Detecting Perivascular Spaces: a Geodesic Deep Learning Approach

Kimberlin van Wijnen

Technische Universiteit Delft

TU Delft Delft University of Technology

Challenge the future

Detecting Perivascular Spaces: a Geodesic Deep Learning Approach

by

Kimberlin van Wijnen

in partial fulfillment of the requirements for the degree of

Master of Science
in Biomedical Engineering

at the Delft University of Technology,
to be defended publicly on Monday November 5, 2018 at 12:30 PM.

Thesis committee:

Prof. dr. Wiro Niessen,	Chairman	TU Delft, Erasmus MC
Dr. Marleen de Bruijne,	Supervisor	Erasmus MC, UCPH
Florian Dubost Msc.	Daily Supervisor	Erasmus MC
Dr. Frans Vos,	Committee member	TU Delft, AMC
Dr. Anna Vilanova,	Committee member	TU Delft
Dr. ir. Marius Staring,	Committee member	TU Delft, LUMC

TU Delft Delft University of Technology

Detecting Perivascular Spaces: a Geodesic Deep Learning Approach

Abstract

Perivascular spaces (PVS) visible on MRI are currently emerging as an important potential neuroimaging marker for several pathologies in the brain like Alzheimers disease and cerebral small vessel disease. PVS are fluid-filled spaces surrounding vessels as they enter the brain. Although PVS are normally not noticeable on MRI scans acquired at clinical field strengths, when these spaces increase in size they become increasingly visible and quantifiable. To study these spaces it is important to have a robust method for quantifying PVS.

Manual quantification of PVS is challenging, time-consuming and subject to observer bias due to the difficulty of distinguishing PVS from mimics and the large number of PVS that can occur in MRI scans. Many promising (semi-)automated methods have been proposed recently to decrease annotation time and intra- and inter-observer variability while providing more information about EPVS. However there are still various limitations in the current methods that need to be overcome. An important limitation is that most of the methods are based on elaborate preprocessing steps, feature extraction and heuristic fine-tuning of parameters, making the use of these methods on new datasets cumbersome. Furthermore the majority of the currently proposed methods have been evaluated on small sets of barely 30 images, as most of these methods aim to segment PVS and require voxel-wise annotations for evaluation.

In this thesis we propose a method for automated detection of perivascular spaces that combines a convolutional neural network and geodesic distance transform (GDT). We propose to use dot annotations instead of voxel-wise segmentations as this is less time-consuming than fully segmenting PVS while still providing the location of PVS. This enables us to use a considerably larger dataset with ground truth locations than is used in all previously proposed (semi-)automatic methods that provide the location of PVS. We investigated two approaches of using geodesic distance transform to optimize the CNN to detect PVS. The first approach focuses on optimizing the CNN for voxel-wise regression of the geodesic distance map (GDM) computed from the dots and the intensity image. The second approach aims to predict segmentations of the PVS using a CNN that is trained on approximated segmentations obtained by thresholding GDMs. We use 1202 proton density-weighted (PDw) MRI scans to develop our methods and 1000 other scans are used to evaluate the performance of the methods. We show that our methods match human intra-rater performance on detecting PVS without the need for any user interaction. Additionally we show that GDMs are extremely useful for capturing complex morphologies when computed from dot annotations. Our experiments indicate that GDMs can be used to provide valuable additional information to CNNs during training.

Keywords: Deep learning, perivascular spaces, detection, geodesic distance transform, dot annotations, weighted loss

1. Introduction

Magnetic resonance imaging (MRI) with its ability to provide images of the brain in vivo and non-invasively, is an invaluable technique for neuroscience. It has enabled a rapid increase in our understanding of neurophysiology and neuropathology (Adams et al., 2015; Annese, 2012). Advancing MRI quality has made it increasingly feasible to study smaller and more subtle structures in the brain and uncover their physiology as well as their pathology and association with neurological diseases (Groeschel et al., 2006; Adams et al., 2015).

An example of such structures is the perivascular space (PVS), also referred to as the Virchow-Robin space, which is currently emerging as an important potential neuroimaging marker for several pathologies in the brain like Alzheimers disease and cerebral small vessel disease (Wang et al., 2016b; Adams et al., 2015; Charidimou et al., 2013). Although perivascular spaces are normally not noticeable on MRI scans acquired at clinical field strengths, when these spaces increase in size they become increasingly visible and quantifiable (Ramirez et al., 2016; Kwee and Kwee, 2007; Wardlaw et al., 2013; Doubal et al., 2010).

There is still relatively little known about PVS and a lot of inconsistencies exist in the literature (Wuerfel et al., 2008; Kilsdonk et al., 2015; Wardlaw et al., 2013). Multiple reasons have been suggested for this. Firstly enlarged PVS were assumed to be benign for a long time (Zhu et al., 2010b; Adams et al., 2015). Additionally image quality has improved a lot over the years, where it was first hardly possible to visualize PVS that were enlarged, now it is even possible to discern very faintly enlarged PVS (Groeschel et al., 2006; Adams et al., 2015). However, distinguishing PVS from other structures visible on MRI scans (e.g. lacunar infarcts) remains difficult due to the similarity of PVS to these structures (Valdés Hernández et al., 2013; Kwee and Kwee, 2007; Potter et al., 2015). Furthermore, many studies based the definition of enlarged PVS on the visibility of PVS on MRI scans. As this is dependent on MRI sequence parameters this is not a very robust definition as well as quite arbitrary and not based on clinical significance (Valdés Hernández et al.,

2013; Wardlaw et al., 2013; Adams et al., 2015). Currently more studies specify the diameters of PVS that are examined, which improves the reliability of comparing studies. Lastly, studying these spaces and comparing different studies is also difficult because of the absence of a general, robust method that is not unreasonably time-consuming for assessment of PVS burden. Efforts are ongoing to improve this (Ikram et al., 2017; Adams et al., 2015, 2013; Wardlaw et al., 2013; Potter et al., 2015; Valdés Hernández et al., 2013).

Currently manual quantification of PVS is still the golden standard, however this is very time-consuming and challenging due to the difficulty of distinguishing PVS from mimics and the number of PVS that can be large. Many studies have suggested that computational methods could improve reliability, generalization and speed of PVS quantification (Valdés Hernández et al., 2013; Adams et al., 2015; Park et al., 2016; Boespflug et al., 2018; Dubost et al., 2018b). Various (semi-)automated methods have been proposed for different types of PVS quantification, such as PVS counting (Dubost et al., 2018b,a), PVS burden categorization (González-Castro et al., 2016b, 2017), or PVS segmentation (Lian et al., 2018; Park et al., 2016; Ballerini et al., 2018; Zhang et al., 2017; Boespflug et al., 2018). These promising methods clearly demonstrate the potential of computational methods for PVS quantification. However, currently proposed methods still have important limitations (see section 4.1).

An important limitation is that most of the methods are based on elaborate preprocessing steps, feature extraction and heuristic fine-tuning of parameters, making the use of these methods on new datasets cumbersome. To overcome this limitation we use a convolutional neural network (CNN). CNNs are taking over the field of medical image analysis by storm with their extraordinary improvements in performance and applicability. CNNs are a type of artificial intelligence that are especially useful and promising for image analysis. Instead of relying on manually designed features, these models learn their own customized features. By optimizing a specified error function referred to as the loss, CNNs adapt their parameters to improve their performance on training images with a given ground truth. The ground truth labels and the loss together specify the objective that the CNN tries to accomplish (Ronneberger et al., 2015; Zhou et al., 2016; Kamnitsas et al., 2017; Litjens et al., 2017; Long et al., 2017). Originally CNNs were used for image classification, predicting a single value representing the class based on the image that is given as input.

Fully convolutional neural networks (FCNs) introduced by Long et al. (2017) are CNNs that perform voxel-wise prediction, meaning their output has the same shape as the given input image and every voxel in the output contains the prediction of the network for the corresponding input voxel. FCNs are consequently aimed towards segmentation. Two of the most recently proposed methods for localizing PVS use CNNs (Dubost et al., 2017; Lian et al., 2018). Both use architectures based on FCN and show promising results, Dubost et al. (2017) on detection of PVS and Lian et al. (2018) on segmentation of PVS. However, both are evaluated on a small dataset. Furthermore, Dubost et al. (2017) focus on detection only in the basal ganglia and Lian et al. (2018) develop and evaluate on MRI scans of the brain at 7 Tesla (T) which is higher than the clinical field strength (1.5 or 3 T).

In this thesis we propose to detect PVS using a CNN that is trained and validated on a large set of 1202 MRI scans and tested on a separate set of 1000 images. This test set is considerably larger than the ones used by Dubost et al. (2017) and Lian et al. (2018) that contained 30 and 20 images, respectively. Furthermore, these MRI scans are acquired at 1.5 T which is also used in clinical practice. 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; Adams et al., 2015). Like the annotations used for testing by Dubost et al. (2017), our annotations are dots that indicate the location of the every PVS. Dot annotations indicate PVS locations, while being considerably less time-consuming than PVS manual segmentation.

Recent studies on cell detection have shown the potential of using CNNs to regress score maps based on distance transforms as a way to use dot annotations for detection (Xie et al., 2018a,b; Kainz et al., 2015). These methods use score maps based on euclidean distance, which does not take image intensity values into account. As the morphology of PVS helps to distinguish PVS from other structures, incorporating the intensity information is necessary (Valdés Hernández et al., 2013; Boespflug et al., 2018). Geodesic distance incorporates image context by combining spatial distance and intensity difference in the image. The intensity is seen as an extra dimension that defines the landscape. Intuitively distances between voxels that are connected by flat terrain are shorter than voxels that have hills and valleys in the space between them. As a consequence strong edges are emphasized in geodesic distance maps, making them very useful for segmentation of structures (Toivanen, 1996; Gaonkar et al., 2015; Wang et al., 2018; Criminisi et al., 2008). Geodesic distance maps show potential for PVS detection as they could incorporate their complex morphology (Park et al., 2016; Valdés Hernández et al., 2013; Boespflug et al., 2018).

In this thesis we address the hypothesis that dot annotations can be used to develop an automated method for PVS detection. This method should be as good as human performance and supply information on quantity and spatial distribution of PVS without need for user interaction. We investigate this hypothesis by developing two approaches to optimize a CNN using geodesic distance maps and comparing performance of these methods to the performance of an expert rater.

The following sections will further discuss PVS (2), geodesic distance transform (3) and current methods for PVS assessment (4.1) and similar methods using (geodesic) distance maps (4.2). Section 5 will present our contributions and in section 6 the data will be introduced, followed by the method in section 7 and the experimental setup in section 8. The results will be presented in section 9, discussed in section 10 and section 11 will conclude this thesis. The appendices contain further information on PVS, distance transforms, additional (exploratory) experiments that were done and appendix Appendix D contains supplementary results.

2. Perivascular Spaces

2.1. Anatomy and Physiology

PVS, also known as Virchow-Robin spaces, surround arteries, arterioles, veins and venules as they enter and emerge from the brain (see Figure 1). The brain is enveloped by three membranes, the dura mater, arachnoid mater and the pia mater. Between the arachnoid mater and the pia mater which covers the cerebral cortex lies the subarachnoid space. Vessels entering the brain from the subarachnoid space or emerging from the brain into the subarachnoid space are enveloped by pia mater. The spaces between the pia mater and vessels is referred to as PVS (Zhang et al., 1990; Braffman et al., 1988; Barkhof, 2004; Kwee and Kwee, 2007; Valdés Hernández et al., 2013). Controversy exists as to whether PVS are filled with cerebrospinal fluid (CSF) (Ramirez et al., 2016; Potter et al., 2015) and/or interstitial fluid (ISF) (Fanous and Midia, 2007; Öztürk and Aydingöz, 2002).

PVS are believed to be involved in the clearance of fluid and solutes in the brain as well as play an important role in immunological and inflammatory responses in the brain (Wang and Olbricht, 2011; Bacyinski et al., 2017; Faghih and Sharp, 2018; Ramirez et al., 2016; Cserr and Knopf, 1992; Esiri and Gay, 1990; Fanous and Midia, 2007; Zhang et al., 1990). The exact role of PVS is still unknown, research about this is still ongoing (Wang and Olbricht, 2011; Bacyinski et al., 2017; Valdés Hernández et al., 2013).

2.2. Pathology

Formerly the enlargement of PVS was assumed to be benign (Zhu et al., 2010b; Adams et al., 2015). Recent studies however support the contrary and an increasing amount of research is being done on this emerging neuroimaging marker. PVS have been associated with worse cognition, hypertension, as well as with markers of cerebral small vessel disease namely white matter hyperintensities and lacunar infarctions (Maclullich et al., 2004; Zhu et al., 2010a; Potter et al., 2015; Chen et al., 2011; Zhu et al., 2010b; Charidimou et al., 2013). PVS are seen in individuals of all ages and in the elderly population they are highly prevalent (Adams et al., 2015; Zhu et al., 2011, 2010b). With increasing age the number and size of PVS in the brain has been shown to increase (Doubal et al., 2010; Kwee and Kwee, 2007; Dubost et al., 2018b). Additionally a high number of PVS has been associated with many neurological conditions including cerebral small vessel disease, cerebral arteriosclerosis, traumatic brain injury, poststroke depression, Parkinson's disease, incident dementia and Alzheimer's disease (Zhu et al., 2010a; Maclullich et al., 2004; Zhu et al., 2011; Ramirez et al., 2016; Zhu et al., 2010b; Chen et al., 2011; Hurford et al., 2014; Potter et al., 2015; Liang et al., 2018; Cai et al., 2015; Doubal et al., 2010).

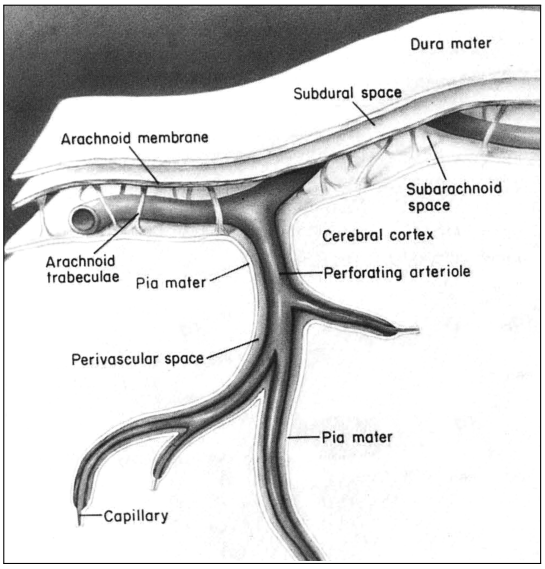


Figure 1: **Anatomy of perivascular spaces** Illustration of a perivascular space surrounding an arteriole penetrating the brain. This is a simplified representation (adapted from Heier et al. (1989))

The cause and mechanism of the enlarging of PVS are not clear yet. Various mechanisms have been proposed that may contribute to the enlargement of PVS, e.g. atrophy of the brain, microvascular or lymphatic obstruction, hypertension and inflammation (Chen et al., 2011; Adams et al., 2015; Groeschel et al., 2006).

