorithm was optimized using the training/validation set, which was randomly split into 240 training scans and 60 validation scans to monitor over-fitting. The algorithm was then evaluated on the test set of 100 scans. The experiments with the MAR framework were conducted using the complete GASROS dataset excluding the 300 scans used to optimized the ventricle segmentation algorithm and 41 scans with strong motion artifacts, but excluding the 100 scans of the test set for ventricle segmentation, hence resulting in 791 scans.

We assessed the automatic segmentation of the ventricular system in the FLAIR sequences based on 11 different metrics. These metrics included the Dice similarity coefficient (Dice), Jaccard index (Jaccard), true positive rate (TPR), mutual information (MI), Cohen's kappa (KAP), intraclass correlation coefficient (ICC), volumetric similarity (VS), adjusted Rand index (ARI), probabilistic distance (PBD), detection error rate (DER) and outline error rate (OER). VS was computed as the absolute volume difference divided by the sum of both volumes. ARI is Rand index corrected for chance. Rand index measures similarity between clusters. PDB measures the distance between fuzzy segmentations. DER measures the disagreement in detecting the same regions, namely the sum of the volumes of regions detected in only one of both segmentations. OER measures the disagreement in outlining of the regions, namely the difference between union and intersection of regions detected in both segmentations. A detailed description of the metrics is given elsewhere (Taha and Hanbury, 2015; Wack et al., 2012).

PBD, DER, and OER are a measure of dissimilarity, where smaller values represent better agreement. As DER and OER are bounded metrics, we rescaled them between 0 and 1, and reported 1-DER and 1-OER. In case of PBD (not bounded), we reported 1/(1+PBD). Subsequently, all similarity metrics are bound between 0 and 1, where 1 indicates a perfect segmentation. Results are visualized as radar plots.[1]

3.2 Evaluation of the multi-atlas registration framework

We compared the proposed multi-atlas registration method to a direct registration to the general atlas and quantified the gain in registration performance by the difference $\Delta_{b,g} = Q_{x,b} - Q_{x,g}$, where Q represents the Dice coefficient of ventricle overlap. We computed Wilcoxon tests on all subjects, in order to evaluate the efficacy of the proposed MAR framework. Additionally, we investigated the effect of utilizing different registration quality assessment metrics and the dependency of age and ventricle volume on the selection of the best atlas.

3.3 Spatial maps of WMH burden

Utilizing the manual WMH segmentations from GASROS, we generated an average voxelwise map of WMH burden in template space. After using the MAR framework, we selected subjects for which registration quality was above a threshold T. Using three different thresholds T = 0, 0.6, and 0.9, we visually assessed the quality of WMH maps constructed.

[1] Github link - https://github.com/marconardin/spider-plotting

4 Results

4.1 Ventricle segmentation

The results of evaluating the automated ventricle segmentation (see Figure 8.3) show good agreement between the manual and automated ventricle segmentations, with Dice coefficients of 0.89 for the single-site GASROS dataset and 0.83 for the multi-site MRI-GENIE dataset. Results of the ventricle segmentation for the MRI-GENIE data set, stratified by site, are shown in Appendix A.

4.2 Multi-atlas registration

4.2.1 Atlas creation

Figure 8.4 shows the age-specific atlases created from the healthy controls from the ADNI dataset. Computing the mean squared intensity difference between the age-specific atlases and the general atlas revealed that atlas 75-80 was the closest to the general atlas, and atlas 80-85 was the most dissimilar.

4.2.2 Gain in registration performance

The gain in registration performance $\Delta_{b,g}$ is shown for each dataset in Figure 8.5 and Appendix E. We observed age-dependent improvements with increases of ventricle overlap by up to 0.15 Dice points. Wilcoxon tests showed that the proposed MAR method reached a significantly higher registration quality – measured as ventricle overlap – than that of the direct registration to the general atlas (Figure 8.6) in N=430 GASROS subjects (54%) and 93 MRI-GENIE subjects (65%). However, when using cross-correlation instead of ventricles overlap for intermediary atlas selection, the proposed MAR method did not reach a significantly higher registration quality than that of the direct registration to the general atlas (Figure 8.6; Appendix C and D). As expected, younger patients with lower ventricle volume were assigned to atlases of younger categories (Figure 8.7).

4.2.3 Manual versus automated ventricle segmentation

To assess the validity of using the CNN results as reference for the registration, we evaluated the difference of results for the MAR framework in each dataset when using manually versus automatically segmented ventricles and found no large difference (Figure 8.8 and Table 8.1).

4.2.4 Spatial WMH maps of WMH burden

Figure 8.9 shows that increasing the threshold of registration quality (rejecting more subjects) reduces, e.g., the erroneous extension of the WMH into the CSF compartments of the brain.

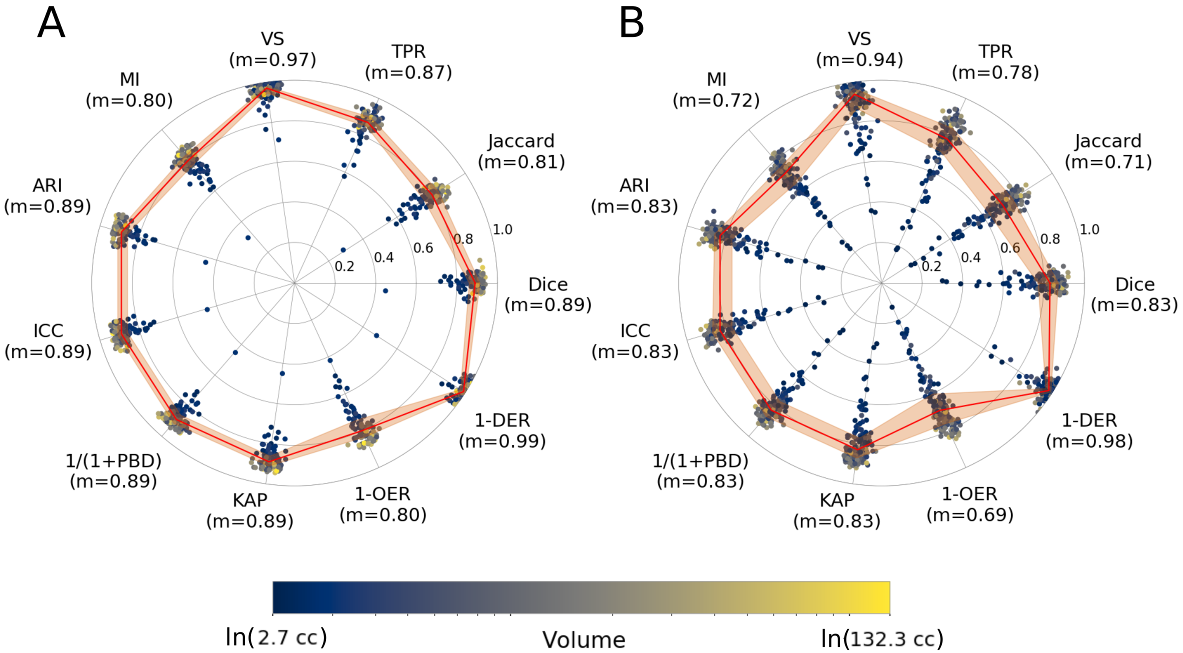


Figure 8.3: **Comparison of automated and manual ventricle segmentations in A) GASROS (N=100; left) and B) MRI-GENIE (N=142; right).** The reported metrics are Dice coefficient (Dice), Jaccard index (Jaccard), true positive rate (TPR), volumetric similarity (VS), Mutual information (MI), Adjusted Rand Index (ARI), intraclass correlation coefficient (ICC), probabilistic distance (PBD), Cohen's kappa (KAP), Detection Error Rate (DER) and Outline Error Rate (OER). The solid line is based on the median of each measure, while the ribbon represents the interquartile range.

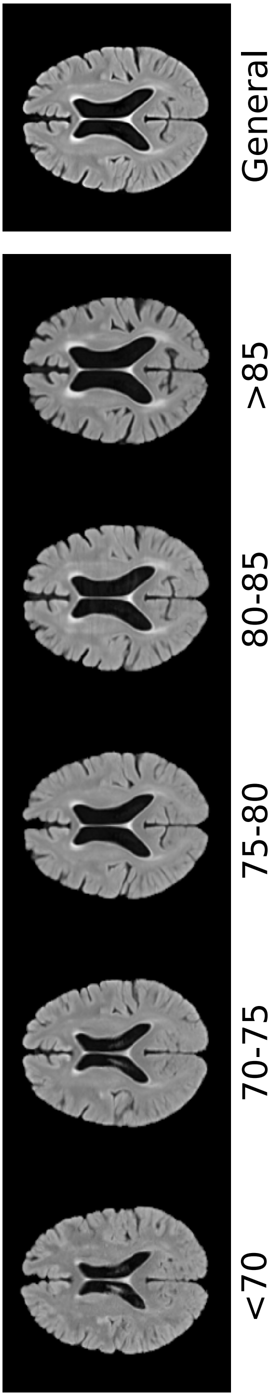


Figure 8.4: **Age-specific atlases and the general atlas registered to MNI space.**

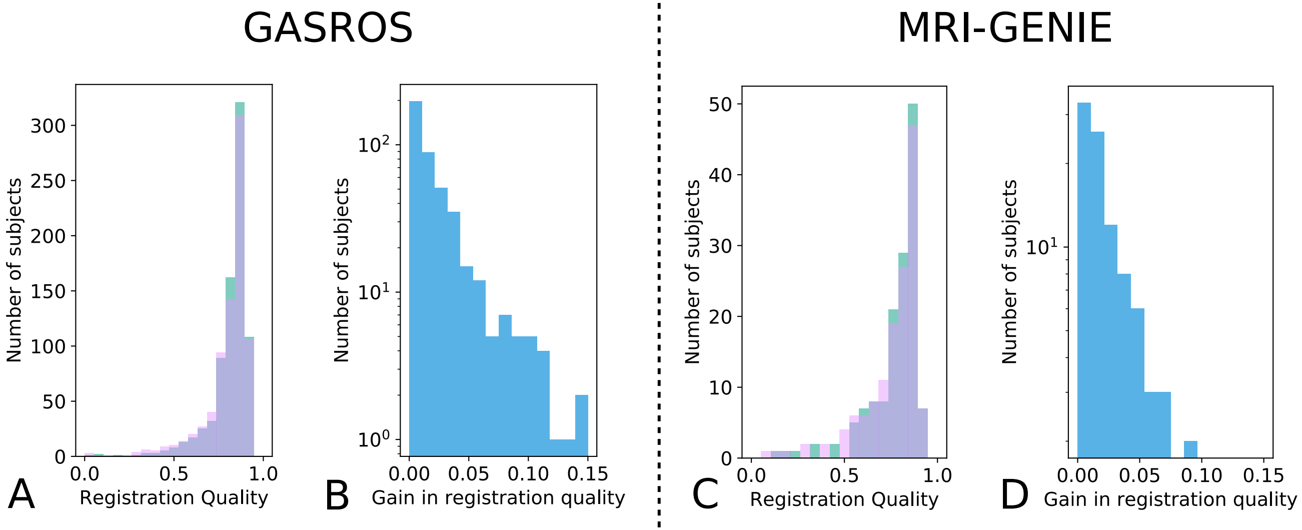


Figure 8.5: **Gain in registration performance measured as ventricle overlap by using the proposed MAR method in comparison to a direct pairwise registration to the general atlas g for each dataset (Left: GASROS; Right: MRI-GENIE).** A/C: registration quality histograms using either direct registration to the general atlas (pink) or the MAR (green; improvement of registration quality). The overlap of both methods is shown in purple. B/D: Gain in registration quality $\Delta_{b,g}$. Scatterplots are also available in Appendix F.

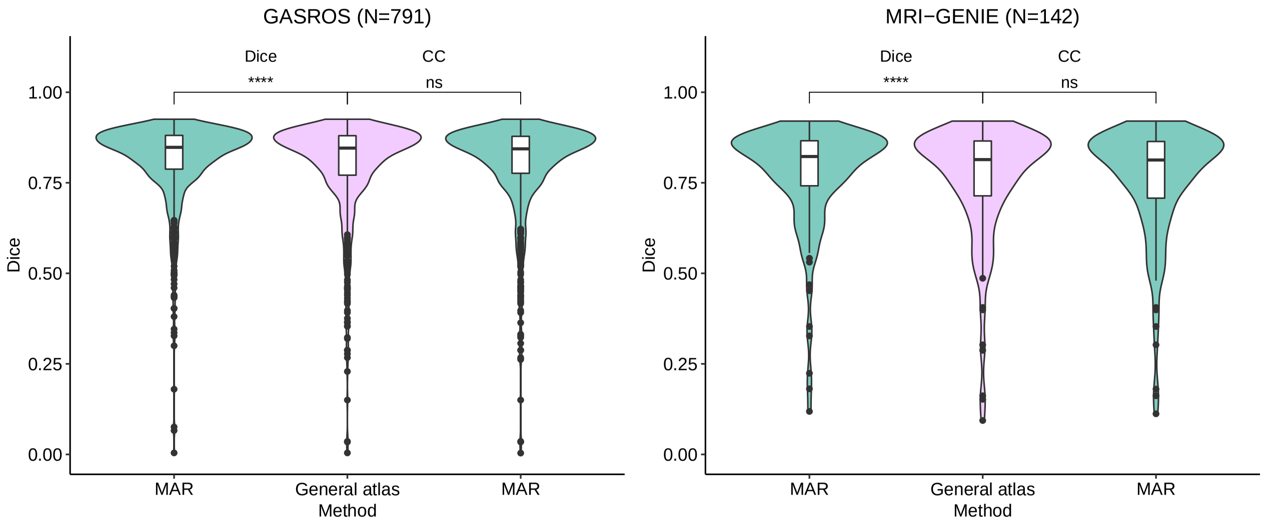


Figure 8.6: **Comparison of the proposed MAR with a direct registration to the general atlas.** Instead of the proposed selection strategy for the intermediary atlas (ventricle Dice), we also experimented using the more standard selection criterion: cross-correlation (CC), computed after the elastic registration and normalization of intensity values. **** indicates a p-value lower than 0.0001 for the Wilcoxon test, and n.s. Indicates a non significant difference.

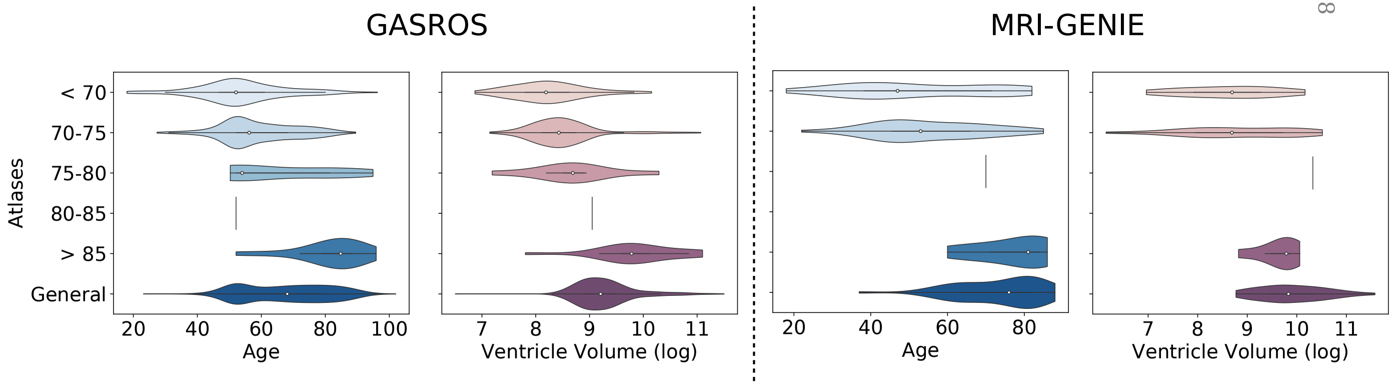


Figure 8.7: **Effect of age and ventricle volume on the selection of the atlases using ventricle overlap as registration quality metric.** Violin plots show the distribution of the subjects' age – and ventricle volume – according to the best atlas the subjects were assigned to in the MAR framework. A vertical line indicates that only n=1 subject has been assigned to the template.

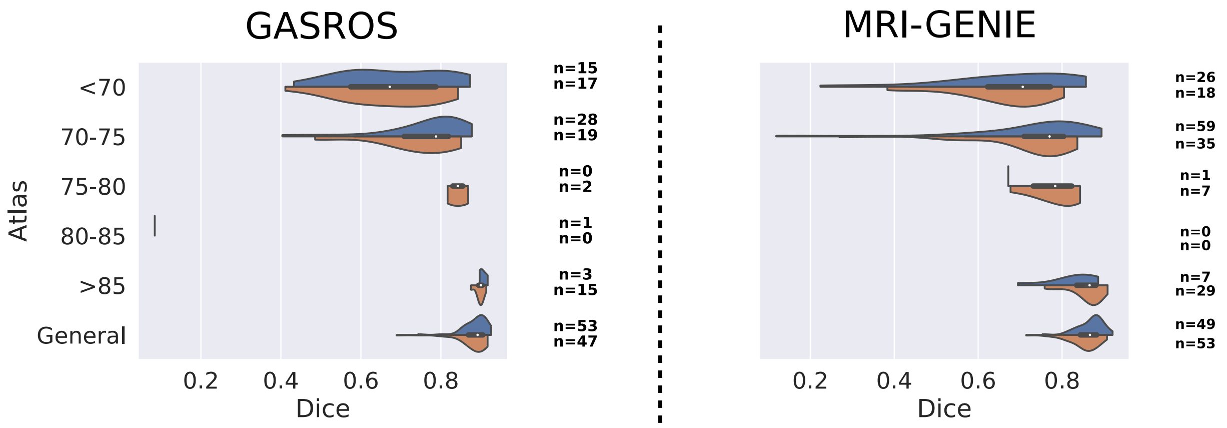


Figure 8.8: **Comparison of multi-atlas registration using automated (blue) and manual (orange) segmentation of the ventricles in subject space.** The number of scans assigned to each atlas is indicated on the right of each plot for both automated and manual ventricle segmentations.

Table 8.1: **Gain in registration performance comparing the proposed multi-atlas registration framework using either the manual or automated ventricle segmentations to compute the registration quality.** Results are displayed as mean Dice coefficient of ventricle overlap. The number of scans assigned to each age-specific atlas is indicated between brackets.