2.3. Visualization

Normal PVS are too small to be noticed on MRI scans at clinical field strengths, however when PVS increase in size they become more visible and quantifiable (Ramirez et al., 2016; Kwee and Kwee, 2007; Doubal et al., 2010; Wardlaw et al., 2013). As there is no clear threshold in terms of diameter yet for when PVS are clinically enlarged, the term enlarged perivascular spaces was often used in studies to refer to any PVS visible on MRI scans. However the visibility of PVS on MRI is dependent on the MR sequence characteristics so this is very study dependent and arbitrary in terms of clinical significance (Valdés Hernández et al., 2013; Wardlaw et al., 2013; Adams et al., 2015). For this reason referring to all PVS (visible on MRI or not) as PVS and specifying the range of PVS diameters that is examined in the study is recommended now (Wardlaw et al., 2013; Adams et al., 2015, 2013).

As PVS follow the course of the vessel they surround, they appear as elongated structures on 3D MRI scans. In a 2D image slice PVS can be round, ovoid or linear dependent on what the orientation of the PVS is with respect to the image slice (Wardlaw et al., 2013). PVS with a diameter between 1 mm and 3 mm is an often used range for examining PVS that are enlarged. PVS with a smaller diameter than this are barely enlarged and larger PVS ($>$ 3 mm) might be dependent on different pathology. PVS have a similar intensity to CSF on all MR sequences (Ramirez et al., 2016; Kwee and Kwee, 2007).

The appearance of PVS on MRI scans bears most resemblance to lacunar infarcts, lacunes and small punctual white matter hyperintensities (WMH) (Potter et al., 2015; Valdés Hernández et al., 2013; Kwee and Kwee, 2007; Bokura et al., 1998). As sulci contain CSF, they can look similar to PVS and motion artifacts can look similar as well. Especially the shape, location, size, and spatial distribution are thought to be important descriptors for distinguishing PVS from mimics (Valdés Hernández et al., 2013; Dubost et al., 2018b; Chen et al., 2011;

Boespflug et al., 2018) (see Appendix B.4 for more information on this).

PVS appear mainly in the basal ganglia, hippocampus, centrum semiovale en mesencephalon. The CSO is seen as the most difficult region to distinguish PVS as it is the largest and contains many similar structures (see Figure 4b) (Kwee and Kwee, 2007; Barkhof, 2004; Adams et al., 2015; Dubost et al., 2018b).

2.4. Assessment

MRI is without a doubt an invaluable tool for research on PVS. However, to be able to study PVS associations, a way of measuring is needed to compare PVS burden in the brain. Until recently most studies used a visual scoring system, categorizing PVS burden in an image into in general 4 to 6 burden levels (Doubal et al., 2010; Maclullich et al., 2004; Hurford et al., 2014; Chen et al., 2011; Potter et al., 2015). Almost every study had its own method of assessing PVS burden, making it difficult to compare studies. Efforts have been made to establish a more general and robust way of evaluating burden of PVS (Ikram et al., 2017; Adams et al., 2015; Wardlaw et al., 2013; Doubal et al., 2010; Potter et al., 2015; Valdés Hernández et al., 2013; Ballerini et al., 2018).

Visual scoring systems are a fast way of PVS assessment. However, these scales are based on subjective classification. Furthermore, clustering the PVS burden into few categories results in floor and ceiling effects. Evidently these scales do not directly indicate any information on location, morphology or volume of PVS (Wang et al., 2016b; Ballerini et al., 2018; Boespflug et al., 2018; Ramirez et al., 2015). Established visual scoring systems that currently appear to be most used are the Patankar scale and the Potter scale (Patankar et al., 2005; Ramirez et al., 2015; Potter, 2011; Wang et al., 2016b; González-Castro et al., 2017).

Other proposed measures of PVS burden are counting the number of PVS per slice or per full brain region, voxel-wise binary labels with either a dot per PVS or segmentations of the PVS (in order of containing increasing information about PVS). Besides containing more information about PVS, these measures also pose a more objective way of assessing PVS burden. However, as PVS are difficult to distinguish from other structures like lacunes, these measures still do suffer from some subjectivity (Adams et al., 2015; Dubost et al., 2018b; Valdés Hernández et al., 2013; Wang et al., 2016b). An important disadvantage of these measures is that obtaining them manually is very time-consuming. Especially for a large brain region like the CSO this would take very long. However the number of PVS found in one slice of the CSO was shown to be highly correlated with the number of PVS in the whole brain region. This considerably decreases the annotation time (Adams et al., 2015).

3. Geodesic Distance Transform

A distance map is an image that shows for every pixel how far away it is from a chosen subset of pixels and was first presented by Rosenfeld and Pfaltz (1968) (Cárdenes et al., 2010; Saito and Toriwaki, 1994). A distance map is computed from a binary image using an operation called a distance transform (see Figure 3). The binary image distinguishes between pixels belonging to the background (0) and foreground (1). The distance transform calculates for every pixels the closest distance to the specified foreground. The output is an image with pixels that have a value corresponding to their distance to the chosen subset defined by the binary image. Essentially a distance map is a composition of distance isocontours, each contour containing all pixels that are a certain distance from the foreground (Paglieroni, 1992; Borgefors, 1986; Rosenfeld and Pfaltz, 1966; Grevera, 2007; Wang and Tan, 2013).

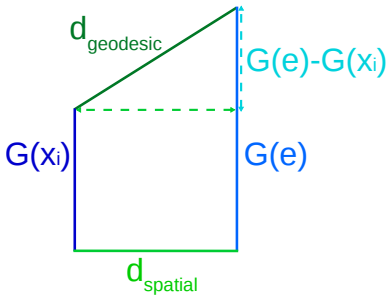
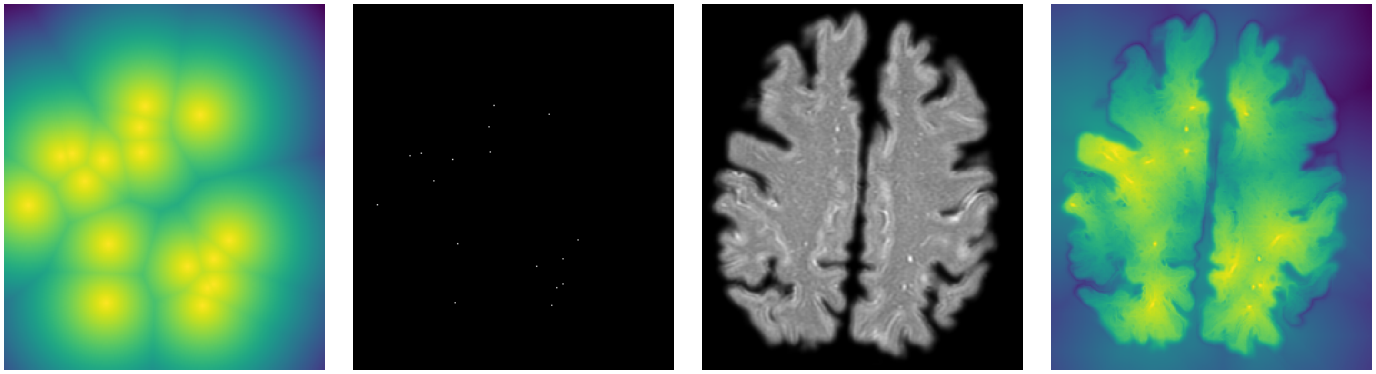


Figure 2: **Displacement on curved space.** To compute the geodesic distance map value for pixel 'e', the geodesic distance $d_{geodesic}$ to all neighboring pixels x_i is calculated. The intensity is perceived as an extra dimension and is defined in the gray-scale image G(x). The spatial distance $d_{spatial}$ for WDTOCS $d_{spatial}$ is 1 for horizontal and vertical neighbors ($N_4(e)$) and $\sqrt{2}$ for diagonal neighbors ($N_8(e) \setminus N_4(e)$). Using Pythagoras $d_{geodesic}$ is calculated from G(e) - G(x_i) and $d_{spatial}$. A 2D input image is assumed, but 3D input images are possible as well with a different connectivity grid (adapted from Toivanen (1996))

The definition of the distance in a distance transform greatly effects the resulting distance map. For gray-scale images taking the intensity into account besides the spatial information is useful to encode the image context into the distance map. Toivanen (1996) developed a measure for this called the weighted distance on curved space (WDOCS) also referred to as the geodesic distance. Combining the spatial and intensity information by using the intensity as an extra dimension, the image is seen as a curved space defined by the spatial coordinates and one intensity coordinate (see Figure 2). The shortest path on curved space is referred to as the geodesic distance, as the path is restricted to the top surface of this height map. Intuitively distances between pixels that are connected by flat terrain are shorter than pixels that have hills and valleys in the height map between them (Grazzini et al., 2007; Toivanen, 1996). The corresponding transform called the weighted distance transform on curved space (WDTOCS) requires a binary image F(x) defining the foreground as well as a gray-scale image G(x). The WDOCS between pixel e and its neighboring pixel $x_i \in (N_8(e)$, with intensities $G(e)$ and $G(x_i)$ respectively, is defined as

$$d(e, x_i) = \begin{cases} \sqrt{\left(G(e) - G(x_i)\right)^2 + 1} & \text{if } x_i \in N_4(e) \\ \sqrt{\left(G(e) - G(x_i)\right)^2 + 2} & \text{if } x_i \in (N_8(e) \setminus N_4(e)) \end{cases} \quad (1)$$

Geodesic distance maps are used a lot in medical image



(a) (b) (c) (d)

Figure 3: **Distance Transforms** Euclidean distance transform (EDT) computes a euclidean distance map (EDM) (a) from a binary image (b). The weighted distance transform on curved space, also referred to as geodesic distance transform, combines the spatial distance and the intensity differences between pixels to compute a geodesic distance map (d). It uses a binary image to define foreground and background (like EDT) and a gray-scale image (c) to define the 'curved space'. In this example the binary image are the dot annotations and the gray-scale image is an axial slice of a proton density-weighted (PDw) MRI scan. Both transformations were computed on 2D images with an assumed pixel connectivity of 8 (N_8, meaning horizontal, vertical and diagonal neighbors in a 2D grid) was used, but can be adapted to 3D and different connectivities.

analysis especially for segmentation as these maps show large differences in values at edges in the intensity image (Wang et al., 2018; Criminisi et al., 2008). Geodesic distance maps clearly show potential to use for incorporating important information from the intensity image with the dot annoations. As annotations are only available for one slice of the data (see section 6) this section assumed 2D images, however 3D images are also possible and different connectivity grids.

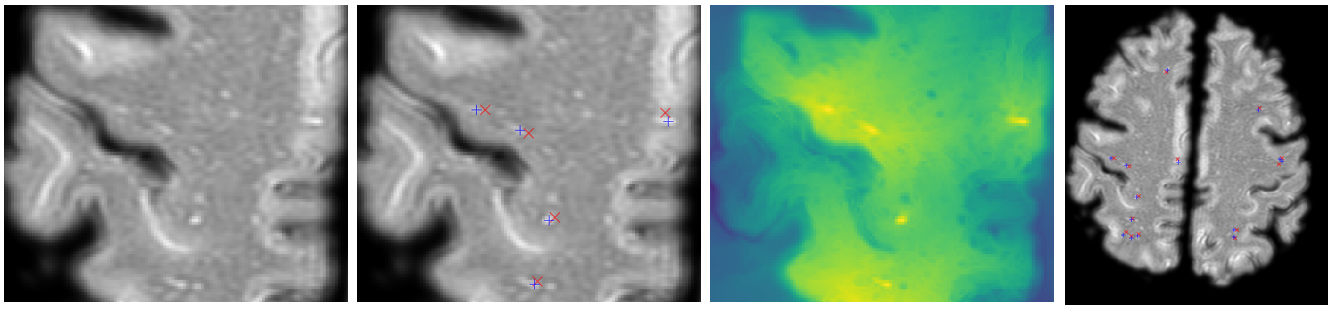
4. Related Work

4.1. Computational Methods for PVS burden

Several promising (semi-)automated methods have already been proposed to decrease annotation time and intra- and inter-observer variability while providing more information about PVS (Lian et al., 2018; Dubost et al., 2018b; Park et al., 2016; Boespflug et al., 2018; Ballerini et al., 2018).

However, current methods still suffer from at least one of the following issues. The proposed methods that are semi-automated still rely on some user interaction, making these methods inconvenient to use for large datasets as well as more susceptible to inter-observer variation (Wuerfel et al., 2008; Ramirez et al., 2015; Wang et al., 2016b). Furthermore, the proposed unsupervised methods depend on elaborate preprocessing steps and heuristic fine-tuning of parameters which hinders the use of these methods on new datasets (Wuerfel et al., 2008; Uchiyama et al., 2008). Some methods use patches instead of full images, reducing the (spatial) information the method gets (Lian et al., 2018; Jung et al., 2018). Moreover, several of the proposed methods require MR images acquired at higher field strengths than the current standard used in practice (1.5 T or 3 T), greatly limiting clinical applicability of these methods (Lian et al., 2018; Zhang et al., 2017; Ballerini et al., 2018). Additionally, the majority of the proposed algorithms is evaluated on a relatively small set namely less than 30 images due to requiring voxel-wise annotations for testing (and training) (Lian et al., 2018; Park et al., 2016; Boespflug et al., 2018). It would be very useful to evaluate these methods on larger datasets to evaluate their true potential and robustness. Besides this, most of the methods are developed and evaluated on data from the same hospital, acquired by the same scanner (Dubost et al., 2018b; Hou et al., 2017; Cai et al., 2015). Evaluating these methods on images from other hospitals and scanners would be beneficial to see if the methods generalize well. Similarly some methods are evaluated using annotations from only one observer (Dubost et al., 2018b; Ballerini et al., 2018). As it is difficult to identify PVS, annotations are subject to observer bias (Adams et al., 2015; González-Castro et al., 2016a; Hou et al., 2017). Comparing annotations computed by a method developed on annotations of one observer, with the annotations of another observer would be valuable to evaluate to what extent the method is overfit on annotations of one observer. In addition many papers do not mention any intra- or inter-observer correlation of the annotations they used (Park et al., 2016; Lian et al., 2018; Uchiyama et al., 2008). However, it is important to know the confidence of the ground truth to compare methods to human performance. Lastly, the amount of information about PVS provided by methods vary from binary classification of the general PVS density to segmentation of PVS. All are useful for research on PVS burden, however additional information given on PVS like location and morphology enables additional research possibilities on e.g. associations with spatial distribution of PVS or with size of PVS (Boespflug et al., 2018). Even though a lot of promising methods have been proposed already on PVS assessment, there is clearly still room for improvement.

In this thesis we use dot annotations as ground truth to develop and evaluate our proposed methods. As this is less time-consuming than fully segmenting PVS this enables us to use a considerably larger dataset with ground truth locations than is used in all previously proposed (semi-)automatic methods that provide the location of PVS. We use 1202 images to develop our methods and 1000 images are used to evaluate the performance of the methods.



(a) (b) (c) (d)

Figure 4: **Shifting dots inside the volume of their corresponding enlarged perivascular space (PVS)** Zoom in of axial slice of intensity image without annotated dots (a), with annotations (indicated by red Xs) and shifted annotations to the PVS center as described in section 7.4 (indicated by blue +s) (b), the geodesic distance map computed from shifted annotations described in the same section (c), and the intensity image of one full axial slice with annotations (indicated by red Xs) and shifted annotations (indicated by blue +s) (d).