	GASROS 100 manual	GASROS 100 automated	MRI-GENIE manual	MRI-GENIE automated
Mean gain dice	0.01 (100)	0.011 (100)	0.010 (142)	0.014 (142)
Mean gain dice when improvement	0.019 (53)	0.024 (43)	0.016(90)	0.022 (93)
Under70	0.034 (17)	0.04 (15)	0.036 (18)	0.033 (26)
70-75	0.019 (19)	0.019 (28)	0.016 (35)	0.019 (59)
75-80	0.001 (2)	(0)’	0.004 (8)	0.0005 (1)
80-85	(0)’	0.001 (1)	(0)’	(0)’
Above 85	0.005 (15)	0.006 (3)	0.006 (29)	0.008 (7)

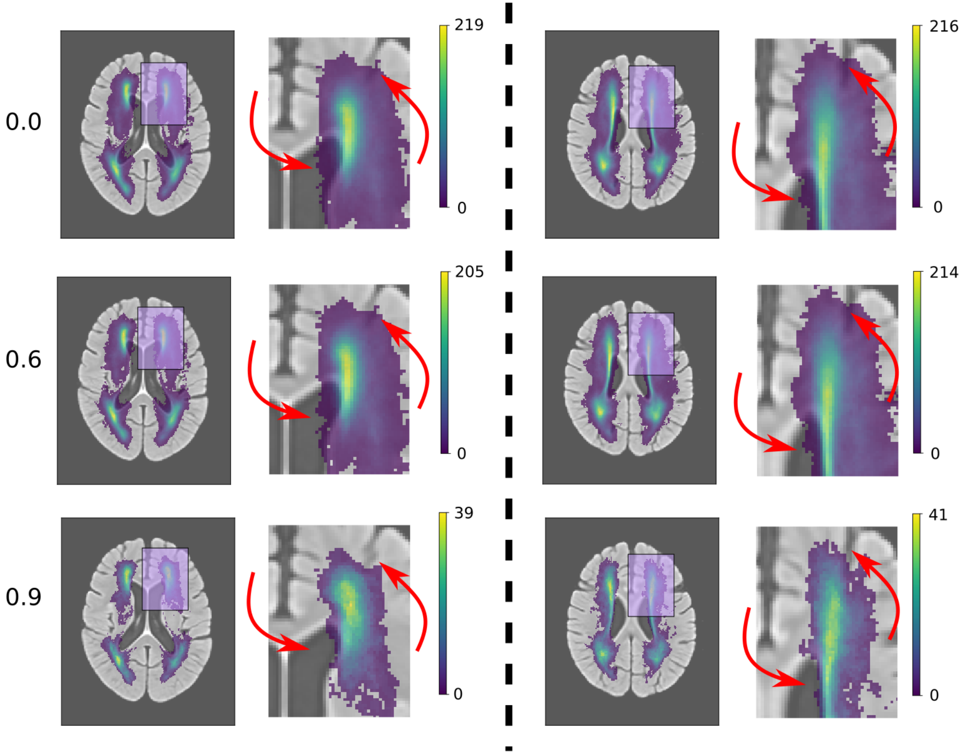


Figure 8.9: **White matter hyperintensity (WMH) burden overlayed with the general atlas.** Rows correspond to different thresholds T for the quality of the registration measure Q used to create WMH maps: from top to bottom: $Q \geq 0$ (all images = 791 images), $Q > 0.6$ (748 images), and $Q > 0.9$ (83 images). The columns correspond to two different brain slices in the axial plane. On the left of each column is the full image and on the right a zoomed in version of the region highlighted in pink. Red arrows indicate regions with a visible improvement in WMH maps.

5 Discussion

In this chapter, we demonstrated the use of a ventricle segmentation algorithm using clinical FLAIR sequences, for automated registration quality assessment, and validated the proposed quality assessment metric in a multi-atlas registration (MAR) framework.

The registration quality assessment method compared the ventricles of a subject, segmented with a machine learning algorithm, to the ventricles of the atlas, propagated to subject space. A ventricle segmentation algorithm that is robust to variations in scanners, sites and image resolutions is consequently a keypoint of its applicability. Here, we demonstrated that the proposed algorithm performed well in a multi-site scenario, while being trained with data from a single site. While, as expected, the algorithm reached a higher performance for the dataset it was optimized on (GASROS), the performance dropped by less than 6 percentage points of Dice coefficient when used on multi-site data. Importantly, the segmentation method generalized well to the other, multi-site data by designing appropriate data augmentation procedures, and without employing advanced transfer learning algorithms. Using manually or automatically segmented ventricles using the proposed deep learning algorithm, led to similar results with the MAR framework in each dataset (Figure 8.8 and Table 8.1), with a difference in mean gain in Dice coefficient of 0.001 in GASROS dataset, and 0.004 in the MRI-GENIE dataset. The largest differences were that: (1) when using the automated segmentation, more scans were assigned to atlas of age range 70-75 instead of atlas under 70 or the general atlas, and (2) when using the manual segmentation, more scans were assigned to atlas of age range above 85 instead of the general atlas.

Klein et al. (2009) showed that for multiple registration algorithms (including

ANTS) the registration error of the ventricles correlates with registration errors in other regions. Manually annotated landmarks describing brain structures in the atlas could help to monitor more globally the registration quality than using the ventricles alone. However, automatically detecting such landmarks in clinical data remains a difficult task, and might lead to more erroneous cases, in contrast to segmenting a large, reliable structure, such as the ventricles. However, our framework can be extended to use multiple segmentations such as grey and white matter segmentations in the future.

We used the automated registration quality assessment method to design a multi-atlas registration (MAR) framework for improving registration quality. Instead of being directly and only registered to a general atlas, scans were first registered to atlases corresponding to several age categories. The best of these atlases was then chosen using the registration quality assessment method, and registration to the selected atlas was used as an intermediary registration step. In our dataset, using the MAR framework with ventricle overlap significantly improved the registration quality. Patients were often assigned to an intermediate atlas that was closer to their chronological age. However, we observed a shift, where, on average, subjects were matched to age-specific atlases of an older age category than their chronological age. This most probably resulted from the specific cohort in our analyses: all subjects had a prior acute ischemic event, which may reflect brains with increased biological age. This is further supported by studies which suggested that biological age, in contrast to chronological age, can play a key role in susceptibility to disease (Wang et al., 2019).This suggests also that selecting the age-specific atlas using the patient's chronological age would be a suboptimal strategy.

We further observed a positive correlation between ventricle volumes and the age category of the atlas the scans were assigned to. This relationship was expected, considering that age is positively correlated with ventricles volume in the general population (Walhovd et al., 2011), which can also be seen on the age-specific atlases themselves (Figure 8.4). The age-specific atlases also showed expected behavior of increased WMH volume and cortical atrophy with increasing age (Earnest et al., 1979). In all experiments, only a few scans were assigned to the atlases of age category 75-80 and 80-85. Computing the mean squared intensity difference between the age-specific atlases and the general atlas revealed that atlas 75-80 was the closest to the general atlas, and atlas 80-85 was the most dissimilar. Consequently, scans most similar to atlas 75-80 were more likely to be assigned to the general atlas instead.

Other researchers have successfully used age-specific atlases (Sanchez et al., 2012; Fillmore et al., 2015; Liang et al., 2015; Schirmer et al., 2019b,a). Liang et al. (2015) proposed to construct age-specific templates, and observed an improvement for hippocampi segmentation. And Fillmore et al. (2015) observed an improvement in segmentation of white matter, gray matter and cerebrospinal fluid using an age-appropriate brain template. It is often impossible to find a single atlas, which works best for studies across the entire lifespan. Instead, using multiple age-specific atlases allows a more accurate description of the lifespan and can improve registration quality. In this chapter, we utilized five age groups, which already demonstrated improvement in overall registration quality. By using even more atlases, i.e. additional or smaller spaced age groups, could lead to further improvements. Intermediary registration to a template has also been used to accelerate multi-atlas segmentation (Dewey et al., 2017), or to improve registration from one image modality to another. For example, Parthasarathy et al. (2011) used a full-volume ultrasound image as intermediary image for the registration of live-3D ultrasound to MRI. Later, Roy et al. (2014) used an synthesized CT image as intermediary image for the registration from

MRI to CT. Groupwise registration (Joshi et al., 2004; Fletcher et al., 2009) could be another strategy to register all scans of a dataset to the same space. No template image needs to be selected in advance, and transformation fields are estimated simultaneously for all scans. One of the main disadvantages of groupwise registration is that the initial common space is estimated as the mean of all scans in the dataset. This mean image can be fuzzy and not provide enough guidance for the iterative optimization process (Wu et al., 2010). Aligning the images to the MNI template instead of only aligning them to the general atlas created from ADNI healthy controls might be of interest, for example, to compare with other datasets already registered to the MNI template. For this purpose, a registration step to MNI template could be added as a last step of the MAR framework, after the registration to the general atlas. The general atlas would then need to be registered to the MNI template. This approach would guarantee a smoother and more controlled transformation than registering the age-specific atlases directly to the MNI template, and would provide a more precise monitoring of potential registration errors: ventricle overlap could be computed both when registering to the general atlas and when subsequently registering the MNI template, and errors in the pipeline could be more easily identified.

The proposed MAR framework using ventricle overlap could be categorized as a feature-based registration method. Segmentations in feature-based registration methods have already been used as initialization (Vemuri et al., 2003), or have been optimized jointly with an intensity similarity metric for registration (Yezzi et al., 2003; Pohl et al., 2006; Chen et al., 2010). More recently, Balakrishnan et al. (2019) proposed to use a deep learning registration approach where segmentations of anatomical structures can be used as auxiliary data during the optimization. This would allow to include the ventricle segmentation in the optimization of the registration, instead of the proposed MAR framework. However, to date, utilizing auxiliary data for registration has not been tested in clinical scans, which are known to be substantially more challenging to segment and register. With the presented ventricle segmentation, and the segmentation of other structures and the entire brain, the extension of such approaches to clinical scans becomes more feasible and is of key interest for future studies. In Appendix G, we compared a registration method in which ventricle segmentation was added as auxiliary objective with equal weight during registration to the proposed MAR and, as expected, obtained higher ventricles overlap. However, by utilizing the ventricle segmentation for registration, we cannot utilize it anymore for objectively assessing registration quality. Additionally, Balakrishnan et al. (2019) have done similar experiments with brain registration and observed that when using the overlap of a single structure as auxiliary objective, the overlap of the other brain structures stayed either the same or even decreased when using larger weight for the auxiliary objective. In addition to, or instead of, using the ventricles to assess registration quality, it might also be interesting to inspect subcortical structures on T1-weighted MRI sequence, and attempt to exploit features based on the intensity difference between white and gray matter in, for example, the basal ganglia.

In our application, we demonstrated that it becomes feasible to automatically select only scans with high registration quality, leading to more globally accurate – but also possibly more noisy as computed from a smaller set – maps of WMH burden. Using automated assessment of registration quality to compute more accurate spatial patterns of disease could further help to relate spatial information to global phenotypes such as stroke severity or hypertension. For instance research has been done on how WMH distribution differs between patients with lobar intracerebral hemorrhage and healthy elderly (Zhu et al., 2012), or on differences between deep and periventricular WMH in relation to stroke (Buyck et al., 2009). However, discarding scans with a lower registration quality might also introduce a bias if the quality of the registration is related to one of the studied determinants or outcomes. Alternatively, a more rigorous quality control procedure might also be triggered for those scans.

There are limitations to this study. Our proposed method requires reliable automated segmentation of a key structure in the image, which can subsequently serves as a reference. This can be challenging with smaller structures in the image. Here, we focused on the ventricular system, which represents a structure that is relatively easy to segment consistently across subjects. While such a discriminative structure might not appear in every body part or with every imaging modality, further methodological advances in image segmentation will improve the generalizability of the proposed framework. Examples of structures that are suited to the proposed method could be large blood vessels in magnetic resonance angiography, or fetus in fetal MRI. The premise of our registration quality assessment lies in ventricles being visible on the clinical images. In particular in stroke cases, mass effects can alter the appearance of the ventricles, sometimes rendering the lateral ventricles invisible in the image. Additionally, the posterior horns of the ventricles may be masked due to the low resolution of the acquired clinical scans. If ventricles cannot be identified on the image, our proposed metrics may indicate insufficient registration quality. However, this assessment can be used to flag this subset of the registered scans as potentially erroneous, which can then be manually assessed by an expert rater rather than being completely rejected from the analysis. If the registration is erroneous, the third and fourth ventricles in particular are less likely to overlap with the atlas, reducing the probability of high dice for incorrect registration. We observed some outliers with low ventricle overlap between the automated and manual ventricle segmentation. The majority of these outliers – for instance 2 out of 100 scans in GASROS dataset – were scans with substantial motion artifacts, where the segmentation of ventricles was challenging even for human raters. Such scans are usually excluded from most neuroimaging pipelines. In addition, in some sites of the MRI-GENIE dataset, sulci were sometimes misclassified as ventricles. Another limitation is that the proposed MAR framework also multiplies the computation time by the number of atlases used: in our case, the registration is six times longer. However, each registration can be run in parallel, and in cases where immediate results are not necessary, this approach can help improve registration quality. Additionally, with the recent development of deep-learning based registration frameworks (Balakrishnan et al., 2019), time concerns may become negligible.

Instead of using segmentation to perform automated quality control of registration, Robinson et al. (2019) proposed to use registration to perform quality control of segmentation. This assumes that the registration is more robust to the variations present in the dataset than the segmentation. Using segmentation to perform automated quality control of registration assumes the opposite. Whether segmentation or registration can be considered more robust depends on the region of interest, imaging modality, and image resolution. The full ventricular system in the brain has a complex shape with substantial inter-subject variability due to, for example, brain atrophy and/or pathological processes. This makes the registration difficult when the shape of subject's ventricles deviate from the expected ventricle shape. Conversely, image intensity on FLAIR-weighted MRI is a substantially more discriminative feature than shape. The high contrast between intensities inside and outside the ventricles is present in all subjects, scanner and FLAIR protocols. Segmentation of the ventricular system can therefore be expected to be more robust than registration.

In contrast the structures composing the heart, as seen on MRI, have a simple ovoid shape with similar image intensities, making registration approaches more reliable as a reference. The other key aspect is that registration of clinical scans to templates is difficult and remains an open research question. Registration could potentially be more reliable if we had a more homogeneous, high-resolution dataset such a the UK-biobank, as Robinson et al. (2019) used in their analyses.

Strengths of our work include segmentation of the four ventricles in clinical scans evaluated in multi-center data and more than 1000 scans. We introduced a multi-atlas registration framework based on this segmentation algorithm, and employed it to compute more accurate maps of WMH burden.

No single registration tool, or set of registration parameters, will perform best on all types of image qualities or sequences. By implementing an automated registration assessment step in large scale image analyses, it becomes feasible to test multiple registration pipelines and select the registration with the best performance. This can increase the number successful registrations, and potentially increase the sample size of a study without the need for time intensive manual quality assessment.

In this work, we demonstrated the utility of an automated tool for assessing image registration quality in clinical scans. Importantly, in addition to extracting an additional phenotype from clinical scans – namely the ventricle volume – this image quality assessment step can be implemented in large-scale, automated processing pipelines of clinical MRI data, increasing the utility of such pipelines and offering improved quality of subsequent analysis, ultimately assisting in the translation of such pipelines to the clinic.

Acknowledgements

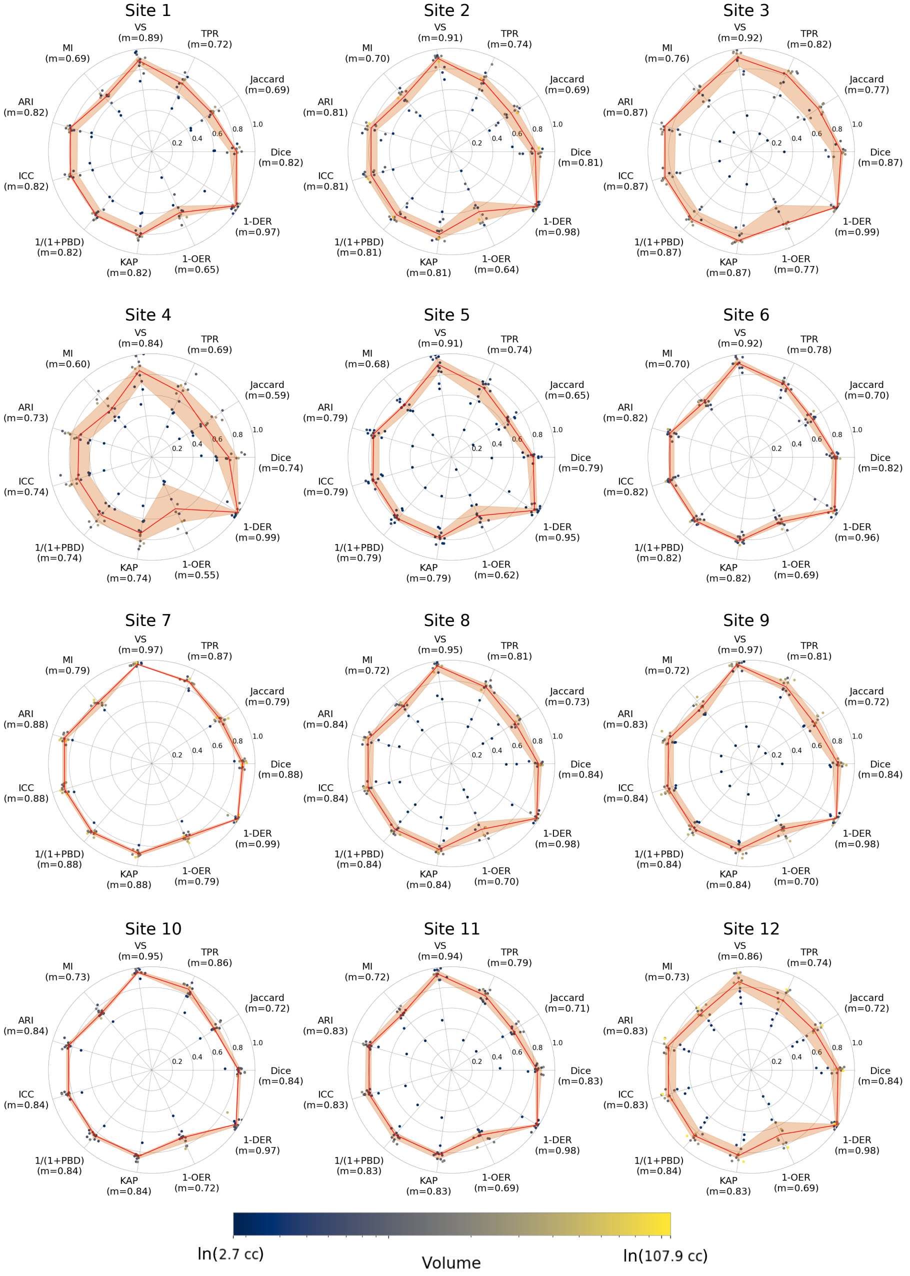
This project has been in part supported by the Foundation “De Drie Lichten” in the Netherlands (F.D.); The Netherlands Organisation for Health Research and Development (ZonMw) Project 104003005 (M. de B.); the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 753896 (M.D.S.); the NIH-NINDS K23NS064052, R01NS082285, and R01NS086905 (N.S.R.); American Heart Association/Bugher Foundation Centers for Stroke Prevention Research and Deane Institute for Integrative Study of Atrial Fibrillation and Stroke (N.S.R.). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Part of the Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda

Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Chapter 8: Appendix A

Ventricle segmentation results for the 12 sites of MRI-GENIE. The reported metrics are Dice coefficient (Dice), Jaccard index (Jaccard), true positive rate (TPR), volumetric similarity (VS), Mutual information (MI), Adjusted Rand Index (ARI), intraclass correlation coefficient (ICC), probabilistic distance (PBD), Cohen's kappa (KAP), Detection Error Rate (DER) and Outline Error Rate (OER).


Site 1
VS (m=0.89)
TPR (m=0.72)
Jaccard (m=0.69)
Dice (m=0.82)
1-DER (m=0.97)
1-OER (m=0.65)
KAP (m=0.82)
1/(1+PBD) (m=0.82)
ICC (m=0.82)
ARI (m=0.82)
MI (m=0.69)
0.2 0.4 0.6 0.8 1.0
Site 2
VS (m=0.91)
TPR (m=0.74)
Jaccard (m=0.69)
Dice (m=0.81)
1-DER (m=0.98)
1-OER (m=0.64)
KAP (m=0.81)
1/(1+PBD) (m=0.81)
ICC (m=0.81)
ARI (m=0.81)
MI (m=0.70)
0.2 0.4 0.6 0.8 1.0
Site 3
VS (m=0.92)
TPR (m=0.82)
Jaccard (m=0.77)
Dice (m=0.87)
1-DER (m=0.99)
1-OER (m=0.77)
KAP (m=0.87)
1/(1+PBD) (m=0.87)
ICC (m=0.87)
ARI (m=0.87)
MI (m=0.76)
0.2 0.4 0.6 0.8 1.0
Site 4
VS (m=0.84)
TPR (m=0.69)
Jaccard (m=0.59)
Dice (m=0.74)
1-DER (m=0.99)
1-OER (m=0.55)
KAP (m=0.74)
1/(1+PBD) (m=0.74)
ICC (m=0.74)
ARI (m=0.73)
MI (m=0.60)
0.2 0.4 0.6 0.8 1.0
Site 5
VS (m=0.91)
TPR (m=0.74)
Jaccard (m=0.65)
Dice (m=0.79)
1-DER (m=0.95)
1-OER (m=0.62)
KAP (m=0.79)
1/(1+PBD) (m=0.79)
ICC (m=0.79)
ARI (m=0.79)
MI (m=0.68)
0.2 0.4 0.6 0.8 1.0
Site 6
VS (m=0.92)
TPR (m=0.78)
Jaccard (m=0.70)
Dice (m=0.82)
1-DER (m=0.96)
1-OER (m=0.69)
KAP (m=0.82)
1/(1+PBD) (m=0.82)
ICC (m=0.82)
ARI (m=0.82)
MI (m=0.70)
0.2 0.4 0.6 0.8 1.0
Site 7
VS (m=0.97)
TPR (m=0.87)
Jaccard (m=0.79)
Dice (m=0.88)
1-DER (m=0.99)
1-OER (m=0.79)
KAP (m=0.88)
1/(1+PBD) (m=0.88)
ICC (m=0.88)
ARI (m=0.88)
MI (m=0.79)
0.2 0.4 0.6 0.8 1.0
Site 8
VS (m=0.95)
TPR (m=0.81)
Jaccard (m=0.73)
Dice (m=0.84)
1-DER (m=0.98)
1-OER (m=0.70)
KAP (m=0.84)
1/(1+PBD) (m=0.84)
ICC (m=0.84)
ARI (m=0.84)
MI (m=0.72)
0.2 0.4 0.6 0.8 1.0
Site 9
VS (m=0.97)
TPR (m=0.81)
Jaccard (m=0.72)
Dice (m=0.84)
1-DER (m=0.98)
1-OER (m=0.70)
KAP (m=0.84)
1/(1+PBD) (m=0.84)
ICC (m=0.84)
ARI (m=0.83)
MI (m=0.72)
0.2 0.4 0.6 0.8 1.0
Site 10
VS (m=0.95)
TPR (m=0.86)
Jaccard (m=0.72)
Dice (m=0.84)
1-DER (m=0.97)
1-OER (m=0.72)
KAP (m=0.84)
1/(1+PBD) (m=0.84)
ICC (m=0.84)
ARI (m=0.84)
MI (m=0.73)
0.2 0.4 0.6 0.8 1.0
Site 11
VS (m=0.94)
TPR (m=0.79)
Jaccard (m=0.71)
Dice (m=0.83)
1-DER (m=0.98)
1-OER (m=0.69)
KAP (m=0.83)
1/(1+PBD) (m=0.83)
ICC (m=0.83)
ARI (m=0.83)
MI (m=0.72)
0.2 0.4 0.6 0.8 1.0
Site 12
VS (m=0.86)
TPR (m=0.74)
Jaccard (m=0.72)
Dice (m=0.84)
1-DER (m=0.98)
1-OER (m=0.69)
KAP (m=0.83)
1/(1+PBD) (m=0.84)
ICC (m=0.83)
ARI (m=0.83)
MI (m=0.73)
0.2 0.4 0.6 0.8 1.0
ln(2.7 cc)
Volume
ln(107.9 cc)

Chapter 8: Appendix B

List of ADNI 3 IDs used for the computation of the age-specific atlases.

< 70 :

23_S_4448, 128_S_4607, 141_S_6008, 014_S_6076, 007_S_6120, 029_S_4384

70 – 75 :

068_S_4340, 031_S_4021, 094_S_4649, 003_S_4644, 019_S_4367, 070_S_4856, 009_S_4388, 100_S_4469,
135_S_4446, 068_S_4424, 116_S_4453, 016_S_4952, 137_S_4520, 127_S_4604, 037_S_4028, 129_S_4369,
014_S_4401, 135_S_6104, 029_S_4585, 037_S_4410, 024_S_4084, 135_S_4598

75 – 80 :

127_S_4148, 011_S_4105, 002_S_4225, 099_S_6038, 016_S_4951, 099_S_4076, 006_S_4357, 014_S_4576,
037_S_4308, 002_S_6007, 023_S_4164, 032_S_4277, 021_S_4335, 018_S_4400, 041_S_4427, 003_S_4288,
129_S_4422, 098_S_4275, 098_S_4506, 116_S_4483, 007_S_4488, 021_S_4276, 006_S_4485, 082_S_4428,
098_S_4003, 941_S_4292, 013_S_4580, 035_S_4464, 007_S_4637, 141_S_6061, 041_S_4200

80 – 85 :

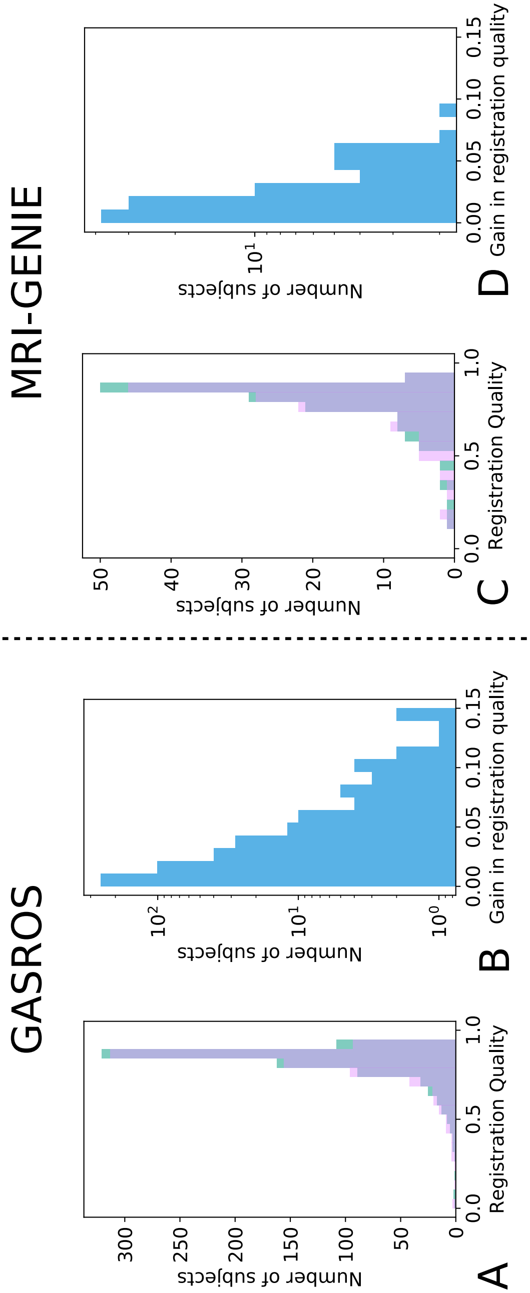
041_S_4037, 011_S_4278, 127_S_0259, 068_S_0473, 018_S_4313, 019_S_4835, 002_S_1280, 032_S_0677,
006_S_0498, 067_S_0056, 007_S_4387, 070_S_5040, 068_S_0210, 141_S_0767, 007_S_1222, 123_S_0106
032_S_4429, 005_S_0602, 130_S_0969, 082_S_4224, 009_S_0751, 033_S_0734, 002_S_4213, 068_S_0127
002_S_1261, 027_S_0120, 137_S_4482, 067_S_0059, 006_S_0731, 033_S_1098, 941_S_4100, 123_S_0072
007_S_4620, 032_S_1169, 128_S_0272, 129_S_4396, 018_S_4399, 941_S_4376, 011_S_0021

> 85 :

100_S_1286, 037_S_0303, 033_S_4177, 941_S_1195, 114_S_0416, 023_S_0031, 130_S_4343, 037_S_4071,
036_S_4491, 036_S_4389, 021_S_0337, 116_S_0382, 005_S_0610, 035_S_0156, 137_S_4466, 037_S_0454
123_S_0298, 099_S_4086, 033_S_1016, 941_S_4365, 033_S_4176, 126_S_0605, 002_S_0413, 126_S_0680
035_S_0555, 116_S_4855, 098_S_0896, 116_S_4043, 033_S_4179, 100_S_0069, 023_S_1190, 021_S_4254

Chapter 8: Appendix C

Gain in registration performance by using the proposed multi-atlas registration method with ventricles overlap instead of the more standard cross-correlation for atlas selection. Left: GASROS. Right: MRI-GENIE. The registration quality with the proposed multi-atlas registration method $Q_{x,b}$ is in green; the registration quality with the proposed multi-atlas registration method using cross-correlation instead ventricle Dice to select the best atlas $Q_{x,bcc}$ is in pink; the overlay of both is purple. $\Delta_{b,bcc} = Q_{x,b} - Q_{x,bcc}$, the gain in registration quality by using the proposed multi-atlas registration method with ventricles overlap instead of cross-correlation for the selection of the intermediary atlas is in blue.


GASROS
MRI-GENIE
A
Number of subjects
300
250
200
150
100
50
0
0.0
0.5
1.0
Registration Quality
B
Number of subjects
10^2
10^1
10^0
0.00
0.05
0.10
0.15
Gain in registration quality
C
Number of subjects
50
40
30
20
10
0
0.0
0.5
1.0
Registration Quality
D
Number of subjects
10^1
0.00
0.05
0.10
0.15
Gain in registration quality

Chapter 8: Appendix D

Gain in registration performance $\Delta_{b,bcc} = Q_{x,b} - Q_{x,bcc}$. Sample size is indicated in brackets.

	GASROS	MRI-GENIE
Mean gain dice	0.011 (791)	0.014 (142)
Mean gain dice when improvement	0.018 (468)	0.021 (98)

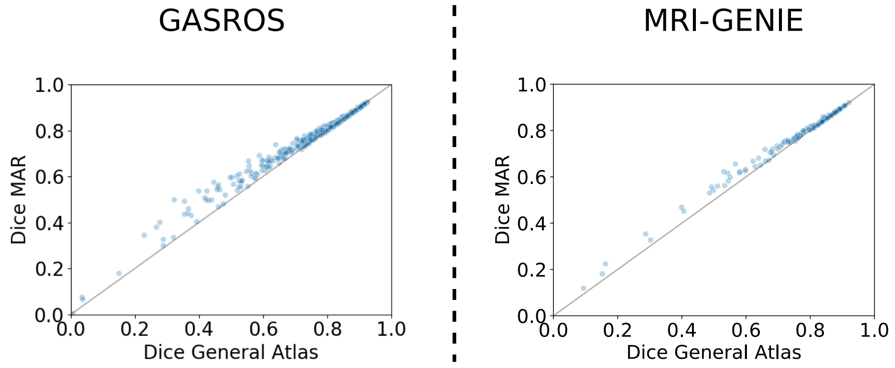
Chapter 8: Appendix E

Gain in registration performance $\Delta_{b,g}$. Sample size is indicated in brackets.

	GASROS	MRI-GENIE
Mean gain dice	0.012 (791)	0.014 (142)
Mean gain dice when improvement	0.022 (430)	0.022 (93)
Under70	0.038 (128)	0.033 (26)
70-75	0.014 (275)	0.019 (59)
75-80	0.006 (9)	0.0005 (1)
80-85	0.14 (1)	’(0)’
Above 85	0.019 (17)	0.008 (7)

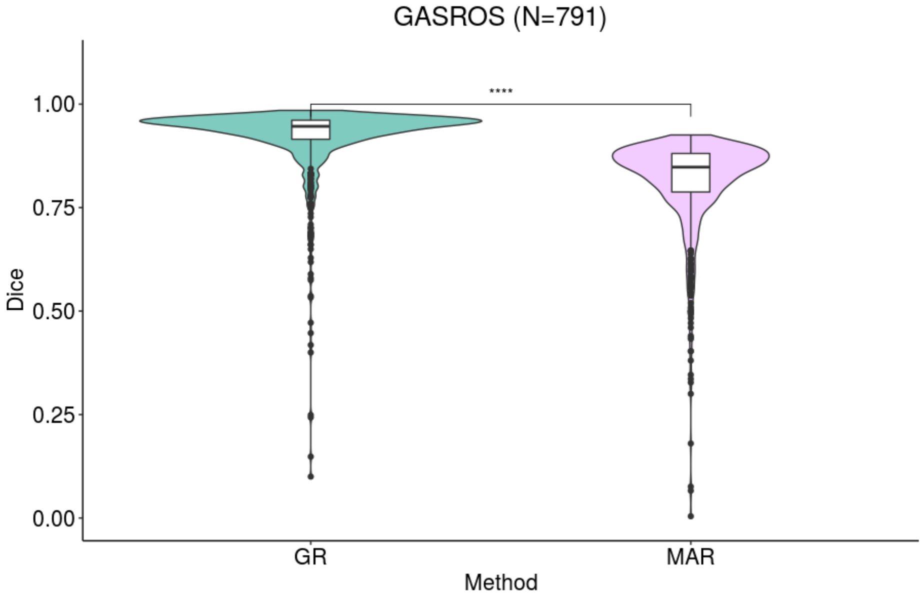
Chapter 8: Appendix F

Dice score of ventricle overlap with direct registration to the general atlas (x-axis) vers registration with the proposed multi-atlas framework (y-axis) in GASROS and MRI-GENIE datasets.



Chapter 8: Appendix G

Comparison of registration to the general atlas in which ventricle segmentation was added as auxiliary objective with equal weight during registration (GR – guided registration – left) to the proposed MAR (right). The value on the y-axis is the overlap between ventricles of the general atlas propagated to subject space and the ventricles segmented in subject space. **** indicates a p-value lower than 0.0001 for the Wilcoxon test.



Part F

General Discussion

Part F

In this part, I summarize and discuss the main findings of this thesis, discuss methodological considerations and limitations, and multiple directions for future research, illustrated with preliminary results.

1 Main findings and position in the field

Throughout this thesis, I studied the methodology of neural networks and developed automated methods for MR brain image analysis, some of which could also be deployed in clinical practise. While both methodological and medical applications are combined in each chapter, Parts B and C focus more on he methodology, and Parts D and E focus more on the medical application. This section is organised in a similar manner, first discussing weakly supervised neural networks optimized with image-level labels, and second, based on those findings, the development of an automated method for the quantification of a specific type of structural brain changes visible in MRI: enlarged perivascular spaces (PVS).

1.1 Methodological findings

In Chapter 1, I showed that neural networks with an architecture similar to that of the VGG classification network (Simonyan and Zisserman, 2015b) can be modified to perform counting in 3D volumetric data. I removed the non-linear activation function after the last fully connected layer, and optimized the mean squared error loss function instead of a classification loss function such as the binary cross entropy. I validated this method for counting the number of PVS in the basal ganglia from 3D MRI data. In computer vision, researchers usually use bounding boxes, dot annotations or scribbles to optimize neural networks for counting tasks (Lempitsky and Zisserman, 2010; Walach and Wolf, 2016; Ren and Zemel, 2017a). Seguí et al. (2015) proposed an approach similar to ours, but, contrary to our work (Chapter 1 and 3), only evaluated their method on 2D datasets and did not verify quantitatively that the network learnt to recognize the target objects. They evaluated their approach for counting handwritten digits and pedestrians in 2D datasets, but did not experiment with 3D medical data. In their article, no information was given about the loss function used, and it seems that the authors did not use a global pooling layer after the convolutional layers, contrary to our methods in Chapter 2, 5, 6 and 7. Not using global pooling to transition from image features to vector features with fully connected layers requires much GPU memory, which is a bottleneck for 3D neural networks. Using global pooling also forces position invariance and can help to improve the generalisability of the network. Secondly, I showed that not only the network predicted the count accurately, but also that it focused on the target objects in the image, in our case PVS. For this empirical demonstration, I highlighted the focus of the network in the input images by computing *attention maps* using a visualization technique proposed by Simonyan et al. (2014). The principle of this method is to compute the derivative of the output, the predicted count, with respect to the network's input, the MRI image. Regions in the image are highlighted proportionally to their contribution to the network's prediction. In another experiment, I masked some of the PVS before inputting the image to the network and observed a decrease in the predicted PVS count, which confirmed that the network was indeed focusing on PVS to predict the count.

One of the main limitations of Chapter 1 is that the network had to be optimized using training datasets of several hundred of images to reach a performance similar to the interrater agreement. In Chapter 2, I proposed a method to use available training data more efficiently. In this chapter I trained neural networks with only 25 images with each a single image-level label (which had to represent a countable quantity). The principle of the method was to combine real training samples to create many more new virtual training samples. To combine those training samples, the network processed simultaneously multiple samples, considering them as a new larger sample, and was optimized to predict the sum of their labels. Consequently, at the expense of having weaker labels, the network could use more samples for its optimization, and, most importantly, potential errors or noise in the labels were averaged. I validated this method for the quantification of PVS burden and white matter hyperintensity (WMH) volume, and showed that using only 25 training samples was sufficient to reach a performance similar to interrater agreement for the quantification of PVS in the basal ganglia. For PVS quantification, this method still needs to be evaluated in the other brain regions. For WMH quantification, we noticed that using this method could stabilize the optimization. While the proposed method consistently reached high performance, the performance of the baseline method showed more fluctuations and was worse on average. The baseline method could sometimes reach a performance similar to that of the proposed method depending on the random initialization of the network's parameters and on the randomness of the on-the-fly data augmentation.

The visualization technique by Simonyan et al. (2014) used in Chapter 1 to highlight the focus of the network in its input image has been criticized by multiple researchers. Springenberg et al. (2015) have noticed that backpropagating the gradient through the ReLU activation could create an interference pattern in the attention maps, which resulted in false positive detections. Springenberg et al. (2015) proposed an improvement of the method by changing the backpropagation through the rectified linear unit activations. Adebayo et al. (2018) showed that both of these gradient methods (Simonyan et al., 2014; Springenberg et al., 2015) could be biased to focus on the more salient objects in the image, independently of the importance of those objects in the computation of the prediction. I decided to investigate multiple types of visualization methods, including class activation map methods (Zhou et al., 2016), which compute attention maps using the features maps of the network instead of backpropagating the gradient. Adebayo et al. (2018) showed that class activation map methods were less sensitive to that bias. Class activation map methods were originally proposed to be used in combination with a global pooling layer placed after the last feature maps of the network. Attention maps could then be computed as a linear combination of those feature maps. The major problem with this method is that, as the attention maps are usually computed using the last downsampled feature maps of the network, they have a resolution that can be several times lower than that of the original input image. Detecting small objects becomes therefore difficult. In Chapter 3, I proposed to use such class activation map methods in combination with segmentation networks. I used variants of U-Net (Ronneberger et al., 2015), a now widely used network architecture for medical image segmentation, that I adapted for 3D processing. I called this method GP-Unet, where GP refers to global pooling. In Chapter 3, I compared GP-Unet with the visualization methods mentioned above and several of their variants. I performed experiments in two datasets: a dataset of images of handwritten digits (MNIST) (LeCun et al., 1998) and a brain dataset for the detection of PVS. In the MNIST experiments, GP-Unet outperformed the other weakly supervised detection methods. The experiments in the brain did not exhibit such a strong trend. GP-Unet together with the methods of Springenberg et al. (2015) and Selvaraju et al. (2017) were significantly better than other methods but the ordering of their performance depended on the brain region. We confirmed the results of Adebayo et al. (2018) that gradient methods, such as the method of Springenberg et al. (2015), got worse results than class activation maps methods in datasets where saliency alone is not enough to discriminate the target object from other salient objects.

In Chapter 3, I also empirically showed that networks optimized with a global regression objective, such as the count of the target object, reach higher detection performance than networks optimized with global (binary) classification objectives such as the presence or absence of the target object in the image. For that purpose, I used a dataset of handwritten MNIST digits, where the labels indicated either the number of occurrences of a target digit in the image, or the presence of at least one occurrence of the target digit. This finding is intuitive as networks optimized with count labels use more information for their optimization. In computer vision, many researchers use classification networks to perform detection tasks (Girshick, 2015). The need for either classification or regression networks may also depend on the average number of target objects in an image. If the target object appears many times within a single image, a regression approach will probably work best. If there are only few or single occurrences of the target object, a classification network might be more appropriate.