4.2. Methods using Distance Transform

Several methods have been proposed using spatial maps to optimize neural networks for the detection of cells. These methods use maps that are based on Euclidean distance. For these methods this makes sense as the cells they aim to detect are mainly circular (Xie et al., 2018b,a; Raza et al., 2018). Furthermore, these methods use 2D images causing problems with occlusion that are tackled using e.g. a density surface (Xie et al., 2018a).

PVS on the other hand are elongated structures and have a complex morphology (Park et al., 2016; Valdés Hernández et al., 2013; Boespflug et al., 2018). As Euclidean distance does not take image context into account, using a different distance measure would make sense. Furthermore, the data we use in our study is 3D. Without the problem of occlusion, a density surface is less advantageous. However, geodesic distance combines spatial distance and intensity differences (see section 3) and show potential for PVS detection as they could incorporate the complex morphology of PVS.

Geodesic distance maps are used in medical image analysis in a lot of different ways. A lot of promising methods have already been proposed using GDMs, like a method proposed by Gaonkar et al. (2015) that uses geodesic distance maps combined with thresholding for tumor volume segmentation. Furthermore, (Jang et al., 2016) use a combination of Hough transform and geodesic distance maps for aorta segmentation. Kontschieder et al. (2013) tackle semantic segmentation with a forest-based model using geodesic distance to compute connectivity features. Moreover, Krähenbühl and Koltun (2014) present a method that produces objects proposals in images using critical level sets in geodesic distance maps computed using seeds placed by trained classifiers. Another interesting approach is a method presented this year by Wang et al. (2018) that combines geodesic distance transformation and deep learning for interactive segmentation. User interactions are encoded as geodesic distance maps and together with the intensity image and the current segmentation these maps are given to a CNN that outputs a refined segmentation. The GDMs provide a useful way to provide the network with extra spatial and intensity information about the foreground or background that can be derived from the user input.

In our proposed methods we also use geodesic distance transformation in combination with a CNN. However, we use the GDT for creating label images for optimization of the network, instead of using the GDM as input to the network. We train the CNN to either predict the geodesic distance map computed from the dot annotations or a thresholded version of it. Park et al. (2016) mention as a recommendation for enlarged PVS segmentation that geodesic distance might be useful for capturing the complex patterns of enlarged PVS. In our experiments we show that this is the case.

5. Contributions

We present two approaches for optimizing a CNN for automated detection of PVS. Both approaches use a geodesic distance transform to extract the complex morphology of PVS from the MRI scan and its corresponding dot annotations. The first approach focuses on optimizing the CNN for voxel-wise regression of the geodesic distance map computed from the dots and the intensity image. The second approach aims to predict segmentations of the PVS using a CNN that is trained on segmentations approximated using GDMs. Both approaches provide more information to the network than just the dot annotations would have. Various label images and losses are presented per approach to optimize the CNN to automatically detect PVS as accurately as possible. Both approaches brought forth methods that match human performance on detecting PVS without the need for any user interaction. To the best of our knowledge we are the first to compare and match human performance on the detection of PVS.

A substantial dataset of 1202 MRI scans was used to develop the methods. The methods were tested on a set of 1000 MRI scans, which is considerably larger than any other papers on automated methods for PVS quantification have reported as far as we know.

Recently many methods use MRI scans acquired at 7 T which has a better spatial resolution (Lian et al., 2018; Park et al., 2016; Zhang et al., 2016), raising the question how these models

perform on scans obtained at clinical field strength (Ballerini et al., 2018; Boespflug et al., 2018). All 2202 scans used in this study have been acquired at a clinical field strength of 1.5 T.

Lastly, even though geodesic distance transformation itself has been used repeatedly in medical image analysis applications, hardly any methods use GDMs in combination with neural networks. We show that GDMs are extremely useful for capturing complex morphologies when computed from dot annotations. Our experiments indicate that GDMs can be used to provide valuable additional information to CNNs during training.

6. Data

6.1. MRI scans

For this study we used 2202 proton density-weighted (PDw) MRI scans from the third cohort of the Rotterdam Study (Ikram et al., 2017). This is a prospective cohort study that investigates diseases in the elderly population in Rotterdam (Ikram et al., 2017). All scans were from different individuals and were acquired on a 1.5 T MRI scanner (General Electric Healthcare, Milwaukee, USA) using an 8-channel head coil. A fast spin echo sequence was used to obtain scans with a slice thickness of 1.6 mm (echo time (TE) = 17.3 ms, repetition time (TR) = 12,300 ms, flip angle = 90-180°, bandwidth = 17.86 KHz, field of view = 25 cm^2, matrix size = 416×256, scan time = 369 seconds). After reconstruction 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 our data is discussed by (Ikram et al., 2015)

Preprocessing of the images was performed as described by (Dubost et al., 2018b). For every MRI scan the CSO was segmented with the FreeSurfer multi-atlas segmentation algorithm (Desikan et al., 2006) producing a binary mask. The scans were subsequently cropped around the center of mass of the mask to reduce the size of the images and with that the memory requirements. As only one slice of the CSO was annotated, we further reduce the size of the scans by selecting only the slices surrounding the annotated slice. This slice is located 1 centimeter above the lateral ventricles as described by Adams et al. (2013). The slice number is estimated automatically by segmenting the lateral verntricles with the FreeSurfer algorithm and choosing the slice located 1 centimeter above the ventricles. The resulting scans with their size of $256 \times 292 \times 16$ fit in the memory of our GPU and contain (part of) the segmented CSO surrounded by zeros. An example slice is shown in Figure 3c.

6.2. Annotations

PVS have been annotated for 2202 MRI scans. Only one axial slice per scan was annotated, as the number of PVS in one slice is highly correlated with the number of PVS in the whole CSO (Adams et al., 2015) and annotating the whole CSO would be much more time-consuming. This slice was defined by Adams et al. (2013) as the slice 1 cm above the lateral ventricles. The location of the annotated slice in the final 3D image that is used for the automated methods is dependent on the FreeSurfer segmentation of the lateral ventricles and the brain region.

Sensitivity	**FPPI**
0.560 (± 0.300)	4.54 (± 3.69)
0.553 (± 0.289)	4.32 (± 3.70)

Table 1: **Annotation variability.** A separate set of 40 MRI scans was annotated twice by the same expert rater in a different random order with two weeks in between. For both time points the sensitivity and average amount of false positives per scan (FPPI) was computed by using the other set of annotations as the ground truth measure. The standard deviation across images is shown in brackets (Maybe not needed because the two points are also plotted in the FROCs)

An expert rater was provided with (full brain) T1w, T2w and FLAIR MR images. The rater selected this slice to annotate defined by Adams et al. (2013) and marked every PVS visible in that slice with a dot near its center (shown as red crosses in Figure 4b for more clarity). The guidelines about PVS discussed by Adams et al. (2013) were used for assessing PVS. PVS are defined to be at least 1 millimeter (mm) in diameter and the maximum diameter is 3 mm, because the pathogenesis of larger PVS might be different.

The variability in the annotations was evaluated on a separate set of 40 MRI scans. The expert rater annotated these scans two times in a different random order with two weeks in between. The sensitivity and average amount of false positives per scan (FPPI) was computed for both sets of annotations by using the other set of annotations as the ground truth measure (see Table 1).

7. Methodology

We propose a method for PVS detection using geodesic distance transform and a convolutional neural network (CNN). This method is optimized using dot annotations. If the dot annotations were directly used as labels, the network would be required to learn the exact voxel chosen by the annotator. Besides the complications this would cause due to the severe class imbalance and contradictory information about exact location, this is highly sensitive to annotation error or bias (see Figure 4b). Another approach is to consider the complete PVS volume as positive. Using the segmentation of the PVS would be ideal to incorporate the prior knowledge that PVS are elongated structures (see section 2). As manual segmentation is very time consuming, we use geodesic distance maps computed from the dot annotations instead. Geodesic distance maps (GDMs) combine spatial distance with intensity differences, they have been used for segmentation before, especially because of the emphasis that is given on sharp edges (Criminisi et al., 2008; Gaonkar et al., 2015; Wang et al., 2018). We investigated two approaches of using the GDM to optimize the CNN to detect PVS. The first approach is using the CNN for voxel-wise regression of geodesic distances to the nearest PVS which is discussed in section 7.2. Secondly we use the CNN for voxel-wise segmentation using a thresholded GDM as approximations of the segmentation of PVS as described in section 7.3. Besides what is described in these sections the rest of the method is the same for both approaches.As we only have

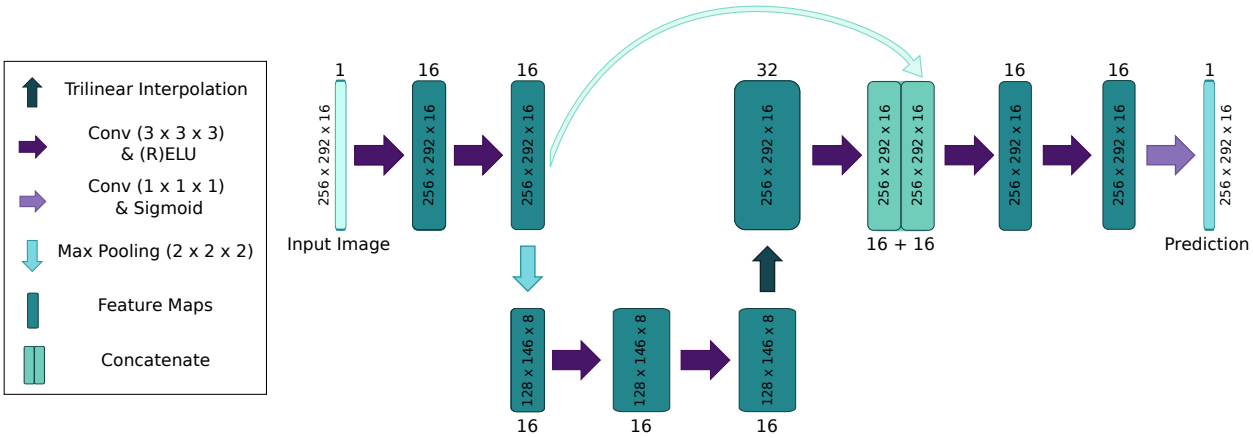


Figure 5: **Architecture of the CNN used for detection.** Preprocessed images of the centrum semiovale are given to the CNN as input images, after which the CNN outputs a prediction map. The CNN consists of seven convolutional layers with kernels of 3 x 3 x 3 and either a ReLU activation or an ELU activation, one max pooling layer, one trilinear interpolation layer, one concatenation and one convolution of 1 x 1 x 1 and a sigmoid activation function.

annotations of one specific axial slice of the CSO we only evaluate the loss on this slice of the volume. The slice 1 cm above the lateral ventricles was chosen for this, Adams et al. (2015) show that this slice is highly correlated with the total number of PVS in the whole CSO. The location of the annotated slice in the volume is different per image, forcing the network to learn to detect PVS in not only one slice but in at least a subset of the slices and possibly even the whole volume.

7.1. Architecture

The architecture of our CNN is inspired by U-Net and FCN (Ronneberger et al., 2015; Long et al., 2017). Our CNN can accept an image of arbitrary size, because it is a fully convolutional neural network with a convolutional layer with a kernel of $1 \times 1 \times 1$ instead of a fully connected layer (see Figure 5). The network consists of first two convolutional layers, a maxpooling layer, followed by again two convolutional layers, a trilinear upsampling layer, another convolutional layer, after which it is concatenated with the output of the second convolutional layer, two more convolutional layers and finally the last convolutional layer with a kernel of size $1 \times 1 \times 1$. The rest of the convolutional layers have kernels with sizes $3 \times 3 \times 3$ and the maxpooling layer has a kernel size of $2 \times 2 \times 2$. The number of feature maps for the convolutional layers are 16 for the full spatial resolution and 32 for the downsampled resolution and padding is applied so the layers output the same sized images as they got as input. For all convolutional layers except the last layer the generally used rectified linear units (ReLUs) are used as activation function to provide non-linear transformations for the regression approach (section 7.2). For the segmentation approach (section 7.3) exponential linear units (ELUs) are used (Clevert et al., 2015).

ELUs were proposed by Clevert et al. (2015) and are said to improve the learning speed of networks and the generalization performance.

The last layer has a sigmoid activation function which scales the output values between 0 and 1, as the label images are also between 0 and 1. Weights for the convolutional layers were initialized by random sampling from a normal distribution with zero mean and unit variance. Biases are initialized at a value of 0.01. Previous exploratory experiments on a subset of the training data indicated sigmoid output worked better than linear output, that having less feature maps slightly harmed performance and that adding more feature maps had no effect on performance. Linear interpolation for upsampling gave better results than transposed convolution and the standard method for upsampling in keras which is just repeating the neighboring voxels. Furthermore, deepening the network did not improve the performance.

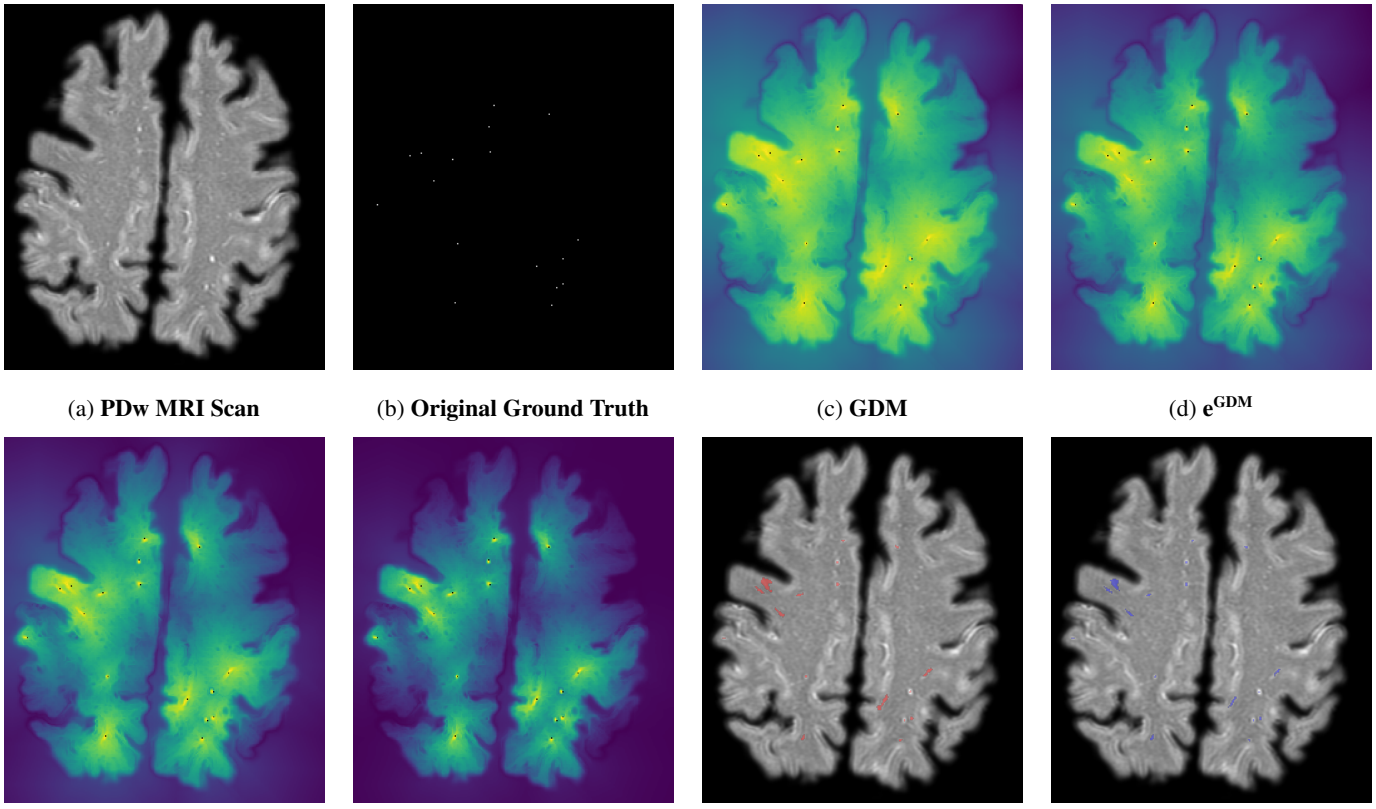
7.2. Regression approach

In this approach we aim to optimize the CNN for voxel-wise regression of the GDM and with this detection of PVS. Optimizing a CNN to predict this geodesic distance map is not straightforward. For this reason we compare how well various loss functions combined with several different label images optimize performance of PVS detection. In line with the work by Xie et al. (2018a) we use mean squared error as the basic loss function and formulated two additional loss functions.