To quantitatively evaluate the weakly supervised method proposed in Chapter 3 in the brain MRI dataset, we had to acquire annotations of the target object. For comparison, in Chapter 4, we developed a neural network method that used these annotations for its optimization. The principle of this method was to first create a distance map using the dots annotations, and optimize the network voxel-wise to regress the values of the distance map. We compared networks optimized with different types of distance maps including Euclidean distance, voxel intensity distance, and geodesic distance which combines Euclidean and voxel intensity distance (Toivanen, 1996). We evaluated the method for the detection of enlarged perivascular space in the centrum semiovale and found that optimizing networks with the intensity distance maps provided the best detection performance. Other researchers have also used such approaches, but most of them focus on Euclidean distance (Sudre et al., 2018; Meyer et al., 2018b).

Optimizing neural networks with global image-level objectives can also be used for biomarker discovery. To detect spatial patterns that are discriminative, registering all images to a common template may be necessary or at least ease the optimization. Optimizing network with image-level labels to find spatial patterns could have applications in lesion-symptom mapping research, such as the work of Biesbroek et al. (2013) who studied white matter hyperintensities and lacunes distribution in relation to cognition. In order to increase the amount of data to study the relation between MR brain patterns and disease, it is useful to learn from clinical datasets. In Chapter 8, I therefore developed a registration method for clinical MRI scans. The method was based on a network that I optimized to robustly segments ventricles in clinical scans. I used the predictions of this model to assess registration quality in clinical scans, and with this registration quality metric, I developed a multi-atlas registration algorithm that uses age-specific atlases to improve the registration quality to a shared target space. This method assumes that the ventricles can be more reliably segmented in clinical scans and that ventricles segmentation is representative of the overall registration quality. Klein et al. (2009) showed that for multiple registration algorithms the registration error of the ventricles was correlated with registration errors in other regions. To provide a more accurate assessment of registration quality, other brain structures that can be robustly segmented in clinical scans could be included in the assessment.

Apart from the weak label training that I performed in this thesis, neural networks can also be optimized without labels. This can be used, for example, for dimensionality reduction. The lower dimensional embeddings of the original images can then be easily manipulated using linear algebra. For example, in Chapter 7, I developed a variational auto-encoder (Kingma and Welling, 2014) that mimics neurodegeneration in the brain and can generate brain regions corresponding to a target pseudo-probability of having Alzheimer's disease. Labels were not used for the optimization of the network, but were used later to navigate in the lower-dimensional space. The main caveat of this approach is that the generated images are on average more blurry than the real images, which is common with variational auto-encoders. Adversarial models such as BiGANs (Donahue et al., 2016) could provide more refined images. We used this generative model to assess the performance of an event based model that exploit voxel-wise image information to estimate spatial progression of dementia. This model was found to have better performance than state-of-the-art event-based model methods using regional volume biomarkers. The same type of variational auto-encoders can be used to model the spatial progression of other types of diseases, or even phenotypes such as age, provided the target brain region is expected to be affected by the disease to an extent that is visible on the scan, and despite the above-mentioned blur in the reconstruction induced by the variational auto-encoder.

1.2 Automated methods for the quantification of enlarged perivascular spaces

As a technical proof of concept, I demonstrated in Chapter 1 that neural networks can be optimized to automatically replicate PVS visual scores in the basal ganglia. In Chapter 5, I extended this work to other relevant regions, namely the midbrain, the hippocampi, and the centrum semiovale, and showed that the associations between twenty potential determinants of PVS and automated PVS scores was similar to associations between the same potential determinants of PVS and visual PVS scores. These findings suggested that automated PVS scores can replace visual PVS scores in neuroepidemiological studies. While this method was developed with scans acquired in a research setting in a population study (Ikram et al., 2017), in Chapter 6 I showed promising results when directly applying this method on heterogeneous clinical data acquired from different scanners. In Chapter 2, experiments in the basal ganglia suggested that this method can be optimised using only 25 scans, which could enable its optimization on small local datasets. In Chapter 3, I proposed an evolution of the automated PVS scoring method to detect PVS in the four above-mentioned brain regions, and showed that the detection performance was close to the intra-rater agreement. In Chapter 4 we developed another automated PVS detection method that was optimized using dot annotations in the center of PVS instead of only using the visual PVS scores. This method has access to more information for its supervision and achieves a higher agreement with dot annotations than the weakly supervised method presented in Chapter 3. The drawback of this method is that it cannot be optimized only using visual PVS score. We only evaluated the performance of this more strongly supervised method in the centrum semiovale in a single dataset, and the generalization to other regions and other datasets has yet to be determined.

Other researchers have also simultaneously developed automated PVS quantification methods (Boespflug et al., 2017; Ballerini et al., 2018; Lian et al., 2018a; Sudre et al., 2018, 2019; Schwartz et al., 2019; Sepehrband et al., 2019). The methods presented in this thesis have been evaluated more thoroughly than those methods. First, I used larger datatsets (more than 2000 participants) than other methods, which used datasets with between 14 and 100 participants. Second, I have also evaluated the proposed methods on clinical scans (Chapter 6), while most of the other methods used scans acquired in a research setting only. Third, I used a wider range of metrics to evaluate the performance of the PVS quantification: correlation with visual scores, voxel-level detection performance, similarity of the association between determinants of PVS and automated and visual scores. Other methods usually considered at most one of these aspects. Finally, I quantified PVS in four separate brain region while other approaches often focus on a single region (mostly the centrum semiovale). The limit of the proposed methods is that PVS segmentation has not been addressed or quantified. Some methods were designed for PVS segmentation but were not evaluated on the voxel-level (Boespflug et al., 2017; Ballerini et al., 2018). Lian et al. (2018a) quantified PVS segmentation using Dice Similarity Coefficient. However their work was evaluated on only 11 7T scans, and it is unsure how the method generalises to clinical scans or scans acquired at lower magnetic fields.

2 Methodological considerations and limitations

In this section I discuss how observer variability can limit the performance of automated methods, and indentify cases of disagreement between the automated PVS quantification methods and expert annotators. I then discuss the deployment of neural networks to external datasets, and address the limitation of neural networks optimized on a single imaging modality. Lastly, I discuss the interpretability of neural networks.

2.1 Observer variability

Observer variability in the creation of labels and annotations for the optimization of automated methods could limit the highest reachable performance of those automated methods. When optimized with the labels of a single rater, the performance of the automated method could be expected to be limited by the intra-rater agreement. In most applications having a performance at least close to the inter-rater agreement is desired. There can be a large observer variability in the assessment of MRI markers, and especially in the assessment of PVS. For example, in our experiments, the intra-rater overlap of identified PVS in the centrum-semiovale almost never reached more than 60 percent Dice similarity coefficient. In case the labels of several raters are available for the same samples, optimizing the automated method to predict the consensus label might be the most appropriate approach, as it is likely to discard PVS that are only subtly enlarged. Sudre et al. (2019) studied different approaches for handling inter-observer variations in the case of PVS and lacune classification. The observer variability in labeling also depends on the type of labels. Although, as mentioned above, there can be poor agreement on the location of individual PVS in the centrum semiovale, observers display less variability in the assessment of the overall PVS burden: for visual PVS scores, the intra-rater agreement in terms of intraclass correlation coefficient (ICC) is 0.88, and the inter-rater agreement is 0.80 ICC (Adams et al., 2013). There are also multiple visual rating methods for the quantification of PVS. For example, in the Rotterdam Scan Study PVS burden is quantified in an anatomically predefined axial slice of the centrum semiovale (Adams et al., 2013), while the Potter scores rate the axial slice having most PVS Potter et al. (2015c). In a study with our collaborators from the Framingham heart study (Ho et al., 1993) at Boston University, we compared visual rating methods to each other and also to the proposed automated PVS scores in the centrum semiovale presented in Chapter 5. Preliminary results show that the automated PVS quantification method correlates well with Potter scores in this dataset. This suggests that either the automated method appropriately estimates the PVS burden overall in the centrum semiovale, and not only in the visually scored slice of the Rotterdam scores, or more directly that the visual scores of Adams et al. (2013) correlate well with the Potter scores.

2.2 Identification of cases of disagreement between automated PVS detection and expert annotations

Expert raters can disagree on the identification of PVS. It is of interest to verify the cases of disagreement between the automated PVS methods and expert PVS annotations, and compare those to annotations of other brain lesions made by other independent raters. In Part C, I proposed a method to automatically detect PVS and showed that it could reach a performance similar to the intrarater agreement. By visually inspecting the predictions of the networks, I noticed that some of the false positive detections of the network were PVS missed by the expert annotator. It can sometimes be challenging even for experts not to miss any PVS, to decide whether a subtle PVS should be accounted for, and also to differentiate between small WMH, lacunes, motion artifacts and PVS. Quantifying and understanding the potential limits of the proposed automated PVS detection method (GP-Unet method in Chapter 3) in that regard is a substantial step in its deployment for medical research and clinical practice. For this purpose, in 1000 images, I compared all automated and manual PVS detections to WMH using automated segmentation of WMH that had previously been visually checked and corrected by experts. In preliminary experiments using those segmentations in the centrum semiovale, I found that both the expert rater and the algorithm mistook small WMH for PVS in less than 1 percent of all detections of PVS (the overall number of automated and manual detection of PVS was similar). According to these automated and corrected WMH segmentations, the number of mistakes made by the algorithm was similar to the number of mistakes made by the PVS rater. In the basal ganglia, due to the masking of the region during preprocessing, the network mistook periventricular WMH for PVS in 1.6 percent of all automated detections of PVS, while this happened in only two out of 3630 cases for the expert rater. Improving the masking could alleviate this problem. Neither the expert nor the network mistook WMH for PVS in the midbrain and hippocampi. In the same set of scans, we compared PVS detection to lacunes segmented by a second independant rater. Out of 15 segmented lacunes, none of them had been mistaken for PVS by either the rater or network. In clinical scans, the prevalence of lacunes is higher, but the proportion of lacunes to PVS is still so low that mistaking a lacune for a PVS could be expected to only have a negligible impact on the quantification of PVS burden. Lastly, as explained in Chapter 5, the first step of the automated PVS quantification method is to mask the region of interest. When quantifying PVS in the centrum semiovale, the white mater is masked first. When deploying the automated PVS method in datasets with scans at a lower voxel resolution, the masking of the sulci sometimes becomes more erroneous. Some of the cut-off sulci resemble PVS, and the network sometimes mistakes those sulci for PVS. To help solve this problem, the network could be optimized in the centrum semiovale without masking the white matter.

2.3 Deployment to other datasets

Machine learning methods generally work well on data that are similar to the training data, but may perform poorly on external datasets with substantially different characteristics such as, for example, data acquired from another scanner, data with different voxel resolution, or data from another population. Researchers have strived to address this issue, creating a research field called transfer learning (Van Opbroek et al., 2014; van Tulder and de Bruijne, 2016). After publishing the automated PVS quantification method, which was evaluated in the Rotterdam Scan Study, I started to deploy it in several other research centers. I found that I could get satisfying results by resampling the scans of the external datasets to match the voxel resolution of the scans in the training dataset. I also evaluated the ventricle segmentation algorithm of Chapter 8 on an international multi-center dataset, while it was trained with clinical data of a single hospital. I had to include additional data augmentation during training to mimic possible protocol differences in the external data. These results suggest that the features learnt by the neural network generalize sufficiently in my applications. In future work, the results may improve further by fine-tuning the neural networks using the local datasets. For PVS quantification, it may also help to add data augmentation during training to imitate the resolution and intensity contrast in external datasets. Generative adversarial networks

(Bowles et al., 2018) could also be used for data augmentation, or domain adversarial objectives could be included in the optimization process to improve generalization to different data (Kamnitsas et al., 2017).

2.4 Single MRI sequence

The networks presented in this thesis used a single MRI sequence for their optimization. Choosing a single sequence for the optimization originates from practical considerations: designing and monitoring such networks is simpler. Another sequence could always be added as input at later development stages. Below, I discuss using a single sequence for the automated quantification of PVS.

In the vast majority of cases, expert raters can differentiate PVS from other similarly appearing lesions only using this sequence. However, in some rare cases, experts need to check T1-weighted or FLAIR sequence to differentiate PVS from WMH, lacunes or motion artifacts. Considering the low prevalence of such cases, they can be expected not to have significant impact on the automated quantification of PVS burden. Another potential issue related the quantification of PVS from the T2-w sequence alone, is that in some studies, only the T1-w sequence may be acquired (for example the UK-biobank (Sudlow et al., 2015) and Generation R data (Jaddoe et al., 2012)). As the networks were not optimized on T1-w scans, they cannot be expected to compute meaningful predictions. I have now also optimized networks on T1-w scans of the Rotterdam scan study, reaching a performance slightly lower than that obtained with the T2-w scans. This was expected as PVS have been rated by inspecting the T2-w scan primarily.

One could also design a neural network that takes the most relevant modalities (T2-w, T1-w and FLAIR) as input. These approaches have the disadvantage of requiring more information and in case of missing sequences for a subject, the network cannot output its prediction. In addition, depending on the protocol, MRI sequences may have a different appearance and contrast, which sometimes makes labelling scans to MRI sequences challenging. Varsavsky et al. (2018) recently proposed a method in which any number of modalities can be given as input to the network in any random order. Using this method, a single model could be used in datasets having T1-w scans alone, T2-w scans alone, both T1-w and T2-w scans, or even more sequences.

2.5 Interpretability of neural networks

Before applying automated quantification methods in new imaging studies, researchers need to verify that they behave as expected on subsets of data by at least checking visually. While enabling faster development of more powerful prediction models, deep neural networks have been criticized by the scientific community as lacking transparency in the rationale of the computation of predictions. It is common for neural networks to have several millions of parameters, which complexify their interpretation. In contrast with, for example, Chapter 2, where I only quantified the performance of networks by computing metrics on the predicted biomarker values, I have attempted to answer some of these concerns by addressing themes such as weakly supervised detection in Chapter 1 and 3. I provided visual explanations for the networks' predictions of PVS burden, performed additional post-processing to obtain automated detections of PVS, and quantified the results by considering the overlap with manual dot annotations of PVS. In Chapter 4, we provided further transparency by directly optimizing the network using the dot annotations. However, for all methods, the use and meaning of the intermediary feature maps inside the network architecture still remains difficult to interpret.

Another line of research to address the interpretability of networks is to more explicitly model how information is processed in the layers. That is for example the objective of capsule networks which are designed to model hierarchical spatial relationships (Sabour et al., 2017). These methods have only recently showed comparable performance to standard convolutional neural networks, and their optimization can be challenging.

3 Clinical implications

Automatically quantifying PVS burden in different brain regions may aid to stratify cerebral small vessel disease (CSVD) subtypes, with for example, the presence of basal ganglia PVS in arteriolosclerotic small vessel disease and centrum semiovale PVS in cerebral amyloid angiopathy. Cerebral amyloid angiopathy has not only been related to severe PVS burden in the centrum semiovale (Charidimou et al., 2017), but also to a significantly higher risk for intracerebral bleeding in face of oral anticoagulant treatment (Wilson et al., 2018). Accounting for PVS burden in the centrum semiovale could presumably have important therapeutic and prognostic implications in terms of prescribing oral anticoagulants and preventing intracerebral hemorrhage. In the basal ganglia, PVS burden might also be an indicator of motor syndrome and relate to elevated urinary glycosaminoglycan levels (Kwee and Kwee, 2007). In the hippocampi, PVS burden might relate to memory and vascularization patterns.

Recent studies have also highlighted the role of perivascular spaces enlargement for glymphatic clearance during sleep (Brown et al., 2018; Mestre et al., 2017; Rasmussen et al., 2018). Quantifying the evolution of volume and shape of PVS from MRI during sleep could help to better understand the relationship between PVS and glymphatic clearance. However segmenting PVS consistently in such large datasets is difficult for experts. Automated methods for PVS segmentation could help segmenting PVS with higher reproducibility.

During inference, the computation of the automated PVS scores lasted only a few minutes on CPU. Most of this time is spent on the segmentation of brain regions and, in the case of prediction using T2 scans, registration from the T1-weighted scans to the T2-weighted scans, as standards brain parcellation algorithms require T1 scans as input. After this preprocessing, the computation of the automated PVS scores from the segmented brain region lasted a few seconds per brain region on CPU depending on the size of the region. This low computation time can facilitate the implementation of the proposed automated PVS quantification methods in clinical practice.

4 Future directions

In this section I discuss and propose future research directions for weak supervision in the optimization of neural networks. I also discuss the potential of using 7T scans for PVS quantification, and finally CSVD research with a specific focus on PVS.

4.1 Towards less supervision for neural networks

In this thesis, I have presented methods for weakly supervised detection. The objective could also be weakly supervised segmentation, where the size and shape of the target objects could be measured and quantified. With Bortsova et al. (2018), we added a sigmoid activation after the last feature map of weakly supervised regression neural networks similar to those presented in Chapter 3. We evaluated this approach on emphysema quantification from CT scans and showed empirically that we could segment diseased lung tissue. This segmentation could subsequently be to discriminate between paraseptal and centrilobular emphysema. For this discrimination task, the method reached a performance similar to the inter-rater agreement. Unfortunately, this intuitive approach did not provide satisfying results for PVS or WMH segmentation. Most often, during training, the loss function could not converge when adding the sigmoid activation. The topic of weakly supervised segmentation with attention maps of neural networks requires further investigation.

When optimizing a neural network with only image-level labels available, instead on focusing on the target object, the neural network may focus on other objects in the input image that are correlated with the target image-level label. In most cases, this situation is not desired. Preliminary experiments with weakly supervised segmentation of WMH have provided us with more insights in this potential problem. In these experiments, the networks were optimized to predict the WMH volume in voxels, and the predicted segmentations were computed during inference by thresholding the attention maps computed as described in Chapter 3 (GP-Unet). When networks were trained using large datasets, we noticed that the performance and features learnt did not seem to be substantially affected by the randomness of the optimization process. However, in small datasets, we noticed that (a) depending on the random initialization of the weights of the networks, and on randomness in the on-the-fly data augmentation, the networks could learn to segment different objects in the image, and that (b) there was no strong correlation between the performance in WMH volume prediction and the performance in WMH segmentation. The networks for example sometimes also segmented the ventricles in addition to the WMH. Both ventricular volume and WMH volume are associated with age (Kaye et al., 1992; De Leeuw et al., 2001). Ventricular volume and WMH volume can consequently be expected to be correlated, and the network probably picks up this correlation to perform its quantification. The generated attention maps are only a limited representation of what the network learns and disentangling errors due to the network itself from the interpretation of the network's attention is a challenge on its own.

Weakly supervised methods could also be used for biomarker discovery by supervising networks on the image level, using for example disease status labels. If a network can be optimized to accurately predict the image label, attention maps can be computed to reveal the imaging features used by the network. These features relate to the image label. It becomes then more difficult to verify whether the network focuses on meaningful information or on potential confounders, and preventing this can be challenging. Recently Zhao et al. (2019) proposed to exclude confounders from attention maps using a partial back-propagation algorithm that incorporates univariate statistical tests. Another interesting strategy could be to force the network to learn multiple explanations simultaneously and enforce these explanations to be different using for instance normalized cross-correlation on feature maps.

4.2 Optimization of neural network

The optimization strategies and behavior of neural networks is an underrepresented research topic. In the medical image analysis literature, and to some extent in the computer vision literature, performance metrics after optimization receive most attention and the optimization itself is less studied. For example, most researchers use Adam optimizer (Kingma and Ba, 2015) and implement some earlier stopping strategies. For many studies, it is very likely that the relative performance after optimization varies depending on the chosen optimizer and the stopping point during training. In many of my experiments the loss function plateaued for many epochs and suddenly decreased exponentially. In these scenarios, early stopping would lead to substantially worse results. Sometimes using Adam instead of Adadelta (Zeiler, 2012) would lead to the networks not converging at all. Depending on the randomness in the initialization and optimization, different features and performance could be obtained for the same architecture and the same training data. Studying and understanding those mechanisms could be of major importance in the coming developments of neural networks research, and more generally of artificial intelligence.

4.3 Emerging availability of 7T

Exploiting the higher image quality in 7T scans could be necessary to understand PVS in more detail. The higher resolution of 7T scans would enable the computation of substantially more precise shape features, more accurate volume estimates, and ease the observation of short term volumetric changes. A few researchers have already started to focus on PVS in 7T scanners (Feldman et al., 2018; Lian et al., 2018a; Zhang et al., 2016). Because of the practical difficulty to obtain 7T scans, those studies often include only a few participants. Another concern is that because of the higher resolution, many more PVS are visible on 7T in comparison to 1.5T or 3T. The decision of the minimum size of PVS for their quantification and inclusion in neuroepidemiological studies will consequently become even more prominent.

4.4 Research on cerebral small vessel disease (CSVD)

In this thesis, I developed methods for the quantification and localization of PVS burden in the brain. It might be of interest to investigate associations between potential determinants of PVS and PVS volume and shape, and especially associations between PVS volume and shape and CSVD subtypes such as arteriosclerosis and cerebral amyloid angiopathy. For the quantification of these PVS features further methodological development is required, with for example automated segmentation of PVS.

In addition to PVS, other structural brain changes visible on MRI are thought to reflect the presence of CSVD. These brain changes include for example cerebral microbleeds, WMH and lacunes (Das et al., 2019; Shi and Wardlaw, 2016). Recent research in CSVD attempts to establish a global marker for the assessment of CSVD burden by creating sum scores encompassing these different CSVD markers. For example, researchers (Huijts et al., 2013; Staals et al., 2015; Yilmaz et al., 2018) proposed to compute a score that aggregates information about these four CSVD MRI markers by assigning one point for each marker above a predefined threshold, resulting in a CSVD sum score ranging scaling from 0 to 4. Machine learning could also be used to learn more advanced combinations of the different CSVD markers to best describe CSVD burden in relation to outcomes, or even predict outcomes using those CSVD markers. Machine learning sum scores could potentially be a real alternative to manually defined sum scores. They may have stronger association with disease and disease status. End-to-end optimization from image data to CSVD-related outcomes could also reveal currently unknown CSVD imaging markers. The challenges with such machine learning approaches are interpretability and generalizability. The best combinations of CSVD markers in the training dataset may be different from the best combination in an independent external dataset.

Part G

Summary

Recent developments in artificial intelligence research have resulted in tremendous success in computer vision, natural language processing and medical imaging tasks, often reaching human or superhuman performance. In this thesis, I further developed artificial intelligence methods based on convolutional neural networks with a special focus on the automated analysis of brain magnetic resonance imaging scans (MRI). I showed that efficient artificial intelligence systems can be created using only minimal supervision, by reducing the quantity and quality of annotations used for training. I applied those methods to the automated assessment of the burden of enlarged perivascular spaces, brain structural changes that may be related to dementia, stroke, multiple sclerosis, and sleep. The proposed methods provide PVS counts and localization in four brain regions: the midbrain, hippocampi, basal ganglia and centrum semiovale.

In Part B, I developed machine learning methods based on convolutional neural networks to count objects or estimate object volume in 3D volumetric data. In Chapter 1, I developed and applied these methods on the quantification of PVS burden in a subcortical brain region: the basal ganglia. I showed that those methods have a performance similar to that of expert raters, and I showed that those methods need at least about 200 rated MRI scans to reach this performance. In Chapter 2, I proposed a methodological improvement that enables the optimization of those networks using substantially fewer rated scans. The principle of the method is to combine training images to create more virtual training images, hence increase the size of the training dataset and average potential errors in the ratings of experts at the price of, on average, weaker labelling of those images. I showed the added-value of those methods on PVS quantification and quantification of white matter hyperintensities, and showed that, for PVS quantification, inter-rater agreement can be reached using only 25 rated images for the optimization.

In Part C, I focused on convolutional neural networks for the detection of objects that appear multiple times in a single 3D volume. In Chapter 3, I proposed a method for weakly supervised object detection. During training, neural networks were supervised only with image-level labels representing the count of target objects in the 3D volume. During inference, the networks could not only predict the count of target objects but also localize them. I designed this method to find small objects in large 3D volumes. The method was evaluated with detection of handwritten digits in 2D images, and of PVS in brain MRI scans. In chapter 4, we proposed another neural network approach for detection that was supervised using generated voxel-wise labels. Those labels corresponded to a distance map computed from manual dot annotations at the center of the target objects. We proposed to use distance maps based on the intensity difference between neighboring voxels instead of the more standard Euclidean distance. We evaluated the method on the detection of PVS in the upper-area of the white matter: the centrum semiovale. These methods reached a performance similar to that of the intra-rater agreement.

In Part D, I focused on the application of methods of Chapter 1 to support automated PVS count in epidemiological studies. In Chapter 5, I developed and evaluated the method for other clinically important brain regions: the midbrain, hippocampi, and the centrum semiovale. I did extensive experiments where I showed that using either manual PVS scores or the proposed automated PVS scores gave similar results for the analysis of the associations between twenty potential determinants of PVS and PVS burden. These results suggested that the automated PVS scores could be used instead of visual PVS scores for neuroepidemiological studies. In Chapter 6, I showed that the methods presented in Chapter 5 could also be applied on 76 clinical scans acquired using

9 different scanners from the University Hospital Magdeburg in Germany for the quantification of PVS burden in the centrum semiovale. Good results were also achieved for the other brain regions, but the performance of the automated method was lower than the intra-rater performance.

Neural networks were not only successful for the quantification of PVS. In Part E, I proposed automated methods for other applications in neurology research. In Chapter 7, we proposed an event-based model to model the progression of dementia. The events considered in the model are biomarkers quantifying neuropathological changes. The objective of this type of model is to estimate the order in which these biomarkers become abnormal. Contrary to current models, which only use scalar biomarkers, with the proposed method we can also use voxel-wise biomarkers. We evaluated the proposed model on the progression of Alzheimer's disease. To validate the performance of the model, we designed a generative machine learning method based on auto-encoders. With this method, we could artificially generate 3D images of brain regions at a given probability of being affected by Alzheimer's disease. Chapter 8 has multiple contributions. First, I designed a method to segment the ventricular system from clinical brain MRI scans. I validated this method in an international dataset of twelve sites with different scanners and protocols. Designing a method that can robustly segment the ventricular system across low resolution, multi-center clinical datasets is a technical challenge. Secondly, the method was then used to assess the quality of pair-wise registration of MRI scans to a common shared brain template, with potential applications in lesion-symptom mapping. Lastly, I used the automated assessment method to build multi-atlas registration framework that uses different age specific atlases as an intermediary registration step to improve registration quality.

Part H

Dutch Summary

Recente ontwikkelingen in de kunstmatige intelligentie hebben ervoor gezorgd dat er grote successen zijn geboekt op het gebied van computer vision, het verwerken van natuurlijke taal en de medische beeldvorming. In veel gebieden presteert kunstmatige intelligentie al op hetzelfde niveau als de mens. Er zijn zelfs gevallen waarin de kunstmatige intelligentie taken beter kan uitvoeren dan de mens. In dit proefschrift heb ik methoden verder ontwikkeld uit een subdomein van kunstmatige intelligentie, namelijk convolutionele neurale netwerken. De nadruk ligt op de automatische analyse van magnetic resonance imaging (MRI) beelden van de hersenen. Ik richt me in het bijzonder op het reduceren van de kwantiteit en de kwaliteit van de annotaties die nodig zijn om robuuste modellen te trainen. Daarnaast heb ik nieuwe methoden ontwikkeld om het aantal verwijde perivasculaire ruimtes automatisch te meten, evenals de locatie. Verwijde perivasculaire ruimtes, ook wel Virchow-Robin ruimtes genoemd, zijn structurele veranderingen in de hersenen. De voorgestelde methoden geven een telling en een lokalisatie van de verwijde perivasculaire ruimtes in vier hersengebieden, namelijk in het mesencephalon, de hippocampus, de basale ganglia en het centrum semiovale.

In deel B heb ik methoden ontwikkeld op het gebied van machine learning die gebaseerd zijn op convolutionele neurale netwerken die gebruikt kunnen worden om objecten te tellen of om het volume van objecten te bepalen op basis van 3D volumetrische data. In hoofdstuk 1 heb ik deze methoden ontwikkeld en toegepast voor het bepalen van het aantal verwijde perivasculaire ruimtes in de basale ganglia, een subcorticaal hersengebied. In hoofdstuk 2 stel ik een methodologische verbetering voor die het mogelijk maakt om de netwerken te optimaliseren met substantieel minder geannoteerde scans. In dit hoofdstuk heb ik laten zien dat deze methoden een toegevoegde waarde hebben voor de kwantificatie van de witte stof hyperintensiteiten die bij verwijde perivasculaire ruimtes worden waargenomen. Daarnaast heb ik laten zien dat voor de kwantificatie van verwijde perivasculaire ruimtes maar 25 geannoteerde afbeeldingen nodig zijn voor de optimalisatie van het netwerk om dezelfde annotatie betrouwbaarheid te bereiken als een menselijke beoordelaar.

In deel C heb ik de focus gelegd op convolutionele neurale netwerken die gebruikt kunnen worden voor de detectie van objecten die meerdere keren voorkomen in een enkel 3D volume. In hoofdstuk 3 stel ik een methode voor die gebruikt kan worden voor object detectie met zwakke labels. Tijdens de optimalisatie van het netwerk wordt alleen gebruik gemaakt van de informatie over het aantal objecten in het 3D volume. Later, als het netwerk wordt toegepast op nieuwe data, wordt ook de locatie van de objecten in de 3D volumes verschaft. In hoofdstuk 4 stellen we een ander neuraal netwerk voor dat gebruikt kan worden voor detectie. De netwerken worden geoptimaliseerd om de afstand tot een object te voorspellen. Deze afstand is gebaseerd op intensiteit tussen aangrenzende voxels, in plaats van de euclidische afstand die over het algemeen gebruikt wordt.

In deel D focus ik op het toepassen van de methoden uit hoofdstuk 1 voor het automatisch tellen van de verwijde perivasculaire ruimtes in hersenscans in epidemiologische studies. In hoofdstuk 5 heb ik een methode ontwikkeld en geëvalueerd voor andere klinisch relevante hersengebieden, namelijk behalve de basale ganglia ook het mesencephalon, de hippocampus, en het centrum semiovale. Dit suggereerde dat de geautomatiseerde telling van de verwijde perivasculaire ruimtes gebruikt kan worden voor neuro-epidemiologische onderzoeken in plaats van de visuele telling van verwijde perivasculaire ruimtes die momenteel gebruikt wordt. In hoofdstuk 6 heb ik aangetoond dat de methoden die in hoofdstuk 5 voorgesteld waren ook toegepast kunnen worden voor het kwantificeren van het aantal verwijde perivasculaire ruimtes in het centrum semiovale in 76 klinische scans die verkregen zijn door negen verschillende scanners van het universitair ziekenhuis Maagdenburg in Duitsland.

In deel E stel ik andere geautomatiseerde methoden voor die kunnen worden toegepast in neurologisch onderzoek. In hoofdstuk 7 stellen we een event-gedreven model voor om het verloop van dementie te modelleren. Daarnaast ontwikkelden we een methode, gebaseerd op het gebied van machine learning, die kunstmatige 3D afbeeldingen van hersengebieden genereert op basis van de waarschijnlijkheid dat de ziekte van Alzheimer wordt vastgesteld. In hoofdstuk 8 heb ik een methode ontworpen om het ventriculaire systeem te segmenteren in klinische MRI-scans van de hersenen. Dit is vervolgens gebruikt om een methode te ontwikkelen die meerdere leeftijdsspecifieke atlassen gebruikt in beeldregistratie als een tussenliggende stap om de kwaliteit van de registratie te bevorderen.

Part I

Acknowledgments

First of all, I would like to thank the Dutch society for supporting the research presented in this book. I also would like to thank the Dutch society on a personal level, as I think I have been, from the start, very welcome in this country. There is no need to say that the comfortable and peaceful life style offered by the Dutch society to PhD students like me significantly contributes to the success of their academic project. I have done in my best in thesis to not only focus on methodological developments but also on the translation of the findings to real world applications. It is very important to me that my research not only creates additional knowledge but also helps to improve the life of the society that financed it. I would like to thank the Netherlands Organisation for Health Research and Development ZonMw for financing my project.

Second I would like to thank my parents for offering me the necessary education to carry on this research. My parents have always tried the best to give me as much as they could, and I am infinitely thankful for that. Without them, this book would not have existed in the first place.

I would like to deeply thank my supervisors Prof. Marleen de Bruijne, Prof. Meike Vernooij and Prof. Wiro Niessen for guiding this research. I think you all three have amazing abilities that perfectly complement each other. You formed a perfect a trio, I honestly do not know how it could have been better.

Marleen, thank you for your close supervision and your availability. I have been feeling extremely comfortable in your management style, which I think significantly contributed to produce the best output for the project. You meet each of your students one hour every week and, at same time, provide them an almost infinite freedom in their research directions and implementations, a perfect balance between freedom and guidance. I think it is a very rare feature for someone of your experience to manage to find the time for those weekly meetings. I very much like that your let you students the freedom to explore the research paths that motivates them most, and that you embrace the versatility of the consequent research landscape in your group by always managing to identify potential pitfalls in their approaches. I have been feeling extremely comfortable with your supervision. I am very satisfied that all the discussions we have had substantially improved the quality of the work for the project. I am also very thankful for your substantial contribution in my education as a scientist. Thank you again.

Meike, thank you very much for your supervision too. Your medical expertise has always been crucial to define the medical objectives of the research project and identify potential pitfalls. I admire your reactivity and organisation. You're extremely fast, precise, and always very positive and enthusiastic. It is extremely pleasant to work with you. I also deeply appreciate that you are open minded about using, understanding and developing technologies to help advance medical research. It is impressive to see how you managed to combine extraordinary academic achievements next to your busy clinical activity, and now even parenting. I have much admiration for you.

Wiro, thank you very much for your supervision. Without you, this research project would definitely not have been as successful. You gave me very nice feedback and helped me significantly in the few situations where I asked for assistance. Your supervision was of course not as close as that of Marleen, but I think I have had the chance to meet with you regularly, and I learnt a lot from each of our encounters. I admire you as a person, I am still trying to find out your secret to successfully lead our research group, translate academic research to practise with the spin-off company Quantib that you founded, and have built a big family. It might take me some additional time, but I will find out and copy it!

Hieab, thank you very much for your advises. You motivated me initially to focus on the quantification of enlarged perivascular spaces, which I think has been very fruitful. The current research project is in great part based on your work, and we would not have reached the current results without you. You also helped significantly at some key moments. I learnt much from you. I admire your ambition, determination, organisation and charisma. Despite your young age, you have already done so much and I don't see how anything can stop you.

I would like to thank the Dutch association "De Drie Lichten", literally translated to "The Three Lights", for financing my research project and visit at the Massachusetts General Hospital in the group of Prof Natalia Rost.

Natalia, thank you very much for welcoming me in your group and for giving me the opportunity to learn about stroke research. I think you are one of the friendliest supervised I have met. Thanks to you, I have felt very welcome in Boston. I deeply enjoyed my experience there, and I think it significantly contributed to further develop my research project and my education as a scientist. I also could not help notice how you place the health of your past, current and future patients at the core of your activity and thoughts. It is very admirable. On a side note, I only discovered later your excellent mastery of the French language. Next time we meet in person, I wish to spectate it in all its magnificence!

Markus, thank you very much for you help and supervision during my project in Boston. I learnt much from you about how to work with clinical data. You always tried to find out what could be the best strategy to accomplish the objectives I had set for my visit in Boston. Following your advices and suggestions, everything worked out extremely well. I am very thankful for your guidance, and also for the work we have continued to do together after my visit. It was also very nice to see you at MICCAI in China!

I will take the opportunity to thank the other people at JPK. Starting of course with Marco, with whom I have had long and intense philosophical discussions;

Shannon, you were extremely helpful in securing the computer parts I ordered; Katy, thank you for arranging all administrative tasks; Suzanne, Anne-Katrin, Mark, Ona, Anand, Andreas, Nicolas, Steven, Jasmine, Edip, Elif, and Devin of course, thank you all for your scientific and less scientific discussions.

Of course, I have also met and worked with other amazing researchers in other groups in Boston. I would like to thank you all too: Sudha, Rafael, Marna, Hugo, Mohamed, Eugenio, Bruce, Polina, Bernhard, Razvan, Clinton, Lila, Charles, Michele, Alfredo. Thank you Sudha, Bruce, Polina and Charles for introducing me to your research group. I am especially very happy to continue the collaborations we started with Sudha and with Charles.

Let's go back to Europe now.

Steffen, thank very much for hosting me in your lab. It was a pleasure working with you on perivascular spaces during my stay in your group, and it is a pleasure to continue this collaboration. My stay in your group was both very productive and very enjoyable, and I only keep good memories from it. Thanks to all MedDigit related people: Max of course, it was a pleasure working in a duo with you; Soumick, Alex, Filip, Denver, Kunal, Sumaiya, Juliane, Yannic, Falk.

Gerda, of course thank you extremely much. For working together, brainstorming, criticizing each other research ideas and making them stronger, reviewing each other papers, sharing tips about what we learnt separately, and maybe most importantly for mutual emotional support. You are my partner in crime, my life friend. Do not take too much time to finish this PhD, I need you at the next place I am going to.

Pinar, thank you very much for collaborating with me on the quantification of enlarged perivascular spaces by giving your medical expert advice and by brainstorming together. Thank you very much for your enthusiasm and availability. And thank you for promoting our research. I think our work has been and will be extremely useful in the translation from methodological development towards usage in medical research. We could not have done this without you.

Kim, thank you very much for brainstorming together. You have immensely contributed to the research project on enlarged perivascular spaces. I think I can say that you have been my most hardcore student, Oliver not being very far behind. I think your efforts completely paid off. I am very happy you got your thesis cum laude and an additional national award. You are raw power, and you are unstoppable once canalized.

Oliver, thank you very much for working together. Your project was extremely interesting and I think is shaping my future research ideas and plans. We have encountered many surprises along the way, and we also learnt much. And eventually your hard work paid off. It is fun to think that we will graduate around the same time! Too bad you are not continuing in academia, you could have brought further significant contributions. As long as you continue to work the common good, whether it is in academia or not, I warmly support of course!

Thanks to all my other students or students I co-supervised: Filipe, Pauline, Florian, Robin, Corentin, Axel, Benjamin, Antonin, Nicolas, Laetitia, Arne.

Gijs van Tulder, thank you so much for being my office mate during almost four years. Things became different after you left, not as fun and philosophical. I have learnt much from you in computer science which greatly helped me in my research project. I am very thankful for that. You are a well of knowledge (sometimes backed up DuckDuckGo) and always ready to help, I was literally blessed to have been sitting in front of you such a long time. You also taught me about economics, Dutch language, Dutch regulations, Dutch customs, and many more things. You have been away only since a few months when I am writing this and I miss you much already! I hope I will be able to hire you once I have built my research group. You are an incredible asset.

Vikram, thank you for brainstorming together and sharing our PhD experience. It is fun to think that we started our PhD on the same day! It was extremely nice to write that work on the dementia project together. It is too bad we could not continue working on it. But yes, it is important to have your priorities set!

Chaoping, thank you for introducing me to a new research field and for giving me the opportunity to develop AI methods for that field. I learnt much from our collaboration about the acceleration of MRI acquisition. I was a pleasure working with you!

Shuai, thank you your Majesty. It was also a pleasure to work with you, and of course to have fun and go to Chinese restaurants together with Hua. Thank you very much for showing us China. My experience there was incredible and you significantly contributed to building it!