As the loss is only evaluated on one slice, a 2D GDM is computed of this slice and its corresponding annotations and used as label image. We also experiment with several adaptations performed voxel-wise to emphasize the edges more in the GDM and create more contrast, by taking the exponential of the GDM, by squaring the values and by taking the GDM to the power 3. All GDM versions are normalized and inverted, resulting in a label image between 0 (farthest voxel from PVS) and 1 (PVS). An example of the four different GDMs used in this thesis are shown in Figure 6c - 6f.

7.2.1. Loss Functions

The goal of optimizing the CNN to predict the geodesic distance map is to handle class imbalance and at evaluation time providing the network with better feedback with which voxels



(a) **PDw MRI Scan** (b) **Original Ground Truth** (c) **GDM** (d) $\mathbf{e^{GDM}}$

(e) $\mathbf{GDM^2}$ (f) $\mathbf{GDM^3}$ (g) **Threshold** = 0.95 (h) **Threshold** = 0.96

Figure 6: **Label images for CNN** An axial slice of the proton density-weighted (PDw) MRI scan (a) and its corresponding dot annotations (b) indicating the location of the PVS in the image. Geodesic distance transform uses these images to compute a geodesic distance map (GDM) (c). To further emphasize the edges surrounding the PVS we modified the GDM by taking the exponential GDM (d), squaring all values in the GDM (e) and by taking all values to the power 3 (f). The increase in contrast and noticeably of PVS is clear to see between the different GDMs. The original dot annotations (b) are shown in black on the GDM images (c-f). Note that this is is an overlay and not part of the label images.Thresholding the original GDM at 0.95 and overlaying this on the intensity image (g) shows the approximate segmentation that is obtained. The segmentations obtained by a threshold of 0.96 are slightly smaller (h) which improves segmentation in some cases and decreases it in others (visually assessed). All these label images are scaled between 0 and 1.

are part of the PVS or close to it. However, the overall objective is to detect the PVS and this is not directly optimized by the loss. In former, exploratory experiments we observed that when using the widely used mean squared error (MSE) loss to optimize the network, there was not a clear improvement of performance in line with the decreasing loss. It seemed like the network was focusing too much on getting the exact background values right. We therefore propose three different ways to improve the MSE by weighting the loss. We compare the ability of these losses to optimize the network combined with the four described label images.

Mean Squared Error. The mean squared error quantifies the difference between the prediction (computed by the CNN) and the ground truth and can be used as a loss function. For every voxel i in the ground truth slice the squared difference between the predicted value $\hat{y}_i$ and the ground truth y_i is computed. To obtain the resulting MSE the mean is taken over the squared distances of all n voxels in this slice.

$$\text{MSE} = \frac{1}{n}\sum_{i=i}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

Label-Weighted Mean Squared Error. To place the focus of the optimization on the most important labels, namely those close to the PVS, the MSE can be weighted by multiplying with the corresponding value in the label image. The voxels in the PVS are approximately 1 and the voxel farthest away in geodesic distance from all PVS is 0. Multiplying this voxel-wise forces the loss to be higher for more important voxels close or in the PVS and voxels further away count less. The label-weighted mean squared error (wMSE) is defined as follows

$$\text{wMSE} = \frac{1}{n}\sum_{i=i}^{n} y_i\,(\hat{y}_i - y_i)^2 \tag{3}$$

Thresholded Label-Weighted Mean Squared Error. Combining advantageous of DSC and wMSE seemed like the next step to further focus the loss on optimizing the overall goal namely detection of PVS. The idea behind this loss is to only optimize for the background if it is "false positive" and otherwise ignore it. This is inspired by DSC as this also does not take into account background values unless they are false positives.

In essence TwMSE is a masked wMSE in a way. The mask defined by the threshold T indicates which voxels of the label

image are important to predict exactly and for which voxels predicting them as any value below the threshold is sufficient. The thresholded label-weighted mean squared error (TwMSE) is defined in the following way

$$\text{TwMSE}_\text{T} = \begin{cases} 0 & \text{if } p \leq \text{T and } y \leq \text{T} \\ \frac{1}{n}\sum_{i=i}^{n} y_i\,(\hat{y}_i - y_i)^2 & \text{otherwise} \end{cases} \tag{4}$$

where T is the chosen threshold. The voxels that are below the threshold in the label slice and in the predicted slice are seen as true negatives and are excluded from the loss (first case in Equation 4). Optimizing these values further is not necessary, as long as these voxels are below the threshold they are seen as unimportant. The voxels in the label slice that are higher than the threshold will always be taken into account as these are important to predict exactly (second case in Equation 4).

Two thresholds are applied for the TwMSE loss, one at 0.5 ($\text{TwMSE}_{\text{T=0.5}}$) and a threshold at 0.8 ($\text{TwMSE}_{\text{T=0.8}}$). The modified GDMs would have a different mask than the other GDMs if the same threshold would be taken for these images without taking into account their modification. For this reason the corresponding thresholds are computed by applying the same modification as for the type of GDM, e.g. T^2 for GDM^2

7.3. Segmentation Approach

Binary cross entropy (BCE) and the Dice similarity coefficient (DSC) were used as voxel-wise segmentation losses in varying combinations with label images. The GDMs are used to segment the PVS by thresholding at two different thresholds producing an approximate segmentation of the enlarged PVS (see Figure 6g and 6g). These approximated segmentations can provide the CNNs with valuable inforation on morphology and spatial distribution of PVS. To

7.3.1. Loss Functions

For the segmentation approach Binary cross entropy (BCE) and the Dice similarity coefficient (DSC) were used, which are both used a lot for optimizing CNNs for medical image segmentation.

Dice Similarity Coefficient. DSC quantifies the amount of agreement between two sets of labels. DSC loss focuses on the true positives and is often used when datasets are highly class imbalanced, which explains why it is frequently used for segmentation in which this is regularly the case (Andrews and Hamarneh, 2015; Trebeschi et al., 2017; Pinto et al., 2016; Sudre et al., 2017). The standard definition for DSC is

$$\text{DSC}(P, Y) = \frac{2|P \cap Y|}{|P| + |Y|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

with $\text{DSC}(P, Y) = 0$ for completely disjoint sets, and $\text{DSC}(P, Y) = 1$ for completely identical sets. Evidently with increasing performance DSC increases as well. However, per definition loss is minimized by neural networks so the DSC has to be flipped. This is in general simply done by defining the DSC loss as negative DSC or by using 1 minus the DSC. As the standard definition of DSC is not differentiable a continuous version of DSC is generally used instead (Shen et al., 2017; Zhou et al., 2016; Andrews and Hamarneh, 2015; Pinto et al., 2016; Sudre et al., 2017). The following DSC loss was used in our experiments

$$\text{DSC} = 1 - \frac{2 * \sum_{i=1}^{n} \hat{y}_i * y_i}{\sum_{i=1}^{n} \hat{y}_i + \sum_{i=1}^{n} y_i} \tag{6}$$

Binary Cross Entropy. BCE is an often used loss in deep learning. It quantifies the difference between the distribution of the data and the distribution of the predictions made by the network. The BCE is defined as

$$\text{BCE} = \sum_{i=i}^{n} \Big(- y_i * log(\hat{y}_i) - (1 - y_i) * log(1 - \hat{y}_i) \Big) \tag{7}$$

As the log of zero is not defined, in the Keras implementation a very small value ϵ is added in both log functions to ensure stability.

7.4. Implementation

The GDM is very dependent on the specified foreground. If the annotated dots are not placed in the volume of the PVS but next to it, the voxels of the corresponding PVS will have very high values (low in the inverted distance map). There will be a large intensity difference between the value at the dot annotation and the intensity of the PVS, which corresponds to a large geodesic distance. In the corresponding normalized and inverted GDM the voxels in the PVS will have low values while the surroundings containing the annotated dot will have high values which would indicate that region corresponds to a PVS. This would give very inconsistent information to the CNN. For this reason, to compute the GDM, dots have to be located in the volume of the PVS. For our dot annotations this was occasionally not the case (see Figure 4b).

To solve this problem, we developed an algorithm to shift all dots inside the volume of their corresponding PVS. In general the PVS' have a relatively high intensity compared to their surroundings in the image and the dots are in general close to the PVS they belong to (see Figure 4). Following these two observations, we shift the dots to the highest intensity value in a cube centered around the dot annotation. The size of the cube is a trade-off between making it large enough to shift to the corresponding PVS and small enough to prevent the dot from shifting to a different PVS. To further prevent the latter, the dots were restricted to only be able to shift to voxels in the same connected component with a connectivity of 8 (N_8, meaning horizontal, vertical and diagonal neighbors in a 2D grid). By visual examination of several images the size of this square was set to 7×7 voxels centered around the dot. The shifted dots were only used to compute a geodesic distance map for the training and validation set. For testing the original ground truth dots were used.

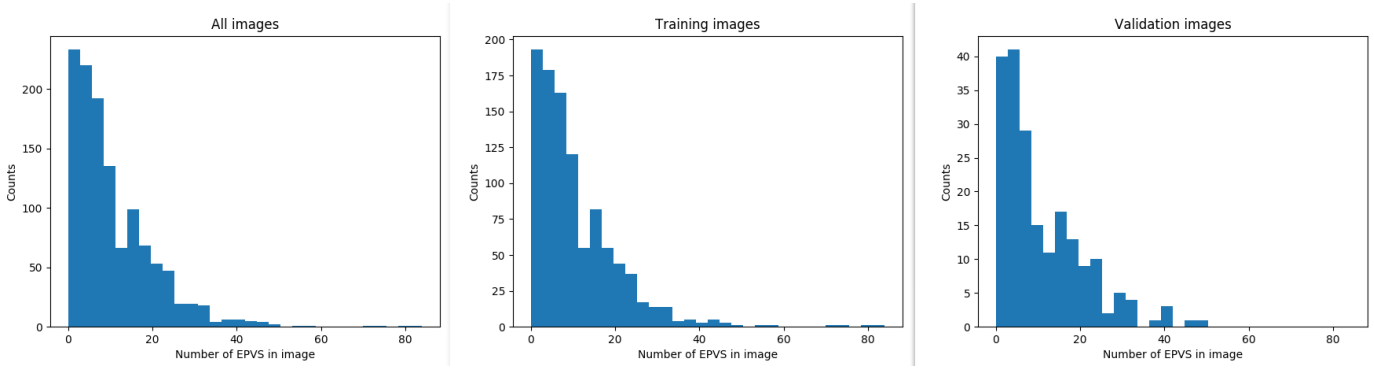


Figure 7: **Histograms of the spread in number of PVS in training and validation set.** The spread in number of PVS over the full set (1202 scans) used for developing the method (a), over the training set (1000 scans) (b) and over the validation set (202 scans) (c)

The geodesic distance maps are computed from the shifted dot annotations by using the geodesic distance transform proposed by Toivanen (1996) also referred to as the WDTOCS (see section 3). We used the same method as in the original paper, which implemented the method by adapting the raster scan method proposed by Rosenfeld and Pfaltz (1966). This algorithm (or a slightly adapted version of it) is often used in methods for medical image analysis (Wang et al., 2018; Criminisi et al., 2008; Kontschieder et al., 2013; Jang et al., 2016; Wei et al., 2012; Zhang et al., 2015; Cerrolaza et al., 2017). Various papers state the method is efficient with an optimal complexitys of O(N), relatively accurate and straightforward to implement (Zhang et al., 2015; Wei et al., 2012). Wang et al. (2018) propose a method for interactive segmentation using this algorithm, clearly showing the algorithm is quite fast.

Raster scanning is an iterative algorithm consisting of two passes over the image per iteration. The minimal geodesic distance per pixel is updated sequentially by passing a mask operation over the image first from the left upper corner to the lower right corner (forward pass using the kernel in Figure A.12b) and the second pass goes over the image in reversed order (backward pass using the kernel in Figure A.12c). All background pixels are initiated with the maximal integer number and the foreground pixels with 0. As the mask moves over the image, the new distance value $F^*(e)$ for pixel e is computed using the neighbors defined in the mask. This is done by first calculating the distance $d(e, x_i)$ (as defined in equation 1) for all neighbors x_i in the mask. Per neighboring pixel this distance $d(e, x_i)$ is added to the already calculated value for this pixel $F^*(x_i)$. The new distance value $F^*(e)$ for pixel e is the minimum value of these calculated distances and the initial value of pixel e $F(e)$

$$F^*(e) = min(F(e), min\{d(e, x_i) + F^*(x_i) \mid i \in mask\}) \quad (8)$$

After computing the geodesic distance transform the resulting GDMs were normalized and inverted as this was convenient for implementation in the neural network pipeline.

We implemented this in Python inspired by the C++ code on GitHub from Wang et al. (2018).

7.5. Model Training

The network is optimized using Adadelta (Zeiler, 2012) to minimize the loss on the training set. Due to memory limits we use a mini-batch of one, so after every training sample the loss is computed and the network is adapted to minimize the loss for the training sample. As this is prone to overfit it is standard practice to monitor the training with a validation set, which is a set of images separate from the training set. This way the performance of the model on new data can be evaluated. The loss on the training set normally either decreases or plateaus as this is the measure that is being optimized. The validation loss generally decreases during training, might plateau for a bit and when the model starts to overfit on patterns specific for the training set the loss on the validation set typically increases again. Early stopping is a general method used for regularization of the network. The ideal timing for stopping the optimized network is just before the overfitting starts. We generally waited for the model to overfit somewhat to be sure we were not looking at a local minimum instead of the global minimum, and stopped training after that. As we save the best model based on the validation loss we can then select the best model before overfitting.

Another method for regularization is to use data augmentation. By shifting, rotating and flipping the data more examples are generated for the network to learn from. As only annnotations for one axial slice are available, we do not do any of the affine transformation with respect to the depth dimension. For the validation set no augmentation is used. We use on-the-fly augmentation for the training set which means we apply random transformations to the training images every epoch. For every image a random shift between -4 and 4 voxels in horizontal and a shift in vertical direction is used, combined with a random rotation around the depth direction with a maximum of 20° either way, and random flipping in horizontal or vertical direction.