Saima, thank you for collaborating on enlarged perivascular spaces. I appreciate your enthusiasm and determination very much. I hope this is the beginning of a long lasting collaboration!

Marius, thank you for your advises at the beginning of my project, and for those during my stay in Boston. Eventually we did not extensively work together but I have learnt from you, and I hope that those moments were fruitful for you too. For some reason, we keep meeting each other by chance at random locations in Rotterdam. It was a pleasure discussing with you every time!

Gena, thank you for brainstorming on the visualization of neural networks. I admire your hard work, intelligence and social abilities. And also, you have made me laugh countless times!

Henri, thank you for introducing me and guiding me through a foreign world: the neuro folder of the cluster. I have used much of your work in my own project. It was also very enjoyable playing board games with you and doing the interview for the Beeldspraak.

Martijn, thank you for supervising students together. I admire your ambition and organisation. If only we have had more time, I am sure we could have done even greater projects together!

Sebastian, thank you for organizing all those events. You did significantly contribute to make life at BIGR super fun, with the board games, escape rooms, movie nights and the many other activities you got yourself involved in!

Marcel and Hakim, thank you for being so helpful with the BIGR-cluster. You definitely helped speeding up things. It was also a pleasure to play boardgames with you!

Thank you to all other current and former BIGR members, Veronika, Annegreet, Zahra, Marco, Hua, Pierre, Jean-Marie, Yuan Yuan, Natalia, Anne, John, Arno, Antonio, Dirk, Esther, Jifke, Adriaan, Stefan, Theo, Danielo, Luisa, Maria, Muhammad, Bo, Emanoel, Wietske, Gokhan, Jiahang, Riwaj, Thomas, Yao Yao, Willem, Mart, Kasper.

Thank you Johan and Gabriel for hosting us. The working space at EMC is amazing. And it is extremely fruitful for our research to work so close to medical doctors. I hope this sort of collaborative workspace between computer scientists and medical doctors flourishes even more in the coming years.

Thank you Desiree, Marise, Petra, Tineke, Andreas for helping for with administrative tasks. Andreas, thank you for being so helpful. It was very nice chatting with you. You also contributed to make me feel welcome in the Netherlands and in EMC.

Thanks to researchers from epidemiology, Reffat, Tavia, Thom, Maria, Arfan. Tavia, it was really nice to work on PVS with you. It was also nice to see you in Boston!

Thanks to all the people in Quantib: Jorrit, Valerio, Reinhard, Jim, Arna, Renske, Rudolf, Arthur, Jaap. I was very nice attending your meetings and brainstorming together. I was also fruitful to collaborate with data annotation on microbleeds and enlarged perivascular. Jorrit, thank you very much for coordinating this.

Thank you Cait and Kim to have helped me to translate the summary of thesis.

I have had the chance to meet so many extraordinary and amazing people and researchers during my PhD. I am very thankful to all of you for helping me in the accomplishment of the research project, for contributing to my development as a scientist, and for all the great conversations and moments we shared.

Part J

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: A system for large-scale machine learning, in: 12th Symposium on Operating Systems Design and Implementation (16), pp. 265–283.

Achiron, A., Faibel, M., 2002. Sandlike appearance of virchow-robin spaces in early multiple sclerosis: a novel neuroradiologic marker. American journal of neuroradiology 23, 376–380.

Adams, H.H., Cavalieri, M., Verhaaren, B.F., Bos, D., van der Lugt, A., Enzinger, C., Vernooij, M.W., Schmidt, R., Ikram, M.A., 2013. Rating method for dilated virchow-robin spaces on magnetic resonance imaging. Stroke 44, 1732–1735.

Adams, H.H., Hilal, S., Schwingenschuh, P., Wittfeld, K., van der Lee, S.J., DeCarli, C., Vernooij, M.W., Katschnig-Winter, P., Habes, M., Chen, C., et al., 2015. A priori collaboration in population imaging: the uniform neuro-imaging of virchow-robin spaces enlargement consortium. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 513–520.

Adams, H.H., Schwingenschuh, P., van der Lugt, A., Katschnig-Winter, P., Hofman, A., Krestin, G., Vernooij, M., Schmidt, R., Ikram, M.A., 2014. Determinants of enlarged virchow-robin spaces: The univrse consortium. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 10, P408.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems, pp. 9524–9535.

Ahrens, W., Hoffmann, W., Jöckel, K.H., Kaaks, R., Gromer, B., Greiser, K.H., Linseisen, J., Schmidt, B., Wichmann, H.E., Weg-Remers, S., 2014. The German

National Cohort: Aims, study des. European Journal of Epidemiology 29, 371–382.

Alam, M.M., Islam, M.T., 2019. Machine learning approach of automatic identification and counting of blood cells. Healthcare technology letters 6, 103–108.

Andrews, S., Tsochantaridis, I., Hofmann, T., 2003. Support vector machines for multiple-instance learning, in: Advances in neural information processing systems, pp. 577–584.

Antonelli, M., Cardoso, M.J., Johnston, E.W., Appayya, M.B., Presles, B., Modat, M., Punwani, S., Ourselin, S., 2019. Gas: A genetic atlas selection strategy in multi-atlas segmentation framework. Medical image analysis 52, 97–108.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Atlason, H.E., Shao, M., Robertsson, V., Sigurdsson, S., Gudnason, V., Prince, J.L., Ellingsen, L.M., 2019. Large-scale parcellation of the ventricular system using convolutional neural networks, in: Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, International Society for Optics and Photonics. p. 109530N.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis 12, 26–41.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage 54, 2033–2044.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging .

Ballerini, L., Lovreglio, R., Hernandez, M.D.C., Gonzalez-Castro, V., Maniega, S.M., Pellegrini, E., Bastin, M.E., Deary, I.J., Wardlaw, J.M., 2016. Application of the Ordered Logit Model to Optimising Frangi Filter Parameters for Segmentation of Perivascular Spaces. Procedia Computer Science 90, 61–67.

Ballerini, L., Lovreglio, R., Hernández, M.d.C.V., Ramirez, J., MacIntosh, B.J., Black, S.E., Wardlaw, J.M., 2018. Perivascular spaces segmentation in brain mri using optimal 3d filtering. Scientific reports 8, 1–11.

Bandos, A.I., Rockette, H.E., Song, T., Gur, D., 2009. Area under the free-response roc curve (froc) and a related summary index. Biometrics 65, 247–256.

Banerjee, G., Kim, H.J., Fox, Z., Jäger, H.R., Wilson, D., Charidimou, A., Na, H.K., Na, D.L., Seo, S.W., Werring, D.J., 2017. Mri-visible perivascular space location is associated with alzheimer's disease independently of amyloid burden. Brain 140, 1107–1116.

Biesbroek, J.M., Kuijf, H.J., van der Graaf, Y., Vincken, K.L., Postma, A., Mali, W.P., Biessels, G.J., Geerlings, M.I., Group, S.S., et al., 2013. Association between subcortical vascular lesion location and cognition: a voxel-based and tract-based lesion-symptom mapping study. the smart-mr study. PloS one 8, e60541.

Bilello, M., Akbari, H., Da, X., Pisapia, J.M., Mohan, S., Wolf, R.L., O'Rourke, D.M., Martinez-Lage, M., Davatzikos, C., 2016. Population-based mri atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma. NeuroImage: Clinical 12, 34–40.

de Boer, R., Vrooman, H.A., Ikram, M.A., Vernooij, M.W., Breteler, M.M., van der Lugt, A., Niessen, W.J., 2010. Accuracy and reproducibility study of automatic mri brain tissue segmentation methods. Neuroimage 51, 1047–1056.

Boespflug, E.L., Schwartz, D.L., Lahna, D., Pollock, J., Iliff, J.J., Kaye, J.A., Rooney, W., Silbert, L.C., 2017. Mr imaging–based multimodal autoidentification of perivascular spaces (mmaps): automated morphologic segmentation of enlarged perivascular spaces at clinical field strength. Radiology 286, 632–642.

Boespflug, E.L., Schwartz, D.L., Lahna, D., Pollock, J., Iliff, J.J., Kaye, J.A., Rooney, W., Silbert, L.C., 2018. Mr imaging–based multimodal autoidentification of perivascular spaces (mmaps): automated morphologic segmentation of enlarged perivascular spaces at clinical field strength. Radiology 286, 632–642.

Bortsova, G., Dubost, F., Ørting, S., Katramados, I., Hogeweg, L., Thomsen, L., Wille, M., de Bruijne, M., 2018. Deep learning from label proportions for emphysema quantification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 768–776.

Bortsova, G., van Tulder, G., Dubost, F., Peng, T., Navab, N., van der Lugt, A., Bos, D., De Bruijne, M., 2017. Segmentation of Intracranial Arterial Calcification with Deeply Supervised Residual Dropout Networks. Springer International Publishing, Cham. pp. 356–364.

Bouvy, W.H., Zwanenburg, J.J., Reinink, R., Wisse, L.E., Luijten, P.R., Kappelle, L.J., Geerlings, M.I., Biessels, G.J., group, U.V.C.I.V.S., 2016. Perivascular spaces on 7 tesla brain mri are related to markers of small vessel disease but not to age or cardiovascular risk factors. Journal of Cerebral Blood Flow & Metabolism 36, 1708–1717.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernández, M.V., Wardlaw, J., Rueckert, D., 2018. Gan augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 .

Brefeld, U., Geibel, P., Wysotzki, F., 2003. Support vector machines with example dependent costs, in: European Conference on Machine Learning, Springer. pp. 23–34.

Brefeld, U., Scheffer, T., 2004. Co-em support vector learning, in: Proceedings of the twenty-first international conference on Machine learning, ACM. p. 16.

Bron, E.E., Steketee, R.M., Houston, G.C., Oliver, R.A., Achterberg, H.C., Loog, M., van Swieten, J.C., Hammers, A., Niessen, W.J., Smits, M., et al., 2014. Diagnostic classification of arterial spin labeling and structural mri in presenile early stage dementia. Human brain mapping 35, 4916–4931.

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2016. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE transactions on medical imaging 35, 1229–1239.

Brown, R., Benveniste, H., Black, S.E., Charpak, S., Dichgans, M., Joutel, A., Nedergaard, M., Smith, K.J., Zlokovic, B.V., Wardlaw, J.M., 2018. Understanding the role of the perivascular space in cerebral small vessel disease. Cardiovascular research .

de Bruijn, R.F., Bos, M.J., Portegies, M.L., Hofman, A., Franco, O.H., Koudstaal, P.J., Ikram, M.A., 2015. The potential for prevention of dementia across two decades: the prospective, population-based rotterdam study. BMC medicine 13, 132.

Buyck, J.F., Dufouil, C., Mazoyer, B., Maillard, P., Ducimetiere, P., Alpérovitch, A., Bousser, M.G., Kurth, T., Tzourio, C., 2009. Cerebral white matter lesions are associated with the risk of stroke but not with other vascular events: the 3-city dijon study. Stroke 40, 2327–2331.

Charidimou, A., Boulouis, G., Pasi, M., Auriel, E., van Etten, E.S., Haley, K., Ayres, A., Schwab, K.M., Martinez-Ramirez, S., Goldstein, J.N., et al., 2017. Mri-visible perivascular spaces in cerebral amyloid angiopathy and hypertensive arteriopathy. Neurology 88, 1157–1164.

Charidimou, A., Meegahage, R., Fox, Z., Peeters, A., Vandermeeren, Y., Laloux, P., Baron, J.C., Jäger, H.R., Werring, D.J., 2013. Enlarged perivascular spaces as a marker of underlying arteriopathy in intracerebral haemorrhage: a multicentre mri cohort study. J Neurol Neurosurg Psychiatry 84, 624–629.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A., 2018. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. NeuroImage 170, 446–455.

Chen, P.F., Krim, H., Mendoza, O.L., 2010. Multiphase joint segmentation-registration and object tracking for layered images. IEEE transactions on image processing 19, 1706–1719.

Chen, Y., Gurol, M., Rosand, J., Viswanathan, A., Rakich, S., Groover, T., Greenberg, S., Smith, E., 2006. Progression of white matter lesions and hemorrhages in cerebral amyloid angiopathy. Neurology 67, 83–87.

Chollet, F., et al., 2015. Keras. `https://github.com/fchollet/keras`.

Cicchetti, D.V., 1994. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. Psychological Assessment 6, 284–290.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G., 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124.

Courand, P.Y., Serraille, M., Grandjean, A., Tilikete, C., Milon, H., Harbaoui, B., Lantelme, P., 2019. Recurrent vertigo is a predictor of stroke in a large cohort of hypertensive patients. Journal of hypertension 37, 942–948.

Dalca, A.V., Bobu, A., Rost, N.S., Golland, P., 2016. Patch-based discrete registration of clinical brain images, in: International Workshop on Patch-based Techniques in Medical Imaging, Springer. pp. 60–67.

Das, A.S., Regenhardt, R.W., Vernooij, M.W., Blacker, D., Charidimou, A., Viswanathan, A., 2019. Asymptomatic cerebral small vessel disease: Insights from population-based studies. Journal of stroke 21, 121.

De Leeuw, F., de Groot, J.C., Achten, E., Oudkerk, M., Ramos, L., Heijboer, R., Hofman, A., Jolles, J., Van Gijn, J., Breteler, M., 2001. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the rotterdam scan study. Journal of Neurology, Neurosurgery & Psychiatry 70, 9–14.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 248–255.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006a. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006b. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31, 968–980.

Dewey, B.E., Carass, A., Blitz, A.M., Prince, J.L., 2017. Efficient multi-atlas registration using an intermediate template image, in: Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging, International Society for Optics and Photonics. p. 101371F.

Dickie, D.A., Shenkin, S.D., Anblagan, D., Lee, J., Blesa Cabez, M., Rodriguez, D., Boardman, J.P., Waldman, A., Job, D.E., Wardlaw, J.M., 2017. Whole brain magnetic resonance image atlases: a systematic review of existing atlases and caveats for use in population imaging. Frontiers in neuroinformatics 11, 1.

Donahue, J., Krähenbühl, P., Darrell, T., 2016. Adversarial feature learning. arXiv preprint arXiv:1605.09782 .

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V.C., Shi, L., Heng, P.A., 2016. Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks. IEEE Transactions on Medical Imaging 35, 1182–1195.

Dubost, F., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., de Bruijne, M., 2019a. 3d regression neural network for the quantification of enlarged perivascular spaces in brain mri. Medical image analysis 51, 89–100.

Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W.J., Vernooij, M., De Bruijne, M., 2017. GP-Unet: Lesion detection from weak labels with a 3D regression network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 214–221.

Dubost, F., Yilmaz, P., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., de Bruijne, M., 2019b. Enlarged perivascular spaces in brain MRI: Automated quantification in four regions. NeuroImage 185, 534–544.

Duperron, M.G., Tzourio, C., Schilling, S., Zhu, Y.C., Soumaré, A., Mazoyer, B., Debette, S., 2019. High dilated perivascular space burden: a new mri marker for risk of intracerebral hemorrhage. Neurobiology of Aging 84, 158–165.

Earnest, M.P., Heaton, R.K., Wilkinson, W.E., Manke, W.F., 1979. Cortical atrophy, ventricular enlargement and intellectual impairment in the aged. Neurology 29, 1138–1138.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al., 2012. 3d slicer as an image computing platform for the quantitative imaging network. Magnetic resonance imaging 30, 1323–1341.

Feldman, R.E., Rutland, J.W., Fields, M.C., Marcuse, L.V., Pawha, P.S., Delman, B.N., Balchandani, P., 2018. Quantification of perivascular spaces at 7 t: A potential mri biomarker for epilepsy. Seizure 54, 11–18.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence 32, 1627–1645.

Fillmore, P.T., Phillips-Meek, M.C., Richards, J.E., 2015. Age-specific mri brain and head templates for healthy adults from 20 through 89 years of age. Frontiers in aging neuroscience 7, 44.

Fletcher, P.T., Venkatasubramanian, S., Joshi, S., 2009. The geometric median on riemannian manifolds with application to robust atlas estimation. NeuroImage 45, S143–S152.

Fong, R.C., Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation, in: The IEEE International Conference on Computer Vision (ICCV).

Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., et al., 2012. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. NeuroImage 60, 1880–1889.

Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A., 1998. Multiscale vessel enhancement filtering, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 130–137.

Ganzetti, M., Liu, Q., Mantini, D., Initiative, A.D.N., et al., 2018. A spatial registration toolbox for structural mr imaging of the aging brain. Neuroinformatics , 1–13.

Ghafoorian, M., Karssemeijer, N., Heskes, T., Bergkamp, M., Wissink, J., Obels, J., Keizer, K., de Leeuw, F.E., van Ginneken, B., Marchiori, E., et al., 2017. Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. NeuroImage: Clinical 14, 391–399.

Ghafoorian, M., Teuwen, J., Manniesing, R., de Leeuw, F.E., van Ginneken, B., Karssemeijer, N., Platel, B., 2018. Student beats the teacher: deep neural networks for lateral ventricles segmentation in brain mr, in: Medical Imaging 2018: Image Processing, International Society for Optics and Photonics. p. 105742U.

Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D., 2016. An Artificial Agent for Anatomical Landmark Detection in Medical Images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 229–237.

Giese, A.K., Schirmer, M.D., Donahue, K.L., Cloonan, L., Irie, R., Winzeck, S., Bouts, M.J., McIntosh, E.C., Mocking, S.J., Dalca, A.V., et al., 2017. Design and rationale for examining neuroimaging genetics in ischemic stroke: The mri-genie study. Neurology Genetics 3, e180.

Girshick, R., 2015. Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

González, G., Washko, G.R., Estépar, R.S.J., 2018. Deep learning for biomarker regression: application to osteoporosis and emphysema on chest ct scans, in:

Medical Imaging 2018: Image Processing, International Society for Optics and Photonics. p. 105741H.

González-Castro, V., Hernández, M.d.C.V., Armitage, P.A., Wardlaw, J.M., 2016. Automatic rating of perivascular spaces in brain mri using bag of visual words, in: International Conference Image Analysis and Recognition, Springer. pp. 642–649.

Gonzalez-Castro, V., Hernández, M.d.C.V., Chappell, F.M., Armitage, P.A., Makin, S., Wardlaw, J.M., 2017. Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance. Clinical Science 131, 1465–1481.

de Groot, M., Vernooij, M.W., Klein, S., Ikram, M.A., Vos, F.M., Smith, S.M., Niessen, W.J., Andersson, J.L., 2013. Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. Neuroimage 76, 400–411.

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M.d.C., Dickie, D., Wardlaw, J., et al., 2018. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. NeuroImage: Clinical 17, 918–934.

Gutierrez, J., DiTullio, M., Cheung, Y.K.K., Alperin, N., Bagci, A., Sacco, R.L., Wright, C.B., Elkind, M.S., Rundek, T., 2019. Brain arterial dilatation modifies the association between extracranial pulsatile hemodynamics and brain perivascular spaces: the northern manhattan study. Hypertension Research 42, 1019–1028.

Gutierrez, J., Rundek, T., Ekind, M., Sacco, R.L., Wright, C.B., 2013. Perivascular spaces are associated with atherosclerosis: an insight from the northern manhattan study. American Journal of Neuroradiology 34, 1711–1716.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: Achievements and challenges. Journal of digital imaging , 1–15.

Hilal, S., Ikram, M.K., Saini, M., Tan, C.S., Catindig, J.A., Dong, Y.H., Lim, L.B.S., Ting, E.Y., Koo, E.H., Cheung, C.Y., Qiu, A., Wong, T.Y., Chen, C.L.H., Venketasubramanian, N., 2013. Prevalence of cognitive impairment in Chinese: Epidemiology of Dementia in Singapore study. Journal of Neurology, Neurosurgery and Psychiatry 84, 686–692.

Ho, K.K., Pinsky, J.L., Kannel, W.B., Levy, D., 1993. The epidemiology of heart failure: the framingham study. Journal of the American College of Cardiology 22, A6–A13.

Huang, J., Alexander, D., 2012. Probabilistic event cascades for alzheimer's disease, in: Advances in neural information processing systems, pp. 3095–3103.

Huijts, M., Duits, A., Van Oostenbrugge, R.J., Kroon, A.A., De Leeuw, P.W., Staals, J., 2013. Accumulation of mri markers of cerebral small vessel disease is associated with decreased cognitive function. a study in first-ever lacunar stroke and hypertensive patients. Frontiers in aging neuroscience 5, 72.

Hussain, S.J., Savitri, T.S., Devi, P.S., 2013. Detection of hydrocephalus lateral ventricles quantitatively in brain mri images of infants. International Journal of Computer Applications 83.

Hwang, S., Kim, H.E., 2016. Self-transfer learning for weakly supervised lesion localization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 239–246.

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: a survey. Medical image analysis 24, 205–219.

Ikram, M.A., Brusselle, G.G., Murad, S.D., van Duijn, C.M., Franco, O.H., Goedegebure, A., Klaver, C.C., Nijsten, T.E., Peeters, R.P., Stricker, B.H., et al., 2017. The rotterdam study: 2018 update on objectives, design and main results. European Journal of Epidemiology 32, 807–850.

Ikram, M.A., van der Lugt, A., Niessen, W.J., Koudstaal, P.J., Krestin, G.P., Hofman, A., Bos, D., Vernooij, M.W., 2015. The Rotterdam Scan Study: design update 2016 and main findings. European Journal of Epidemiology 30, 1299–1315.

Ilse, M., Tomczak, J.M., Welling, M., 2018. Attention-based deep multiple instance learning. ICML .

Inoue, H., 2018. Data augmentation by pairing samples for images classification. arXiv preprint arXiv:1801.02929 .

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, pp. 448–456.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., Clifford R. Jack Jr., MD1,*, Matt A. Bernstein, PhD1, Nick C. Fox, MD2, Paul Thompson, PhD3, Gene Alexander, PhD4, Danielle Harvey, PhD5, Bret Borowski, RTR1, Paula J. Britson, BS1, Jennifer L. Whitwell, PhD1, Chadwick Ward, BA1, Anders M. Dale, PhD6, Joe, M.W.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging 27, 685–691.

Jaddoe, V.W., van Duijn, C.M., Franco, O.H., van der Heijden, A.J., van IIzendoorn, M.H., de Jongste, J.C., van der Lugt, A., Mackenbach, J.P., Moll, H.A., Raat, H., et al., 2012. The generation r study: design and cohort update 2012. European journal of epidemiology 27, 739–756.

Joshi, S., Davis, B., Jomier, M., Gerig, G., 2004. Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23, S151–S160.

Jungreis, C.A., Kanal, E., Hirsch, W., Martinez, A., Moossy, J., 1988. Normal perivascular spaces mimicking lacunar infarction: Mr imaging. Radiology 169, 101–104.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International conference on information processing in medical imaging, Springer. pp. 597–609.

Kaye, J.A., DeCarli, C., Luxenberg, J.S., Rapoport, S.I., 1992. The significance of age-related enlargement of the cerebral ventricles in healthy men and women measured by quantitative computed x-ray tomography. Journal of the American Geriatrics Society 40, 225–231.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. International Conference on Learning Representations .

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. International Conference on Learning Representations .

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. Neuroimage 46, 786–802.

Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J., 2010. Elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Transactions on Medical Imaging 29, 196–205.

Korbar, B., Olofson, A.M., Miraflor, A.P., Nicka, C.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A., Hassanpour, S., 2017. Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE. pp. 821–827.

Kuhn, H.W., 1955. The hungarian method for the assignment problem. Naval research logistics quarterly 2, 83–97.

Kuijf, H.J., Biesbroek, J.M., de Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities; results of the wmh segmentation challenge. IEEE transactions on medical imaging .

Kwee, R.M., Kwee, T.C., 2007. Virchow-robin spaces at mr imaging. Radiographics 27, 1071–1086.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks, in: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE. pp. 239–248.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 .

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets, in: Artificial Intelligence and Statistics, pp. 562–570.

Lee, J.H., Kim, K.G., 2018. Applying deep learning in medical images: The case of bone age estimation. Healthcare informatics research 24, 86–92.

Lempitsky, V., Zisserman, A., 2010. Learning to count objects in images, in: Advances in neural information processing systems, pp. 1324–1332.

Lian, C., Liu, M., Zhang, J., Zong, X., Lin, W., Shen, D., 2018a. Automatic segmentation of 3d perivascular spaces in 7t mr images using multi-channel fully convolutional network, in: Proceedings of the International Society for Magnetic Resonance in Medicine... Scientific Meeting and Exhibition. International Society for Magnetic Resonance in Medicine. Scientific Meeting and Exhibition, NIH Public Access.

Lian, C., Zhang, J., Liu, M., Zong, X., Hung, S.C., Lin, W., Shen, D., 2018b. Multi-channel multi-scale fully convolutional network for 3d perivascular spaces segmentation in 7t mr images. Medical image analysis 46, 106–117.

Liang, P., Shi, L., Chen, N., Luo, Y., Wang, X., Liu, K., Mok, V.C., Chu, W.C., Wang, D., Li, K., 2015. Construction of brain atlases based on a multi-center mri dataset of 2020 chinese adults. Scientific reports 5, 18216.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision 60, 91–110.

Maillard, P., Mitchell, G.F., Himali, J.J., Beiser, A., Tsao, C.W., Pase, M.P., Satizabal, C.L., Vasan, R.S., Seshadri, S., DeCarli, C., 2016. Effects of arterial stiffness on brain integrity in young adults from the framingham heart study. Stroke 47, 1030–1036.

Martinez-Ramirez, S., Pontes-Neto, O.M., Dumas, A.P., Auriel, E., Halpin, A., Quimby, M., Gurol, M.E., Greenberg, S.M., Viswanathan, A., 2013. Topography of dilated perivascular spaces in subjects from a memory clinic cohort. Neurology , 10–1212.

Mestre, H., Kostrikov, S., Mehta, R.I., Nedergaard, M., 2017. Perivascular spaces, glymphatic dysfunction, and small vessel disease. Clinical science 131, 2257–2274.

Meyer, M.I., Galdran, A., Mendonça, A.M., Campilho, A., 2018a. A pixel-wise distance regression approach for joint retinal optical disc and fovea detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 39–47.

Meyer, M.I., Galdran, A., Mendonça, A.M., Campilho, A., 2018b. A pixel-wise distance regression approach for joint retinal optical disc and fovea detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 39–47.

Miao, S., Wang, Z.J., Liao, R., 2016. A cnn regression approach for real-time 2d/3d registration. IEEE transactions on medical imaging 35, 1352–1363.

Mills, S., Cain, J., Purandare, N., Jackson, A., 2007. Biomarkers of cerebrovascular disease in dementia. The British journal of radiology 80, S128–S145.

Mnih, V., Heess, N., Graves, A., et al., 2014. Recurrent models of visual attention, in: Advances in neural information processing systems, pp. 2204–2212.

Moore, E.A., Graves, M.J., Prince, M.R., McRobbie, D.W., 2006. MRI from Picture to Proton. Cambridge University Press.

Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K., 2016. A large contextual dataset for classification, detection and counting of cars with deep learning, in: European Conference on Computer Vision, Springer. pp. 785–800.

Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T., 2000. Text classification from labeled and unlabeled documents using em. Machine learning 39, 103–134.

Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., Askham, H., Romera-Paredes, B., Karthikesalingam, A., Chu, C., et al., 2018. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430 .

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention U-Net: Learning where to look for the pancreas. Medical Imaging with Deep Learning Conference .

Ou, Y., Akbari, H., Bilello, M., Da, X., Davatzikos, C., 2014. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. IEEE transactions on medical imaging 33, 2039–2065.

Oxtoby, N.P., Alexander, D.C., 2017. Imaging plus x: multimodal models of neurodegenerative disease. Current opinion in neurology 30, 371.

Park, S.H., Zong, X., Gao, Y., Lin, W., Shen, D., 2016. Segmentation of perivascular spaces in 7 t mr image using auto-context model with orientation-normalized features. NeuroImage 134, 223–235.

Parthasarathy, V., Hatt, C., Stankovic, Z., Raval, A., Jain, A., 2011. Real-time 3d ultrasound guided interventional system for cardiac stem cell therapy with motion compensation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 283–290.

Patankar, T.F., Mitra, D., Varma, A., Snowden, J., Neary, D., Jackson, A., 2005. Dilatation of the Virchow-Robin space is a sensitive indicator of cerebral microvascular disease: study in elderly patients with dementia. American Journal of Neuroradiology 26, 1512–1520.

Paul Cohen, J., Boucher, G., Glastonbury, C.A., Lo, H.Z., Bengio, Y., 2017. Count-ception: Counting by fully convolutional redundant counting, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 18–26.

Petsiuk, V., Das, A., Saenko, K., 2018. Rise: Randomized input sampling for explanation of black-box models, in: British Machine Vision Conference.

Platt, J., et al., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10, 61–74.

Pohl, K.M., Fisher, J., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2006. A bayesian model for joint segmentation and registration. NeuroImage 31, 228–239.

Pollock, H., Hutchings, M., Weller, R.O., Zhang, E.T., 1997. Perivascular spaces in the basal ganglia of the human brain: Their relationship to lacunes. Journal of Anatomy 191, 337–346.

Potter, G.M., Chappell, F.M., Morris, Z., Wardlaw, J.M., 2015a. Cerebral perivascular spaces visible on magnetic resonance imaging: Development of a qualitative rating scale and its observer reliability. Cerebrovascular Diseases 39, 224–231.

Potter, G.M., Doubal, F.N., Jackson, C.A., Chappell, F.M., Sudlow, C.L., Dennis, M.S., Wardlaw, J.M., 2015b. Enlarged perivascular spaces and cerebral small vessel disease. International Journal of Stroke 10, 376–381.

Potter, G.M., Doubal, F.N., Jackson, C.A., Chappell, F.M., Sudlow, C.L., Dennis, M.S., Wardlaw, J.M., 2015c. Enlarged perivascular spaces and cerebral small vessel disease. International journal of stroke 10, 376–381.

Prince, M., Wimo, A., Guerchet, M., Ali, G., Wu, Y., Prina, M., et al., 2015. The global impact of dementia: an analysis of prevalence, incidence, cost and trends. World Alzheimer Report 2015.

Qi, H., Collins, S., Noble, A., 2017. Weakly supervised learning of placental ultrasound images with residual networks, in: Annual Conference on Medical Image Understanding and Analysis, Springer. pp. 98–108.

Qi, H., Collins, S., Noble, J.A., 2018. Automatic lacunae localization in placental ultrasound images via layer aggregation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 921–929.

Ramirez, J., Berezuk, C., McNeely, A.A., Scott, C.J., Gao, F., Black, S.E., 2015. Visible virchow-robin spaces on magnetic resonance imaging of alzheimer’s disease patients and normal elderly from the sunnybrook dementia study. Journal of Alzheimer’s Disease 43, 415–424.

Rasmussen, M.K., Mestre, H., Nedergaard, M., 2018. The glymphatic pathway in neurological disorders. The Lancet Neurology 17, 1016–1024.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Ren, M., Zemel, R.S., 2017a. End-to-end instance segmentation with recurrent attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6656–6664.

Ren, M., Zemel, R.S., 2017b. End-to-end instance segmentation with recurrent attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6656–6664.

Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., et al., 2019. Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance 21, 18.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roy, S., Carass, A., Jog, A., Prince, J.L., Lee, J., 2014. Mr to ct registration of brains using image synthesis, in: Medical Imaging 2014: Image Processing, International Society for Optics and Photonics. p. 903419.

Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules, in: Advances in neural information processing systems, pp. 3856–3866.

Sanchez, C.E., Richards, J.E., Almli, C.R., 2012. Age-specific mri templates for pediatric neuroimaging. Developmental neuropsychology 37, 379–399.

Sangineto, E., Nabi, M., Culibrk, D., Sebe, N., 2018. Self paced deep learning for weakly supervised object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1.

Sarvaiya, J.N., Patnaik, S., Bombaywala, S., 2009. Image registration by template matching using normalized cross-correlation, in: 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, IEEE. pp. 819–822.

Schirmer, M.D., Dalca, A.V., Sridharan, R., Giese, A.K., Donahue, K.L., Nardin, M.J., Mocking, S.J., McIntosh, E.C., Frid, P., Wasselius, J., et al., 2019a. White matter hyperintensity quantification in large-scale clinical acute ischemic stroke cohorts–the mri-genie study. NeuroImage: Clinical , 101884.

Schirmer, M.D., Giese, A.K., Fotiadis, P., Etherton, M.R., Cloonan, L., Viswanathan, A., Greenberg, S.M., Wu, O., Rost, N., 2019b. Spatial signature of white matter hyperintensities in stroke patients. Frontiers in Neurology 10, 208.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 146–157.

Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention-gated networks for improving ultrasound scan plane detection. Medical Imaging with Deep Learning Conference .

Schott, J.M., Bartlett, J.W., Fox, N.C., Barnes, J., Investigators, A.D.N.I., 2010. Increased brain atrophy rates in cognitively normal older adults with low cerebrospinal fluid aβ1-42. Annals of neurology 68, 825–834.

Schwartz, D.L., Boespflug, E.L., Lahna, D.L., Pollock, J., Roese, N.E., Silbert, L.C., 2019. Autoidentification of perivascular spaces in white matter using clinical field strength t1 and flair mr imaging. NeuroImage , 116126.

Seguí, S., Pujol, O., Vitria, J., 2015. Learning to count with deep object features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 90–96.

Selvarajah, J., Scott, M., Stivaros, S., Hulme, S., Georgiou, R., Rothwell, N., Tyrrell, P., Jackson, A., 2009. Potential surrogate markers of cerebral microvascular angiopathy in asymptomatic subjects at risk of stroke. European radiology 19, 1011–1018.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization., in: ICCV, pp. 618–626.

Sepehrband, F., Barisano, G., Sheikh-Bahaei, N., Cabeen, R.P., Choupan, J., Law, M., Toga, A.W., 2019. Image processing approaches to enhance perivascular space visibility and quantification using mri. Scientific reports 9, 1–12.

Shao, M., Han, S., Carass, A., Li, X., Blitz, A.M., Shin, J., Prince, J.L., Ellingsen, L.M., 2019. Brain ventricle parcellation using a deep neural network: Application to patients with ventriculomegaly. NeuroImage: Clinical , 101871.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Shi, Y., Wardlaw, J.M., 2016. Update on cerebral small vessel disease: a dynamic whole-brain disease. Stroke and vascular neurology 1, 83–92.

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps, in: International Conference for Learning Representations Workshop.

Simonyan, K., Zisserman, A., 2015a. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations abs/1409.1, 1–10. arXiv:1409.1556.

Simonyan, K., Zisserman, A., 2015b. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference for Learning Representations .

Sixt, L., Wild, B., Landgraf, T., 2018. Rendergan: Generating realistic labeled data. Frontiers in Robotics and AI 5, 66.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., et al., 2004. Advances in functional and structural mr image analysis and implementation as fsl. Neuroimage 23, S208–S219.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net, in: International Conference for Learning Representations.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

Staals, J., Booth, T., Morris, Z., Bastin, M.E., Gow, A.J., Corley, J., Redmond, P., Starr, J.M., Deary, I.J., Wardlaw, J.M., 2015. Total mri load of cerebral small vessel disease and cognitive ability in older people. Neurobiology of aging 36, 2806–2811.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al., 2015. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 12, e1001779.

Sudre, C.H., Anson, B.G., Ingala, S., Lane, C.D., Jimenez, D., Haider, L., Varsavsky, T., Smith, L., Jäger, H.R., Cardoso, M.J., 2018. 3D multirater RCNN for multimodal multiclass detection and characterisation of extremely small objects. arXiv preprint arXiv:1812.09046 .

Sudre, C.H., Anson, B.G., Ingala, S., Lane, C.D., Jimenez, D., Haider, L., Varsavsky, T., Tanno, R., Smith, L., Ourselin, S., et al., 2019. Let's agree to disagree: Learning highly debatable multirater labelling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 665–673.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging 15, 29.

Tan, R.x., Zhang, J., Chen, P., Wang, B., Xia, Y., 2018. Cells counting with convolutional neural network, in: International Conference on Intelligent Computing, Springer. pp. 102–111.

Toivanen, P.J., 1996. New geodosic distance transforms for gray-scale images. Pattern Recognition Letters 17, 437–450.

Tokozume, Y., Ushiku, Y., Harada, T., 2018. Learning from between-class examples for deep sound recognition. International Conference on Learning Representations .

van Tulder, G., de Bruijne, M., 2016. Representation learning for cross-modality classification, in: Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging. Springer, pp. 126–136.

Uiterwijk, R., van Oostenbrugge, R.J., Huijts, M., De Leeuw, P.W., Kroon, A.A., Staals, J., 2016. Total cerebral small vessel disease mri score is associated with cognitive decline in executive function in patients with hypertension. Frontiers in aging neuroscience 8, 301.

Van Opbroek, A., Ikram, M.A., Vernooij, M.W., De Bruijne, M., 2014. Transfer learning improves supervised image segmentation across imaging protocols. IEEE transactions on medical imaging 34, 1018–1030.

Varsavsky, T., Eaton-Rosen, Z., Sudre, C.H., Nachev, P., Cardoso, M.J., 2018. Pimms: Permutation invariant multi-modal segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 201–209.

Vemuri, B.C., Ye, J., Chen, Y., Leonard, C.M., 2003. Image registration via level-set motion: Applications to atlas-based segmentation. Medical image analysis 7, 1–20.

Venkatraghavan, V., Bron, E.E., Niessen, W.J., Klein, S., Initiative, A.D.N., et al., 2019. Disease progression timeline estimation for alzheimer's disease using discriminative event based modeling. NeuroImage 186, 518–532.

Wack, D.S., Dwyer, M.G., Bergsland, N., Di Perri, C., Ranza, L., Hussein, S., Ramasamy, D., Poloni, G., Zivadinov, R., 2012. Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates. BMC medical imaging 12, 17.

Walach, E., Wolf, L., 2016. Learning to count with cnn boosting, in: European conference on computer vision, Springer. pp. 660–676.

Walhovd, K.B., Westlye, L.T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Agartz, I., Salat, D.H., Greve, D.N., Fischl, B., et al., 2011. Consistent neuroanatomical age-related volume differences across multiple samples. Neurobiology of aging 32, 916–932.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE transactions on pattern analysis and machine intelligence 35, 611–623.

Wang, J., Knol, M., Tiulpin, A., Dubost, F., De Bruijne, M., Vernooij, M., Adams, H., Ikram, M.A., Niessen, W., Roshchupkin, G., 2019. Grey matter age prediction as a biomarker for risk of dementia: A population-based study. BioRxiv , 518506.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., T O'Brien, J., Barkhof, F., Benavente, O.R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology 12, 822–838.

Wenham, P., Price, W., Blundell, G., 1991. Apolipoprotein e genotyping by one-stage pcr. The Lancet 337, 1158–1159.

Wieberdink, R.G., Ikram, M.A., Hofman, A., Koudstaal, P.J., Breteler, M.M., 2012. Trends in stroke incidence rates and stroke risk factors in rotterdam, the netherlands from 1990 to 2008. European journal of epidemiology 27, 287–295.

van Wijnen, K.M., Dubost, F., Yilmaz, P., Ikram, M.A., Niessen, W.J., Adams, H., Vernooij, M.W., de Bruijne, M., 2019. Automated lesion detection by regressing intensity-based distance with a neural network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–242.