As in our approach (and many other approaches) the loss does not directly focus on detection of PVS, the optimization of the network was also monitored in other ways. Every 10 epochs the predictions of the current network on a few images of the validation set were saved to visually examine if the network was improving its ability to detect PVS. To further monitor the training of the models every 30 epochs the best model was saved (based on the validation loss). All models were subsequently

evaluated on performance of detecting PVS on the validation set. This was useful for evaluating if the best performance on the detection task was also at the best performance of the loss.

7.6. Post-Processing

After optimizing the network to output a GDM or segmentations of PVS, a few steps are followed to obtain the final detections proposed by the method. The 3D intensity images are given as input to the trained model, resulting in a prediction of the GDM or segmentation. The output of the network is a 3D volume, however only for one slice annotations are available (as mentioned in section 6). Only the slice for which annotations are available is used for evaluating performance. Non-maximum suppression is applied to the predicted slice to decrease the amount of false positive detections. This is implemented by applying a 5×5 maximum filter to the predicted slice with a connectivity of 8 (N_8, meaning horizontal, vertical and diagonal neighbors in a 2D grid). The voxels that have the same value in the filtered slice as in the original predicted slice are the local maximums which are referred to as the proposed detections of the network. These detections are ordered by their value, which is assumed to be an indication of the certainty of the network that this is a detection or not and will be referred to as the certainty value.The amount of detections proposed by the network depends on the threshold that is chosen for the certainty value, only detections with a higher certainty value are accepted as detections. Section 8.2 describes how the varying of this threshold is used to evaluate performance of a method.

8. Experiments

8.1. Experimental Settings

Implementation of the methods was done in Python and Keras (Chollet and Others, 2015) with Tensorflow as backend (Abadi et al., 2015). Code provided by Dubost et al. (2017) was used and adapted for the current methods. The exploratory and currently presented experiments were run on Nvidia Tesla K40 GPUs that are available for research purposes at the Dutch national cluster Cartesius and on an Nvidia Geforce GTX 1080 GPU and on 3 different Nvidia Geforce GTX 1070Ti GPUs. An experiment with the current CNN architecture with added batch normalization layers and an experiment with twice as many feature maps per convolutional layer was run on an NVIDIA QUADRO P6000 as this GPU has a higher memory capacity.

Stratified random sampling based on the number of PVS was used to split the 2202 scans into a set of 1202 for development of the method and a separate set of 1000 for testing that we did not use until after the whole development phase had ended.

The set of 1202 scans was split into 1000 training scans and 202 validation scans (distribution of number of PVS per image shown in Figure 7). Dubost et al. (2017) further improved and extended their proposed method that is optimized using weakly supervised labels (the number of PVS in a slice) and detects PVS. The same training and validation set were used by them to further improve and extend their proposed approach of detecting PVS with weakly supervised labels namely the number of PVS in a slice. The test set was used to evaluate and compare the different experiments on as well as to compare to the performance of the weakly supervised methods of Dubost et al. (2017) that they developed further.

A subset of the training set was first used for exploratory experiments to set up an initial pipeline and to investigate which loss functions improved detection performance most. The best performing pipelines on this subset were further developed on a larger training set with additional experiments for verification.

The validation set was used to monitor the validation loss and detection performance during training, to stop the training when the network was overfitting on the training set. The methods reached their optimal performance on PVS detection on average after around 200 epochs.

8.2. Evaluation

The objective of our methods is to detect PVS. We evaluate this by comparing with the annotations provided by the expert rater. The maximum diameter of PVS is defined as 3 mm for our annotations (see section 6). As the voxel-resolution is 0.49×0.49 in the annotated axial slice, this corresponds to a distance of 6 voxels in the annotated slice. Therefore we set the maximum distance for a correct detection to 6 voxels. Using the hungarian algorithm we match the detections proposed by the methods with the expert rater's annotations (Kuhn, 1955).

8.2.1. Free-Response Operating Characteristic

The main evaluation of the detection performance is done by computing the Free-Response Operating Characteristic (FROC) curve and its area under the curve (FAUC). This enables evaluation and comparison of the methods at varying thresholds. The methods output as indicated in section 7.6 a list of proposals ordered by highest certainty of being a detection. To compute the FROC curve for every method we decrease the threshold value from 1.0 (the highest value in the prediction map) in steps of 0.005 and either stop at 0.2 or after more than 500 detections are proposed. Based on the spread of counts seen in the training distribution (see Figure 7) the chance that there are more than 500 PVS in the image is not realistic.

For every threshold we evaluate the corresponding proposed set of detections for every image as discussed with the hungarian algorithm resulting in a number of true positives (TP, the proposed detection and the annotation match), false positives (FP, a proposed detection that matches no annotated dot) and false negatives (FN, an annotated dot that has no proposed detection) per image. The sensitivity per image is calculated by dividing the amount of TPs by the amount of total annotated dots in the image. The mean sensitivity is computed by taking the mean over the sensitivities of all images at that threshold. The average amount of FPs per image (FPPI) is computed by taking the mean over all FPs in all images. When there is no PVS annotated in the image the sensitivity is not defined and the image is not counted in the average sensitivity. The amount of FPs on the other hand is defined, because there can either be proposed detections which are all FP or there are zero FPs in the image when there are no proposed detections. Both give

important information on performance, therefore the amount of FPs is incorporated into the FPPI even if no sensitivity is defined. For every threshold a point of the FROC curve is calculated in this way defined by the FPPI and the average sensitivity and eventually the FROC is plotted for every method.

8.2.2. FAUC

The FAUC was obtained by calculating the area under the FROC curve until 10 FPPI. This amount of false positives was chosen based on the performance of the expert rater at about 0.56 average sensitivity and about 4.4 average FPPI. It was set at approximately twice the FPPI of the rater. As often there is not an exact point at 10 false positives per image but there are values before and after, we interpolate the sensitivity value at 10 FPPI and add this point to the curve. The composite trapezoidal rule was used to numerically approximate the area under the curve. The resulting area is divided by the full possible area to get the ratio. This maximum possible area is 10.0 as the sensitivity ranges from 0 to 1 and we set the limit of the false positives from 0 to 10. So the FAUC corresponds to the area under the curve divided by 10, possibly multiplied by 100% to get the percentage. To provide a measure of the uncertainty of the FAUC we computed the standard deviation of the FAUC using bootstrapping of the test set. Bootstrapping was performed by random sampling with replacement from the test set, which means scans can be included more than once in the resulting set (Efron and Tibshirani, 1993). Sets of 1000 were obtained in this way and evaluated by computing the FAUC. After 1000 runs the mean FAUC and standard deviation of the FAUC were calculated.

9. Results

The proposed methods are compared in four ways. Firstly, to evaluate if the loss is indeed optimizing the CNN for detecting PVS, the FAUC (as described in section 8.2) is computed on the validation set on a regular interval during training. Figure 8 shows how the loss and detection performance quantified in FAUC vary during training of the network. Secondly, the FROC curve is computed per method on the test set which is shown in Figure 9. The FROC curves show how the methods perform at different thresholds in terms of sensitivity and FPPI. Thirdly, the mean FAUC is computed per method along with the standard deviation using bootstrapping on the test set. The results are shown in Table 2. Lastly, the methods are compared visually by computing the predicted map on a random image of the test set and overlaying this predicted map on the corresponding intensity image as shown in Figure 10 (more Figures like this are shown in Appendix D). Figure 11 also shows the intensity image with the predicted map as overlay as well as the separate predicted maps for a different image of the test set only for the four best performing methods (see Table 2 and 3) in terms of FAUC.

To examine how well the different loss functions and label images combine to optimize the CNN for detection of PVS, every 30 epochs the best models based on the validation set were saved and the corresponding FAUC was computed on the validation set. Figure 8 shows per method the training and validation loss as well as the FAUC during training. If the loss optimizes for detection the loss and the FAUC performance are highly correlated and follow the same trend in the plots. For most of the methods the loss and FAUC follow at least a similar trend. A particularly nice example of this is the CNN optimized using GDM^3 and wMSE (Figure 8h) as the loss and performance first improve together steadily after which they both plateau. However, sometimes this is clearly not the case, see e.g. e^{GDM} MSE (Figure 8a) and GDM wMSE (Figure 8e) that reach the maximum in their performance long before convergence and decrease in performance during convergence. The best model per method was chosen based on these FAUC values and evaluated on the test set.

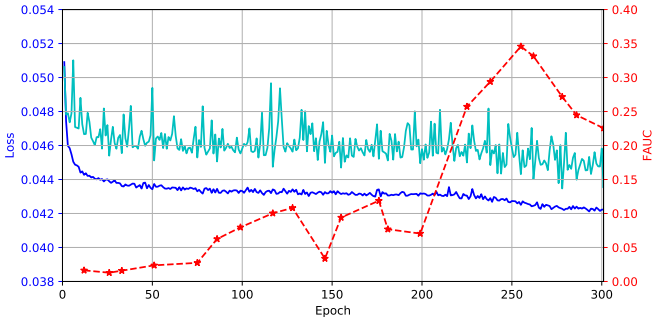
To compare the performance of the proposed methods on the test set of 1000 images the FROC curves were computed as described in section 8.2 and plotted in groups based on the different losses. The FROCs per loss function are shown in Figure 9. The color of the curve indicates the label image that was used as ground truth during optimization. This is shown in the legend along with the CNNs corresponding FAUC value. Learning failed for the two CNNs using DSC loss as well as for the CNN that were supposed to predict the GDM using $TwMSE_{T=0.8}$ loss. For this reason the FROC curves for these methods are missing in the plots.

The performance of the expert rater (shown in Table 1) is added in the FROCs with a red star. The best methods on weakly labels developed by Dubost et al. (2017) are GP-Unet and Grad-CAM and are shown in the FROCs as a blue + and a purple x respectively. These methods were developed on the same training and validation set and tested on the same test set to allow for optimal comparison of performance between the methods.

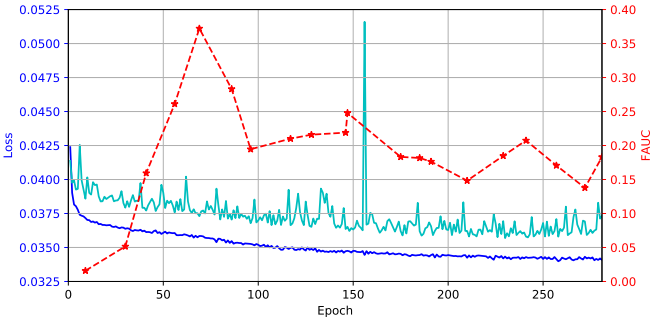
If a curve passes under the points indicating the performance of the rater, this method is considered to be inferior in performance. Curves that pass above these points would seem to show increased performance compared to the rater. However, we consider the performance of the expert rater as the limit, as the performance measure is dependent on the annotations of this rater. Therefore curves that are higher than this point might actually overfit on the annotations.

For the regression approach especially the CNNs trained with the GDM^3 as label image perform very well, matching the expert rater's performance almost exactly for some thresholds. Only the CNN optimized with MSE loss and GDM^3 as label image showed worse performance. In general the CNNs optimized with MSE loss did not perform well. Without the weighting used in the other loss functions the optimization might be too focused on the whole image and not enough on the PVS.

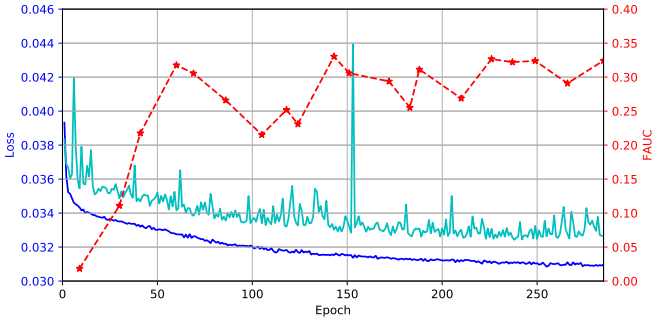
Out of the four methods proposed for the segmentation approach the two using DSC loss failed to learn and were excluded. Meanwhile the remaining two methods for this approach that were both optimized using BCE even surpassed the expert rater's performance by the most distance of all proposed methods. However, as mentioned before this performance is based on the annotations of this expert rater, so better performance on those annotations could mean the CNNs are overfitting on the rater's



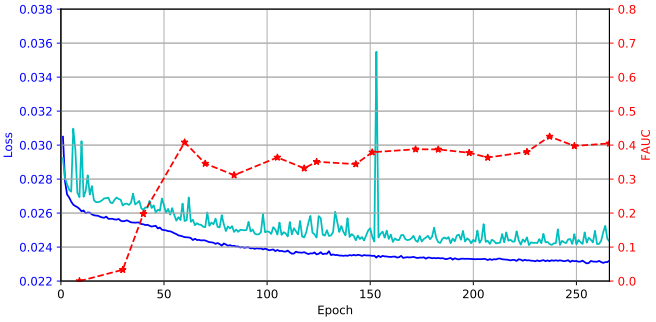
(a) **GDM MSE**



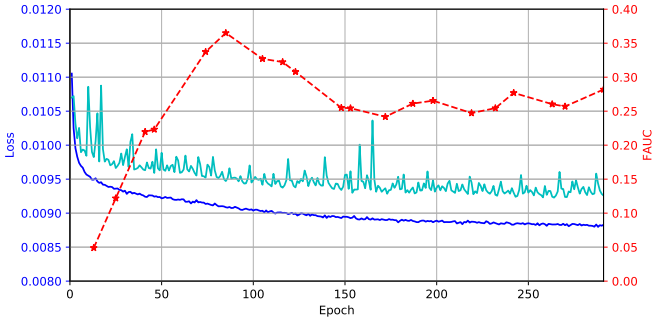
(b) **e$^{\text{GDM}}$ MSE**



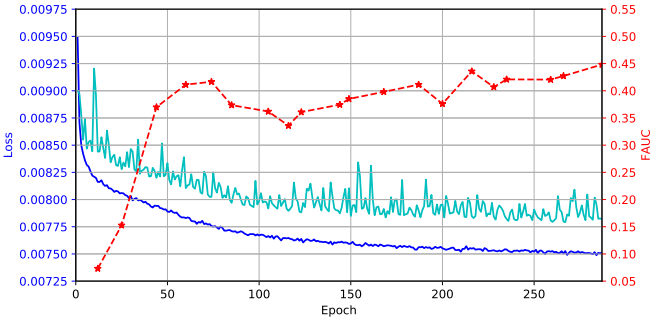
(c) **GDM^2 MSE**



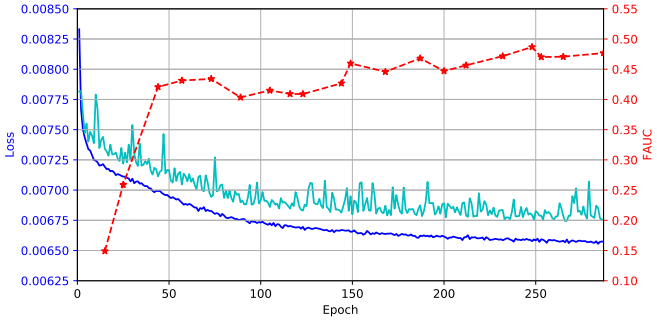
(d) **GDM^3 MSE**



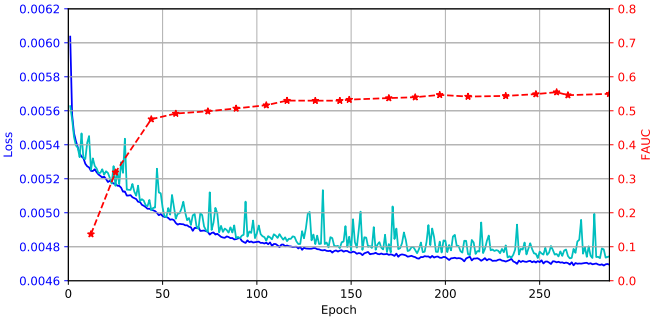
(e) **GDM wMSE**



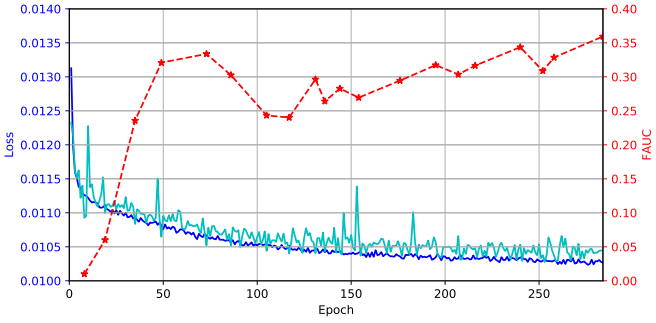
(f) **e$^{\text{GDM}}$ wMSE**



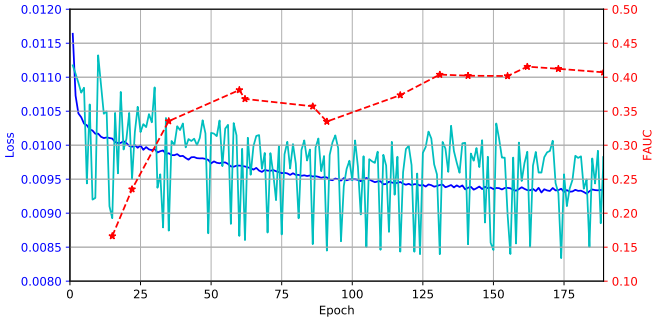
(g) **GDM^2 wMSE**



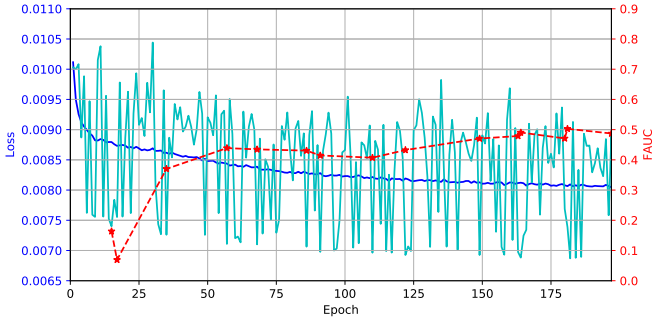
(h) **GDM^3 wMSE**



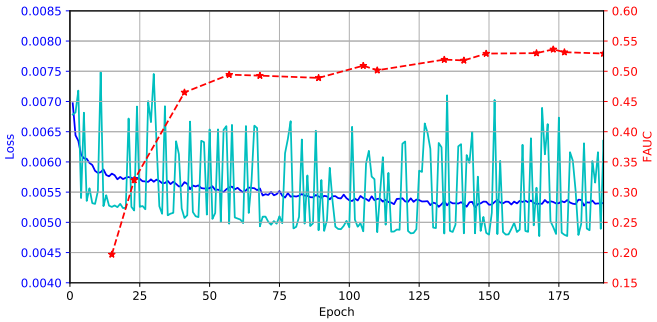
(i) **GDM TwMSE$_{\text{T=0.5}}$**



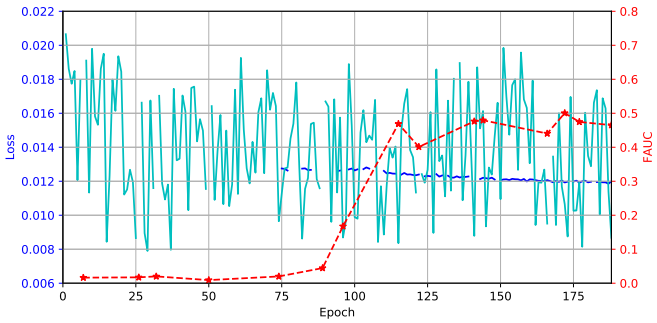
(j) $\mathbf{e^{GDM}\ TwMSE_{T=0.5}}$



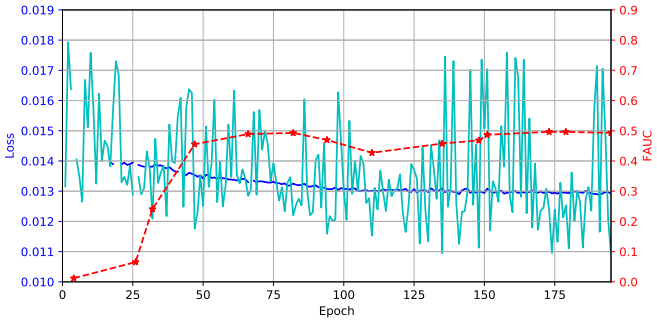
(k) $\mathbf{GDM^2\ TwMSE_{T=0.5}}$



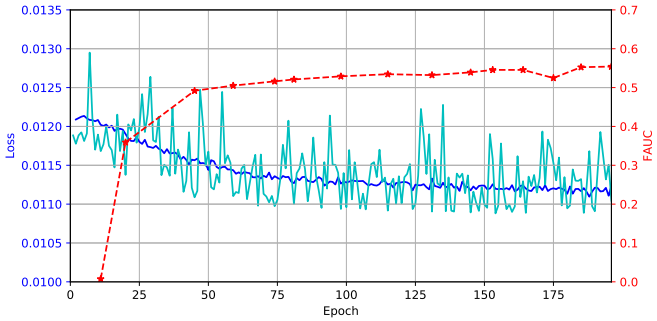
(l) $\mathbf{GDM^3\ TwMSE_{T=0.5}}$



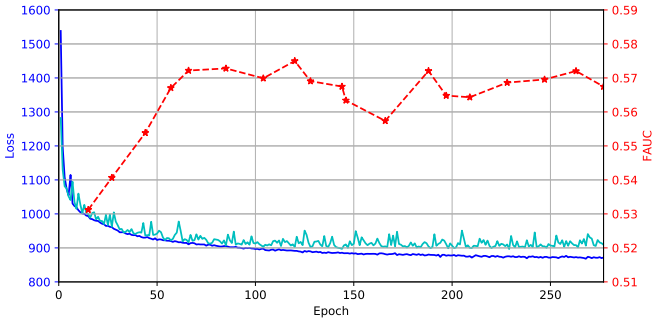
(m) $\mathbf{e^{GDM}\ TwMSE_{T=0.8}}$



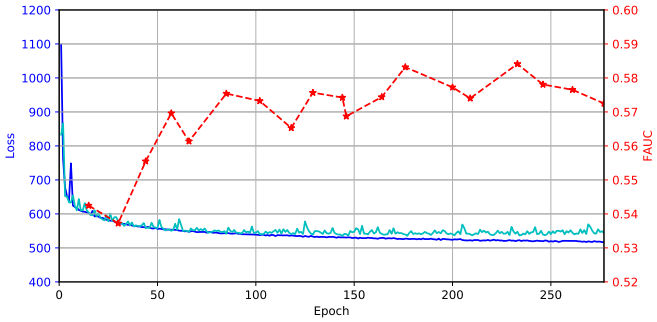
(n) $\mathbf{GDM^2\ TwMSE_{T=0.8}}$



(o) $\mathbf{GDM^3\ TwMSE_{T=0.8}}$

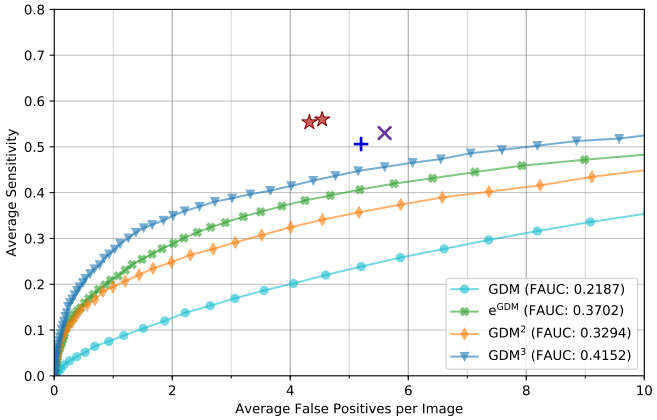


(p) **T=0.95 BCE**

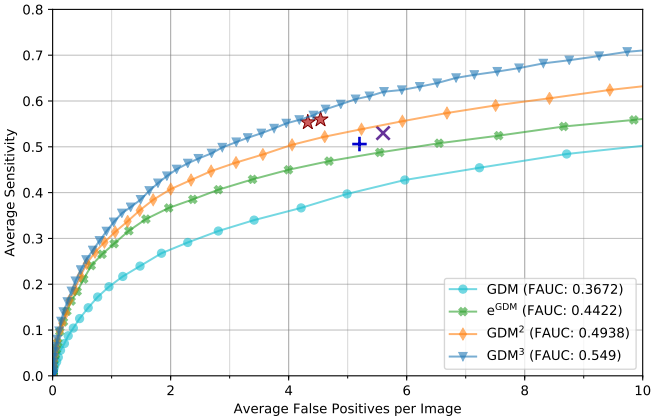


(q) **T=0.96 BCE**

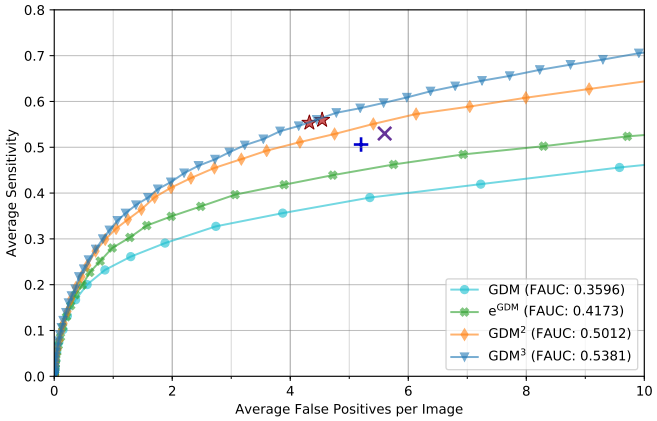
Figure 8: **Loss and detection performance during training.** Training was monitored by looking at the training loss (dark blue, left y-axis) and the validation loss that was computed on the validation set (cyan, left y-axis). Every 30 epochs the best model based on the validation loss was saved. These models were subsequently evaluated on the validation set on detection performance which was quantified by the FAUC value (described in section 8.2. Using these FAUC values the detection performance can be plotted along the epochs (red, right y-axis). These plots show how the detection performance varies during convergence of the network. Ideally if the loss optimizes for detection, the performance on the loss and on detection should be highly correlated and the same trend would be seen in the loss and in the detection performance over the epochs. A nice example of this is the CNN optimized using GDM^3 and wMSE (h) as the loss and performance first improve together steadily after which they both plateau. However sometimes this is clearly not the case, see e.g. e^{GDM} MSE (b) and GDM wMSE (e) that reach the maximum in their performance way before convergence and decrease in performance during convergence. Note that the axes are customized for every plot. This is because the methods vary largely in how high their best performance is, how long it takes for the network to converge based on the optimization and the losses they are optimized for span different ranges.