Wilson, D., Ambler, G., Shakeshaft, C., Brown, M.M., Charidimou, A., Salman, R.A.S., Lip, G.Y., Cohen, H., Banerjee, G., Houlden, H., et al., 2018. Cerebral microbleeds and intracranial haemorrhage risk in patients anticoagulated for atrial fibrillation after acute ischaemic stroke or transient ischaemic attack (cromis-2): a multicentre observational cohort study. The Lancet Neurology 17, 539–547.

Wu, G., Jia, H., Wang, Q., Shen, D., 2010. Groupwise registration with sharp mean, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 570–577.

Xia, Y., Hu, Q., Aziz, A., Nowinski, W.L., 2004. A knowledge-driven algorithm for a rapid and automatic extraction of the human cerebral ventricular system from mr neuroimages. NeuroImage 21, 269–282.

Xie, W., Noble, J.A., Zisserman, A., 2018a. Microscopy cell counting and detection with fully convolutional regression networks. Computer methods in biomechanics and biomedical engineering: Imaging & Visualization 6, 283–292.

Xie, Y., Xing, F., Shi, X., Kong, X., Su, H., Yang, L., 2018b. Efficient and robust cell detection: A structured regression approach. Medical image analysis 44, 245–254.

Xie, Y., Xing, F., Shi, X., Kong, X., Su, H., Yang, L., 2018c. Efficient and robust cell detection: A structured regression approach. Medical image analysis 44, 245–254.

Xue, Y., Ray, N., Hugh, J., Bigras, G., 2016. Cell counting by regression using convolutional neural network, in: European Conference on Computer Vision, Springer. pp. 274–290.

Yezzi, A., Zöllei, L., Kapur, T., 2003. A variational framework for integrating segmentation and registration through active contours. Medical image analysis 7, 171–185.

Yilmaz, P., Ikram, M.K., Niessen, W.J., Ikram, M.A., Vernooij, M.W., 2018. Practical small vessel disease score relates to stroke, dementia, and death: The rotterdam study. Stroke 49, 2857–2865.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks?, in: Advances in neural information processing systems, pp. 3320–3328.

Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., et al., 2018. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. Nature communications 9, 4273.

Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C., 2014. A data-driven model of biomarker changes in sporadic alzheimer’s disease. Brain 137, 2564–2577.

Young, A.L., Oxtoby, N.P., Ourselin, S., Schott, J.M., Alexander, D.C., Initiative, A.D.N., et al., 2015. A simulation system for biomarker evolution in neurodegenerative disease. Medical image analysis 26, 47–56.

Zeiler, M.D., 2012. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 .

Zhang, C.R., Cloonan, L., Fitzpatrick, K.M., Kanakis, A.S., Ayres, A.M., Furie, K.L., Rosand, J., Rost, N.S., 2015. Determinants of white matter hyperintensity burden differ at the extremes of ages of ischemic stroke onset. Journal of Stroke and Cerebrovascular Diseases 24, 649–654.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018a. mixup: Beyond empirical risk minimization. International Conference on Learning Representations .

Zhang, J., Gao, Y., Park, S.H., Zong, X., Lin, W., Shen, D., 2016. Segmentation of perivascular spaces using vascular features and structured random forest from 7t mr image, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 61–68.

Zhang, Q., Wu, Y.N., Zhu, S.C., 2018b. Interpretable convolutional neural networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8827–8836.

Zhang, Q.s., Zhu, S.C., 2018. Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering 19, 27–39.

Zhao, Q., Adeli, E., Pfefferbaum, A., Sullivan, E.V., Pohl, K.M., 2019. Confounder-aware visualization of convnets, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 328–336.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.

Zhou, M., Scott, J., Chaudhury, B., Hall, L., Goldgof, D., Yeom, K.W., Iv, M., Ou, Y., Kalpathy-Cramer, J., Napel, S., et al., 2018. Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. American Journal of Neuroradiology 39, 208–216.

Zhu, Y.C., Chabriat, H., Godin, O., Dufouil, C., Rosand, J., Greenberg, S.M., Smith, E.E., Tzourio, C., Viswanathan, A., 2012. Distribution of white matter hyperintensity in cerebral hemorrhage and healthy aging. Journal of neurology 259, 530–536.

Zijlmans, J.C.M., Daniel, S.E., Hughes, A.J., Révész, T., Lees, A.J., 2004. Clinicopathological investigation of vascular parkinsonism, including clinical criteria for diagnosis. Movement Disorders 19, 630–640.

Part K

List of Publications

Journal Articles

- **Dubost, F.**, Yilmaz, P., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Enlarged Perivascular Spaces in Brain MRI: Automated Quantification in four Regions. Neuroimage. 2019.

- **Dubost, F.**, Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. 3D Regression Neural Network for the Quantification of Enlarged Perivascular Spaces in Brain MRI. Medical Image Analysis. 2019.

- Wang, J., Knol, M., Tiulpin, A., **Dubost, F.**, de Bruijne, M., Vernooij, M., Adams, H., Ikram, M.A., Niessen, W. and Roshchupkin, G. Grey Matter Age Prediction as a Biomarker for Risk of Dementia. Proceedings of the National Academy of Sciences. 2019.

- **Dubost, F.**, de Bruijne, M., Nardin, M.J., Dalca, A.V., Donahue, K.L., Giese, A., Etherton, M.R., Wu, O., de Groot, M., Niessen, W., Vernooij, M.W., Rost, N.S., and Schirmer, M.D. Automated image registration quality assessment utilizing deep-learning based ventricle extraction in clinical data. Submitted.

- **Dubost, F.**, Adams, H., Yilmaz, P., Bortsova, G., van Tulder, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Weakly Supervised Object Detection with 2D and 3D Regression Neural Networks. Submitted.

- Bortsova, G., Bos, D., **Dubost, F.**, Vernooij, M.W., Ikram M.K., van Tulder, G. and de Bruijne, M. Automated Assessment of Intracranial Carotid Artery Calcification Using Deep Learning. Submitted.

- Chen, S., Sedghi Gamechi, Z., **Dubost, F.**, van Tulder, G. and de Bruijne, M. An End-to-end Approach to Segmentation with CNN and Posterior-CRF. In preparation.

- **Dubost, F.**, van Wijnen, K., Yilmaz, P., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M.W. and de Bruijne, M. Detection of Enlarged Perivascular Spaces in four Brain Regions with Deep Networks. In preparation.

- **Dubost, F.**, van Wijnen, K., Yilmaz, P., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M.W. and de Bruijne, M. Automated quantification of enlarged perivascular spaces: a volumetric analysis. In preparation.

- **Dubost, F.**, et al. Reproducibility of Ratings of Enlarged Perivascular Spaces: the Framingham Heart Study. In preparation.

- Lysen, S.T., Yilmaz P., **Dubost F.**, de Bruijne, M., Ikram, M.A., Vernooij, M.W. and Luik, A.I. Sleep and perivascular spaces in the middle-aged and elderly population. In preparation.

- Vinke, E.J., Yilmaz, P., van der Toorn, J., Fakhry, R., Frenzen, K., **Dubost, F.**, Licher, S., de Bruijne, M., Ikram, M.A., Vernooij, M.W. and Bos,D. Intracranial arteriosclerosis and brain structure in aging. In preparation.

Conference Full Articles

- Kayal, S., **Dubost, F.**, Tiddens, H., de Bruijne, M. Spectral Data Augmentation Techniques to quantify Lung Pathology from CT-images. 2020 IEEE International Symposium on Biomedical Imaging (ISBI).

- **Dubost, F.**, Collery, B., Renaudier, A., Roc, A., Posocco, N., Bortsova, G., Niessen, W., Bortsova, G., de Bruijne, M. Automated Estimation of the

Spinal Curvature via Spine Centerline Segmentation with Cascaded Neural Networks. MICCAI workshop CS 2019 (oral).

- **Dubost, F.***, Duennwald M.*, Scheumann, V., Schreiber, F., Huff, D., Vernooij, M., Niessen, W., Skalej, M., Schreiber, S., Oeltze-Jafra, S.**, de Bruijne M**. Automated Quantification of Enlarged Perivascular Spaces in Clinical Brain MRI across Sites. MICCAI workshop MLCN 2019 (oral).

- Zhang, C., **Dubost, F.**, de Bruijne, M., Klein, S., Poot, D.H.J. APIR-Net: Autocalibrated Parallel Imaging Reconstruction using a Neural Network. MICCAI workshop MLMIR 2019 (oral).

- **Dubost, F.**, Bortsova, G., Adams, H., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Hydranet: Data Augmentation for Regression Neural Networks. MICCAI 2019.

- van Wijnen K.*, **Dubost, F*.**, Yilmaz, P., Ikram, M.A., Niessen W., Adams, H., Vernooij, M. and de Bruijne, M. Automated Lesion Detection by Regressing Intensity-Based Distance with a Neural Network. MICCAI 2019.

- Bortsova, G., **Dubost, F.**, Katramados, I., Hogeweg, and de Bruijne, M. Semi-supervised Medical Image Segmentation via Learning Consistency under Transformations. MICCAI 2019.

- Egger, B., Schirmer, M.D., **Dubost, F.**, Nardin, M.J., Rost, N.S., and Golland, P. Patient-specific Conditional Joint Models of Shape, Image Features and Clinical Indicators. MICCAI 2019.

- Venkatraghavan, V.*, **Dubost, F.***, Bron, E.E., Niessen, W.J., de Bruijne, M. and Klein, S. Event-Based Modeling with High-Dimensional Imaging

* and ** indicate equal contribution

Biomarkers for Estimating Spatial Progression of Dementia. IPMI 2019. (oral)

- Bortsova, G., **Dubost, F.**, Ørting, S.N., Katramados, I., Hogeweg, L., Thomsen, L.H., Wille M., and de Bruijne, M. Deep Learning from Label Proportions for Emphysema Quantification. MICCAI 2018.

- Marques F., **Dubost, F.**, Kemner-van de Corput M., Tiddens H. A. W., de Bruijne. M. Quantification of Lung Abnormalities in Cystic Fibrosis Using Deep Networks. SPIE Medical Imaging 2018 (oral).

- **Dubost, F.**, Bortsova, G., Adams, H., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. GP-Unet: Lesion detection from weak labels with a 3D regression network. MICCAI 2017.

- Bortsova, G., van Tulder, G., **Dubost, F.**, Peng, T., Navab, N., van der Lugt, A., Bos, D. and de Bruijne, M. Segmentation of intracranial arterial calcification with deeply supervised residual dropout networks. MICCAI 2017.

- **Dubost, F.**, Peter, L., Rupprecht, C., Becker, B.G. and Navab, N. Hands-Free Segmentation of Medical Volumes via Binary Inputs. In Deep Learning and Data Labeling for Medical Applications, MICCAI workshop LABELS 2016.

Conference Abstracts

- **Dubost, F.**, Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Automated detection of enlarged perivascular spaces on brain MRI. European Congress of Radiology 2019 (oral).

- Bortsova G., **Dubost F.**, van der Lugt A., Bos D., de Bruijne M. Automatic detection of intracranial calcifications in CT using deep learning. European Congress of Radiology 2019 (oral).

- **Dubost, F.**, Adams, H., Bortsova, G., Ikram, M.A., Niessen, W.J., Vernooij, M. and de Bruijne, M. Automatic quantification of enlarged perivascular spaces on brain MRI. European Congress of Radiology 2018 (oral).
- Bortsova, G., van Tulder, G., **Dubost, F.**, van der Lugt, A., Bos, D. and De Bruijne, M. Automatic detection of intracranial calcifications in non-contrast CT. European Congress of Radiology 2018 (oral).
- Bortsova G., Ørting S.N., **Dubost F.**, Katramados I., Hogeweg L., Wille M.M., Thomsen L.H. and de Bruijne, M. Automatic prediction of emphysema extent in low-dose CT by deep learning. European Congress of Radiology 2018 (oral).

Part L

PhD Portfolio

Courses

- 2019: Logframe for grant writing
- 2019: Grant writing for Dutch charities
- 2018: Scientific Integrity
- 2018: MISS summer school
- 2018: Scientific Visualization with Blender
- 2017: Computer Vision by Learning
- 2017: Advanced Pattern Recognition
- 2017: Biomedical English Writing and Communication
- 2017: NFBIA summer school
- 2016: Front-End Vision and Multi-Scale Image Analysis

International and local research meetings

- NCCV 2019, Wageningen (Netherlands): Oral and Poster
- MICCAI 2019, Shenzhen (China): Oral and Poster
- ECR 2018, Vienna (Austria): Oral presentation
- ICT.OPEN 2018, Amersfoort (Netherlands): Oral and Poster
- MICCAI 2017, Quebec City (Canada): Poster
- ICT.OPEN 2017, Amersfoort (Netherlands): Oral and Poster
- MISP 2017, Attendee

- NVPHBV fall 2017, Eindhoven (Netherlands): Oral presentation
- MICCAI 2016, Athens (Greece): Workshop poster
- NVPHBV spring 2016, Rotterdam (Netherlands): Oral presentation
- ECCV 2016, Amsterdam (Netherlands): Attendee
- Erasmus MC - Medical Informatics research lunch meeting (biweekly), 2016-2019
- Erasmus MC - Model Based Medical Image Analysis Group Meeting (weekly), 2016-2019
- Erasmus MC - Biomedical Imaging Group Seminars, 2016-2019
- Erasmus MC - Biomedical Imaging Group Literature Meetings, 2016-2019

Fellowships, Awards and Competitions

- 2019: 3rd place in international competition, Accurate Automated Spinal Curvature Estimation, MICCAI 2019 challenge, China
- 2018: Personal Grant (5.8k€), De Drie Lichten association, Netherlands
- 2018: Research Grant with Prof. MD. M. Vernooij (GPU equivalent 6k€), NVIDIA, U.S.
- 2018: 3rd place in international competition (leaderboard03/09/2018), White Matter Hyperintensity Segmentation Challenge, Netherlands
- 2017 - 2018: Pilot grants (2X450.000 Billing Units = 7.4k€), SURFsara, Dutch National Computing Facilities, Netherlands

- 2017: 3rd Prize Poster Award, ICT.OPEN (Dutch IT research conference), Netherlands
- 2017: 2nd Prize Poster Award, NFBIA summer school (Dutch summer school on medical image analysis), Netherlands

Supervision of Graduate Students

- 2019 – Present: Co-supervision of 3 Master Students, collaboration between Brigham and Women Hospital, Harvard Medical School, U.S. and Erasmus MC, Netherlands
- 2018: Member of a Master's thesis defence committee at TU Delft, Netherlands
- 2017 – Present: Supervision of 13 MSc. thesis students and interns (with one cum Laude Master's thesis), Erasmus MC, Netherlands

Teaching activities

- 2017 - Present: Training of medical students for the rating and annotation of brain lesions from MRI scans, Quantib BV, Netherlands and Otto von Gerricke University Hospital, Germany
- 2019: Teaching assistant for the Clinical Technology bachelor program at Delft University of Technology, Netherlands

Program Committee

2019: Member of program committees (NeurIPS conference workshop: Medical Imaging meets NeurIPS, Canada, and Medical Image Computing and Computer Assisted Intervention conference workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis, China)

INVITED LECTURES

- Massachusetts Institute of Technology (MIT), Computer Science Artificial Intelligence Laboratory, Boston, U.S., 2019 (Invited by Prof. Golland).
- Boston University, Boston, U.S., 2019 (Invited by Prof. Seshadri).
- Harvard Medical School, Massachusetts General Hospital, Boston, U.S., 2019 (Invited by Prof. Rost).
- Harvard Medical school, Athinoula A. Martinos Center, Boston, U.S., 2019 (Invited by Prof. Fischl).
- Otto von Gerricke University Hospital, Department of Neurology, Magdeburg, Germany, 2019 (Invited by Prof. Schreiber and Dr. Oeltze-Jafra).
- Netherlands Organisation for Applied Scientific Research (TNO), The Hague, The Netherlands, 2018 (Invited by Dr. Van Opbroek).

SOCIETAL IMPACT OF OWN RESEARCH

I disseminate the outcomes of my research on Twitter (1.2k followers – fpgdubost) and on YouTube. I also have public access Github repositories.

VOLUNTEER WORK

2016: Staff member for the European Conference on Computer Vision (ECCV), the Netherlands

REVIEWING ACTIVITIES

- 2017 – Present: Reviewer for the journals: Medical Image Analysis; Neuroimage; Transactions on Medical Imaging; and Transactions on Biomedical Engineering

- 2017 – Present: Reviewer for international conferences in medical image analysis (MICCAI, ISBI, MIDL, Med-NeurIPS)

Memberships of Scientific Societies

- 2018 - Present: Member, European Society of Radiology
- 2017 - Present: Member, Dutch Society of Pattern Recognition and Image Processing
- 2016 - Present: Member, Medical Image Computing and Computer Assisted Intervention Society

Major Collaborations

- Prof. Rost, Massachusetts General Hospital, Harvard Medical School, U.S.
- Prof. Guttmann, Brigham and Women Hospital, Harvard Medical School, U.S.
- Prof. Seshadri, Boston University, U.S., and University of Texas Health Science Center at San Antonio, U.S.
- Prof. Schreiber and Dr. Oeltze-Jafra, Department of Neurology, Otto von Gerricke University Hospital, Germany

Part M

About the Author

Florian Dubost was born Clamart, France on the 27th December 1992. He finished high school in July 2010 from Versailles academy, France. He joined preparatory classes in Janson de Sailly, Paris, France, where he followed the Physics and Engineering Sciences program, a program mostly based on mathematics. In Janson de Sailly, he prepared national competitive exams from September 2010 to July 2012. After those competitions, he was admitted in September 2012 to the engineering school Ecole Centrale Marseille, Marseille, France. In Centrale he studied engineering sciences and was also introduced, among others, to finance, management, quantum physics, catalytic chemistry, robotics, biochemistry. During his studies at Centrale, Florian Dubost worked one month in the repair and maintenance center for medical devices in the General Hospital of Vienna (AKH Wien), Vienna, Austria. He was first introduced to machine learning and medical imaging in 2014 in Marseille when he worked six months jointly in the Laboratoire d'Informatique Fondamentale de Marseille and the Timone Neuroscience Institute, and developed machine learning classification models for the auditory cortex. In 2014, he was admitted to a double degree program with the Technical University of Munich (TUM), Munich, Germany, in the Medical Engineering master's program. Shortly after joining the TUM, he was admitted to a one-year mentoring program in Siemens. In 2016, he worked on his master's thesis in Prof. Nassir Navab's medical image analysis group at the TUM, and developed automated methods for the hands-free segmentation of medical volumes via binary inputs, which he published in a workshop of the international conference on Medical Image Computing and Computer Assisted Intervention, held in Athens, Greece. After defending his master's thesis on the 14th April 2016 at the TUM, Florian Dubost received his two master's degrees in Science in Engineering from Ecole Centrale Marseille and in Medical Engineering from the TUM. On the 18th April, he started a PhD in Erasmus MC, Rotterdam, the Netherlands, under the close supervision of Prof. Marleen de Bruijne and that of Prof. Wiro Niessen and Prof. Meike Vernooij. The initial focus of his PhD was the development of machine learning methods for the automated quantification of brain lesions from MRI scans. In 2018, during his PhD, he received a personal grant for a four months research visit in Prof Natalia Rost's group at the Massachusetts General Hospital, Harvard Medical School, Boston, USA. He researched automated methods for the discrimination of spatial patterns of brain lesions in the general population versus a stroke population. During his stay in Boston, Florian was invited to present his work on the automated quantification of brain lesions in Boston University, Harvard Medical School, and the Massachusetts Institute of Technology. After returning to Europe, he started multiple international collaborations for the deployment of his methods. In 2019, his visited the group of Dr. Stefen Oeltze-Jafra affiliated to the Neurology Department of Otto von Guericke University Magdeburg, Magdeburg, Germany. And until now he has also been collaborating with Prof. Sudha Seshadri from the Framingham study, Prof. Charles Guttmann from Harvard Medical School, and Dr. Saima Hilal from the National University of Singapore.