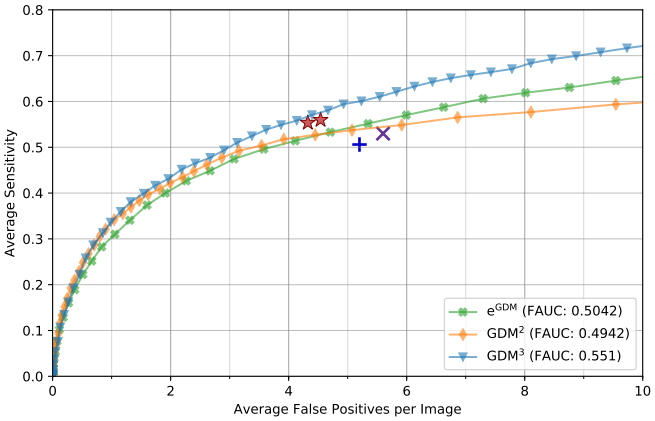
(a) **MSE loss**



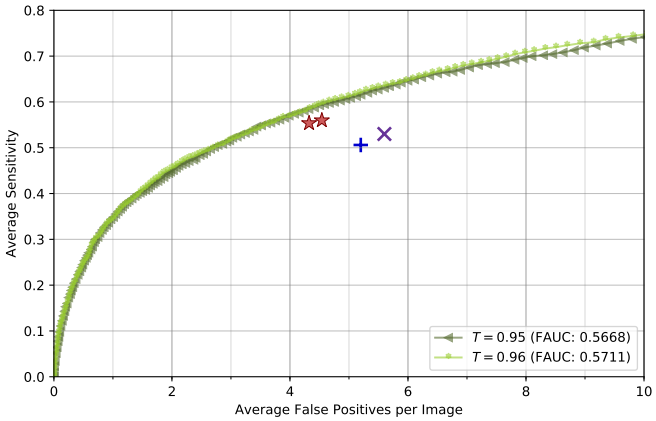
(b) **wMSE loss**



(c) **$\text{TwMSE}_{\text{T=0.5}}$ loss**

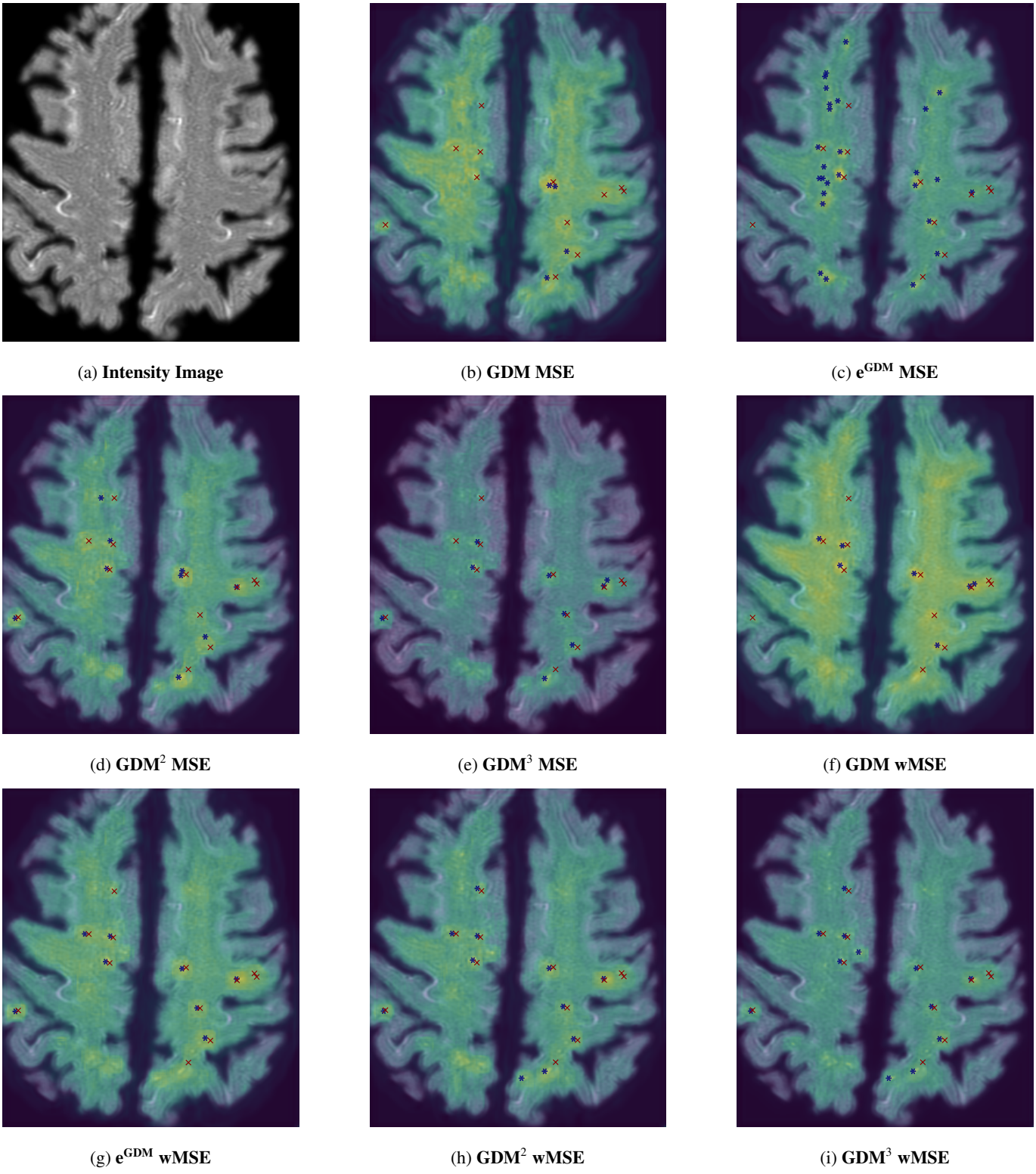


(d) **$\text{TwMSE}_{\text{T=0.8}}$ loss**



(e) **BCE loss**

Figure 9: **Free-Response Operating Characteristic (FROC) curves for CNNs optimized using different losses and label images.** The curves are computed as described in section 8.2 on the test set of 1000 scans. They are grouped in plots according to the loss function that was used to optimize the CNN. The color of the curve indicates the label image that was used as ground truth during optimization. This is shown in the legend along with the CNNs corresponding FAUC value. As described in section 7 for the voxel-wise regression approach CNNs were optimized using mean squared error (MSE) loss (a), label-weighted MSE (wMSE) loss (b), thresholded wMSE loss with a threshold at 0.5 ($\text{TwMSE}_{\text{T=0.5}}$) (c) and with a threshold at 0.8 ($\text{TwMSE}_{\text{T=0.8}}$) (d). The CNNs were optimized to either predict the geodesic distance map (GDM) computed from the dot annotations or a modified GDM (see Figure 6). For the segmentation approach binary cross entropy (BCE) loss (e) and dice similarity coefficient (DSC) loss were used to optimize the CNNs to predict the approximated segmentations of the PVS that were obtained by thresholding the GDM at 0.95 or 0.96 (see Figure 6). Learning failed for the two CNNs using DSC loss as well as for the CNN that was supposed to predict the GDM using $\text{TwMSE}_{\text{T=0.8}}$ loss. For this reason the FROC curves for these methods are missing in the plots. The red stars correspond to the performance of the expert rater on a separate set of 40 images. The blue + corresponds to the performance of GP-UNet and the purple X to Grad-CAM, two detection methods developed and adapted respectively for PVS detection by Dubost et al. (2017). These methods were trained and validated on the same sets as our methods. The plots are scaled with a sensitivity range until 0.8 to make the plots more clear.

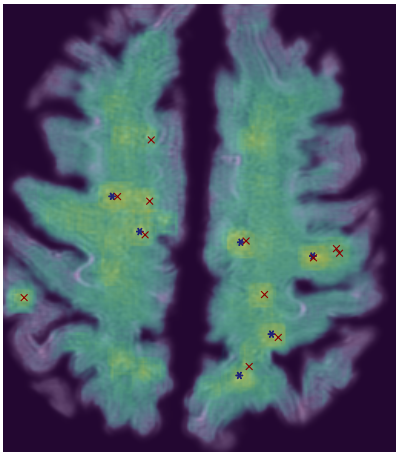


(a) **Intensity Image** (b) **GDM MSE** (c) **e^{GDM} MSE**

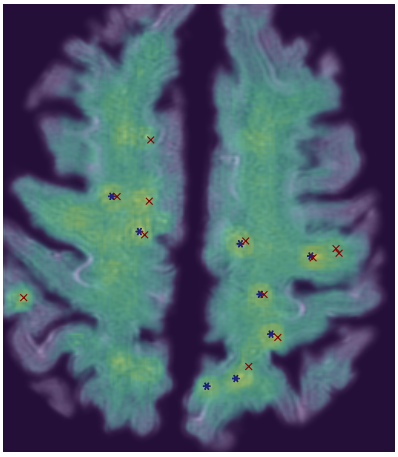
(d) **GDM^2 MSE** (e) **GDM^3 MSE** (f) **GDM wMSE**

(g) **e^{GDM} wMSE** (h) **GDM^2 wMSE** (i) **GDM^3 wMSE**

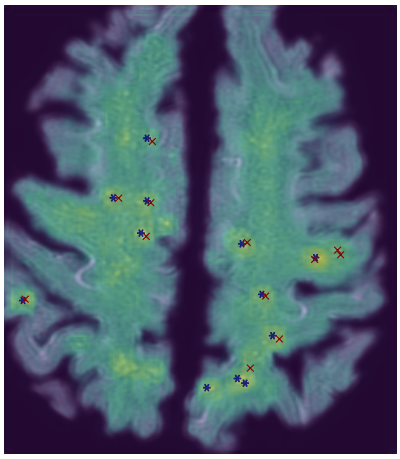
Figure 10: **Predictions on an image from the test set.** For every method the predicted map is overlaid on the intensity image given as input to the model. Red crosses are the ground truth given by the expert rater (so no shifting of dots during inference). The blue asterisks indicate the proposed detections at the threshold closest to the performance of the expert rater on the test set. The overlaid prediction map is shown in a sequential perceptually uniform color scale that ranges from purple for the lowest value in the image to yellow for the highest value. The contrast in the predicted maps illustrate the spread of the values in the predicted maps. Note that the predicted maps are not scaled, so the range of values in predicted maps may vary. Learning failed for the two CNNs using DSC loss as well as for the CNN that was supposed to predict the GDM using $\text{TwMSE}_{T=0.8}$ loss. For this reason the visualizations for these CNNs are missing.



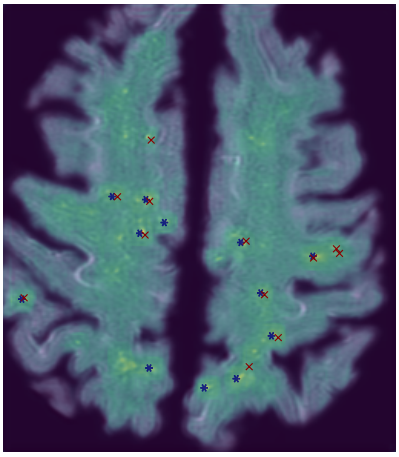
(j) **GDM TwMSE$_{T=0.5}$**



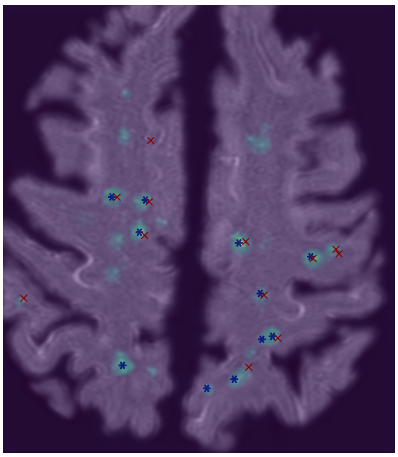
(k) **e^{GDM} TwMSE$_{T=0.5}$**



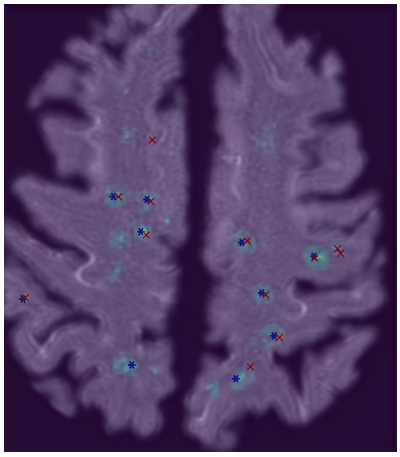
(l) **GDM^2 TwMSE$_{T=0.5}$**



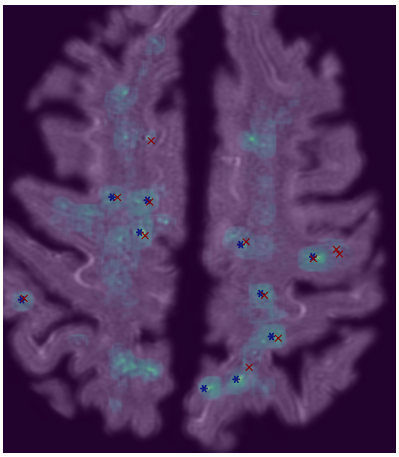
(m) **GDM^3 TwMSE$_{T=0.5}$**



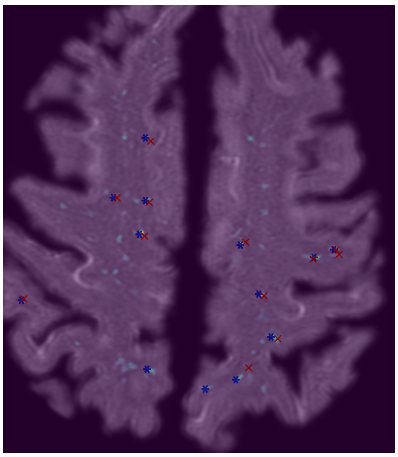
(n) **e^{GDM} TwMSE$_{T=0.8}$**



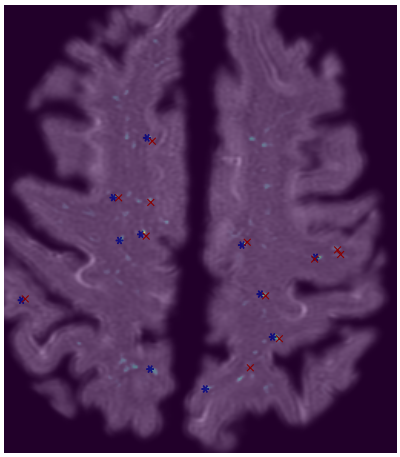
(o) **GDM^2 TwMSE$_{T=0.8}$**



(p) **GDM^3 TwMSE$_{T=0.8}$**



(q) **T=0.95 BCE**

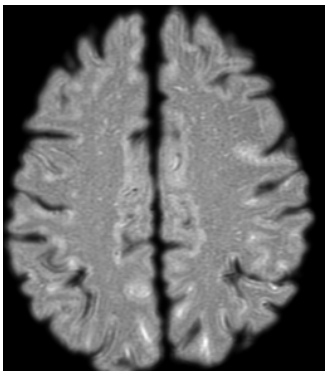


(r) **T=0.96 BCE**

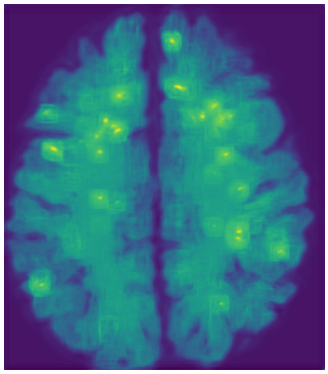
Figure 10: **Predictions on an image from the test set.** For every method the predicted map is overlaid on the intensity image given as input to the model. Red crosses are the ground truth given by the expert rater (so no shifting of dots during inference). The blue asterisks indicate the proposed detections at the threshold closest to the performance of the expert rater on the test set. The overlaid prediction map is shown in a sequential perceptually uniform color scale that ranges from purple for the lowest value in the image to yellow for the highest value. The contrast in the predicted maps illustrate the spread of the values in the predicted maps. Note that the predicted maps are not scaled, so the range of values in predicted maps may vary. Learning failed for the two CNNs using DSC loss as well as for the CNN that was supposed to predict the GDM using TwMSE$_{T=0.8}$ loss. For this reason the visualizations for these CNNs are missing.

Table 2: **FAUCs** Varying losses and ground truth images on test set (1000 images). Bootstrapping is used to quantify the uncertainty, resulting in a mean FAUC and a standard deviation given in the brackets. Note that the FAUCs and corresponding standard deviations are shown in percentages, to improve readability.

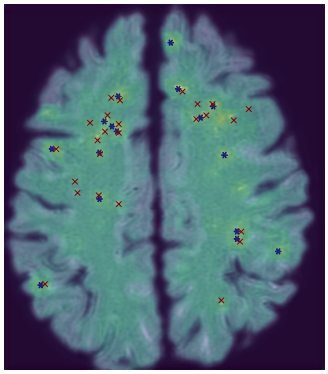
	GDM	**e$^{\text{GDM}}$**	**GDM^2**	**GDM^3**
MSE	21.91 (± 0.63)	37.02 (± 0.73)	32.96 (± 0.81)	41.56 (± 0.83)
wMSE	36.73 (± 0.73)	44.24 (± 0.81)	49.39 (± 0.81)	**54.90 (± 0.84)**
TwMSE$_{\text{T=0.5}}$	35.95 (± 0.75)	41.78 (± 0.78)	50.06 (± 0.85)	53.85 (± 0.83)
TwMSE$_{\text{T=0.8}}$	-	50.45 (± 0.83)	49.38 (± 0.87)	**55.11 (± 0.82)**



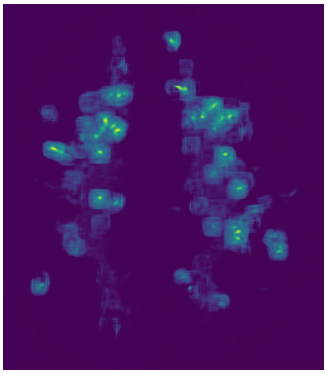
(a) **Corresponding Intensity Image**



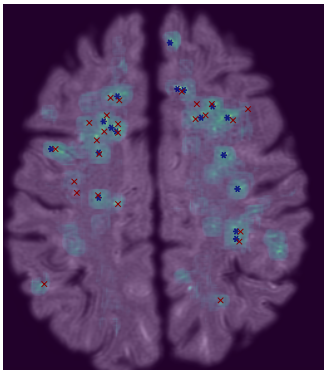
(b) **GDM^3 wMSE**



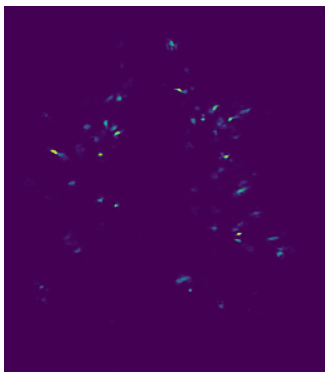
(c) **GDM^3 wMSE**



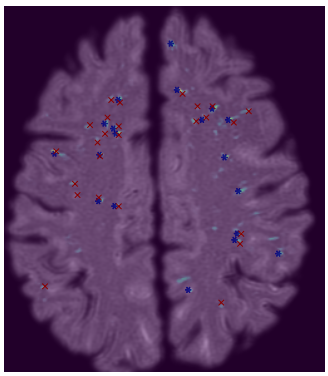
(d) **GDM^3 TwMSE$_{\text{T=0.8}}$**



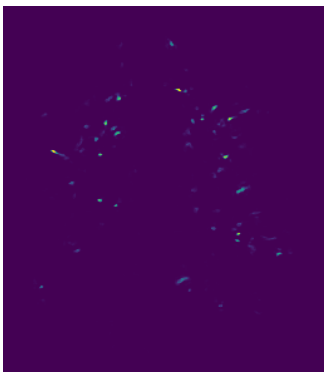
(e) **GDM^3 TwMSE$_{\text{T=0.8}}$**



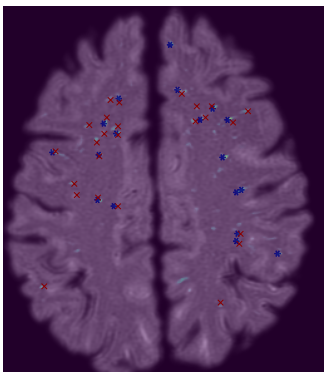
(f) **T=0.95 BCE**



(g) **T=0.95 BCE**



(h) **T=0.96 BCE**



(i) **T=0.96 BCE**

Figure 11: **Predictions by the four best performing CNNs.** For every method the predicted map is separately shown and is overlaid on the intensity image given as input to the model. Red crosses are the ground truth given by the expert rater (so no shifting of dots during inference). The blue asterisks indicate the proposed detections at the threshold closest to the performance of the expert rater on the test set. The overlaid prediction map is shown in a sequential perceptually uniform color scale that ranges from purple for the lowest value in the image to yellow for the highest value. The contrast in the predicted maps illustrate the spread of the values in the predicted maps. Note that the predicted maps are not scaled, so the range of values in predicted maps may vary. Learning failed for the two CNNs using DSC loss as well as for the CNN that was supposed to predict the GDM using TwMSE$_{\text{T=0.8}}$ loss. For this reason the visualizations for these CNNs are missing.

annotation style.

To quantify the uncertainty in the FAUC values bootstrapping was performed on the test set as described in section 8.2.2 and producing a mean FAUC per method as well as a standard deviation. For the regression approach the performance of the CNNs is shown in Table 2 and for the segmentation approach this is shown in Table 3. The best performances in terms of FAUC are shown in bold. Note that the values in the brackets shown in the tables are standard deviations and not confidence intervals.

To visually examine the performance of the CNNs the prediction maps produced by the CNNs are overlaid on a random image of the test set shown in Figure 10. The predicted maps are shown using a color scale that is perceptually uniform. This way the difference in range and distribution of values in the predicted maps can be seen. The most clear difference is between the CNNs optimized to predict the GDM e.g. which have a very gradual increase in values from the borders to the predictions (Figures 10b and 10f), while the CNNs of the segmentation approach seem to have a bimodal distribution of values (Figures 10q and 10r). This corresponds with the label images they were optimized with, namely the GDM that has a gradual spread in values and the approximated segmentations that are binary. It is interesting to see that the CNNs optimized with $\text{TwMSE}_{\text{T=0.8}}$ show a quite bimodal distribution as well. Also valuable to note is that even though the mask voxels can have any value below 0.8, most voxels have the same approximate value (see also 11).

Furthermore, the detections proposed by the CNNs at the threshold closest in performance on the test set (in terms of Euclidean distance) to the rater's performance are shown as blue asterisks. The dot annotations are shown as red crosses. The same visualization is shown in Appendix D for another randomly chosen image from the test set.

Lastly, Figure 11 shows for a different random image of the test set the prediction maps overlaid on the intensity image and the proposed detections and also the predicted map by itself for the four best performing CNN in terms of FAUC (see Table 2 and 3). This figure shows more clearly the difference in predicted maps computed by the CNNs, with the regression CNNs predicting a more smoothed prediction map and the segmentation CNNs predicting the PVS more precisely.

The CNNs took about a week of training to fully converge, some took even longer. However, once trained the whole method takes less than a minute, with the CNNs outputting the predicted map based on the input image and the post-processing steps defined in section 7.6 outputting the final detection proposals.

Table 3: **FAUCs** Binary cross entropy (BCE) and the Dice similarity coefficient (DSC) were used as voxel-wise segmentation losses in varying combinations with a thresholded geodesic distance map (GDM) at two different thresholds producing an approximate segmentation of the enlarged perivascular spaces (PVS). The performance is evaluated on the test set (1000 images). Bootstrapping is used to quantify the uncertainty, resulting in a mean FAUC and a standard deviation given in the brackets. Note that the FAUCs and corresponding standard deviations are shown in percentages, to improve readability. Learning failed for both CNNs using DSC loss.

	BCE	**DICE**
Threshold 0.95	**56.67 (± 0.85)**	-
Threshold 0.96	**57.19 (± 0.83)**	-

10. Discussion

Both approaches for optimizing the CNN to detect PVS brought forth methods that match human performance in detecting PVS. Once optimized the whole method including the CNN takes less than a minute to detect the PVS in a given image. From our experiments we can conclude that for the regression approach modifying the GDMs to emphasize the PVS improved performance as well as weighting the loss to focus more on the PVS. For the segmentation approach only the BCE loss managed to optimize the CNN to detect PVS.

The best method in terms of performance (see Table 2 and 3) is the CNN optimized with BCE and smallest approximated segmentations (T=0.96, Figure 6h). However the other CNN optimized with BCE performs almost as well with less than 1 standard deviation between the FAUCs. However, both CNNs even surpass the performance of the expert rater, raising the question if they are overfitting on the annotation style. We defined the performance of the expert rater to be the limit of what is possible. In this way the CNNs that were optimized for regression with GDM^3 as label image and either wMSE and TwMSE as loss function had the best performance as they almost exactly match the performance of the rater.

The methods that perform best on the test set are all methods that show a similar trend in loss and detection performance during training (see Figure 8) indicating that their combination of loss function and label image really seem to optimize for detection.

Both CNNs that were optimized for segmentation seem to output more precise predictions than the other methods. This is understandable as they are trained with binary images containing the approximated segmentations while the other methods are trained with continuous images. The CNNs optimized for regression show predictions that are more smoothed. This was part of our logic of using GDM regression for detecting PVS. As the annotations are subject to observer bias, our idea was to use a more smoothed ground truth to regularize and decrease the chance of overfitting to the labels. The following observation contributes to this hypothesis. The CNNs optimized for segmentation perform very well, but do seem to overfit on the annotations as they even surpass the rater's performance, while the best CNNs optimized for regression almost exactly match the performance of the rater.

The best trade-off between sensitivity and average false positives per image is not very clear. Due to the difficulty of distinguishing PVS it is more ambiguous how many false-positives are still acceptable because some might actually be false negatives in the ground truth. It would be beneficial to find a way to let the CNN learn a detection threshold like Dubost et al. (2017) show with GP-UNet. However as the current objective was to match the raters performance, we now choose the threshold based on the closest Euclidean distance to the raters performance. Further

research would be beneficial to figure out a sensible threshold for detection.

It is noteworthy that methods that do better in the low FP range do not reach full sensitivity while methods that perform worse at low FP do reach full sensitivity at a lot of false positives. This could indicate that the brightness plays an important role at first but when methods are better at predicting PVS they rely on other image features.

The TwMSE might not optimize correctly because the thresholding in the loss makes this loss function less nicely differentiable. Especially the TwMSE with a higher threshold showed some convergence problems especially for the GDM and exponentially weighted GDM.

This method might also work for similar image analysis applications, especially when structures with a complex morphology are the aim of detection.

The current methods were not evaluated on performance of segmentation of PVS. It would be interesting to see how well these methods trained with dot annotations perform in terms of segmenting the PVS. Furthermore combining the currently proposed methods for detection with (semi-)automated methods for PVS segmentation currently proposed in the literature (Zhang et al., 2016; Ballerini et al., 2018; Park et al., 2016) has the potential to perform very well at segmentation of PVS.

Many neural network architectures require a fixed input image size due to fully connected layers, however since FCNs consist of only convolutional layers and pooling layers these neural networks accept images of arbitrary size. As the CNNs in this study are fully convolutional, once they are trained they can accept images of any size and compute PVS detections. This has not been tried yet. It would be interesting to see how well the CNNs would handle different sized images as well as different spatial resolution. The PDw MRI scans used in this thesis were acquired at a clinical field strength of 1.5 T. It would be interesting to see how the methods perform on 3 T or 7 T scans which have a higher spatial resolution.

Our best CNNs clearly outperform the weakly labeled methods for PVS detection. The dot annotations that we used for optimizing the CNN contained the location information of the PVS while the weak labels (number of PVS per slice) used to optimize Grad-CAM and GP-UNet did not contain this information. Our results show that the location information is important for improving the performance of CNNs for detecting PVS.

Most of the networks took quite long to converge. Various methods have been proposed in the literature to speed up convergence e.g. ELU and batch normalization (Clevert et al., 2015). This would be interesting to look at, if the convergence could be sped up while maintaining the same performance.

As shown in the results PVS detection can be improved by training on the location of the PVS. However it is not clear how much the CNNs overfit on style of the rater. Sensitivity rate is only 0.55 in CSO so there are a lot of false negatives and false positives in both annotations of the rater, which is understandable as PVS are very difficult to distinguish from their mimics. It would be very interesting and valuable to test the current methods on annotations from another rater to see how much these methods have overfit on our current annotations.

As the intra-rater agreement is quite low and this was the limit for the performance of our models, the next step would be to improve the ground truth. Combining annotations from multiple raters could help decrease the variability in the annotations.

Furthermore trying the best methods on the other brain regions would be valuable to see which methods are most robust for PVS detection in the brain.

11. Conclusion

In this thesis we proposed two approaches for optimizing a CNN to detect PVS. For both approaches various label images and loss functions were combined and compared. Both approaches brought forth methods that match human intra-rater performance in detecting PVS without the need for any user interaction.

Once optimized our one-stage detection method takes less than a minute to detect the PVS in a given image.

We can conclude that it is possible to use dot annotations to optimize a CNN for detection of perivascular spaces that matches the performance of an expert rater. To the best of our knowledge we are the first to compare and match human performance on the detection of PVS.

12. Acknowledgment

I had an amazing time working on this research and would like to thank everyone who made it possible.

First of all I would like to thank my daily supervisor Florian Dubost for all our amazing brainstorm sessions & meetings, and for making my thesis a super fun, intense semi-fair competition, which always kept me motivated and enthusiastic! Even though I feel like I was privileged as I had better labels, I do feel like I won that house right? Or was it a brownie? Many thanks for the crash course on Ubuntu & Deep Learning, on why it is important to be very skeptical of Dark Deals even though the idea of GPU power is very tempting, and that Marvel is amazing!

I would like to thank Marleen de Bruijne for our meetings and all her great questions, making me continuously question what I was doing, and her kind, understanding and down to earth attitude that really helped me to take a step back and think about what I had been up to lately in my project.

I would like to thank Frans Vos for all his help and great advice during my time at the TU Delft, from developing a premaster for me so I could switch from Veterinary Medicine to Biomedical Engineering, to being a part of my graduation committee so I can (hopefully) graduate!

Many thanks to everyone at BIGR for the amazing time I had with you! I really enjoyed the great welcoming and kind environment at BIGR and the super fun game nights! Special thanks to the BIGR Sloths for all the fun banana breaks, discussions about deep learning and about how we survive in a world made of cake (still not sure about that insuline idea haha) and definitely thanks for getting me to take breaks when I thought I didn't need a break but in reality definitely needed one.

I would like to thank my family, Roel and my friends for supporting me through my whole journey from Veterinairy Medicine to finally finding my passion in Medical Image Analysis.

Many thanks to Meike Vernooij and Arfan m. Ikram for the great dataset.

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.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Yangqing, J., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems.

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., Seshadri, S., van Duijn, C.M., Ikram, M.K.A.K., Grabe, H.J., Schmidt, R., Ikram, M.K.A.K., 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. doi:10.1016/J.DADM.2015.10.004.

Adams, H.H.H., Cavalieri, M., Verhaaren, B.F.J., 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 , STROKEAHA—-111.

Andrews, S., Hamarneh, G., 2015. Multi-Region Probabilistic Dice Similarity Coefficient using the Aitchison Distance and Bipartite Graph Matching arXiv:1509.07244.

Annese, J., 2012. The importance of combining MRI and large-scale digital histology in neuroimaging studies of brain connectivity and disease. Frontiers in Neuroinformatics 6, 13. doi:10.3389/fninf.2012.00013.

Bacyinski, A., Xu, M., Wang, W., Hu, J., 2017. The Paravascular Pathway for Brain Waste Clearance: Current Understanding, Significance and Controversy. Frontiers in Neuroanatomy 11, 101. doi:10.3389/fnana.2017.00101.

Bai, X., Sapiro, G., 2007. A geodesic framework for fast interactive image and video segmentation and matting, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE. pp. 1–8. doi:10.1109/ICCV.2007.4408931.

Bakker, E.N., Bacskai, B.J., Arbel-Ornath, M., Aldea, R., Bedussi, B., Morris, A.W., Weller, R.O., Carare, R.O., 2016. Lymphatic Clearance of the Brain: Perivascular, Paravascular and Significance for Neurodegenerative Diseases. doi:10.1007/s10571-015-0273-8.

Ballerini, L., Lovreglio, R., Valdés Hernández, M.D.C., 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, 2132. doi:10.1038/s41598-018-19781-5, arXiv:1704.07699.

Barkhof, F., 2004. Enlarged Virchow-Robin spaces: Do they matter? Journal of Neurology, Neurosurgery and Psychiatry 75, 1516–1517. doi:10.1136/jnnp.2004.044578.

Bechmann, I., Kwidzinski, E., Kovac, A.D., Simbürger, E., Horvath, T., Gimsa, U., Dirnagl, U., Priller, J., Nitsch, R., 2001. Turnover of rat brain perivascular cells. Experimental Neurology 168, 242–249. doi:10.1006/exnr.2000.7618.

Boespflug, E.L., Schwartz, D.L., Lahna, D., Pollock, J., Iliff, J.J., Kaye, J.A., Rooney, W., Silbert, L.C., 2018. MR Imagingbased Multimodal Autoidentification of Perivascular Spaces (mMAPS): Automated Morphologic Segmentation of Enlarged Perivascular Spaces at Clinical Field Strength. Radiology 286, 632–642. doi:10.1148/radiol.2017170205.

Bokura, H., Kobayashi, S., Yamaguchi, S., 1998. Distinguishing silent lacunar infarction from enlarged Virchow-Robin spaces: A magnetic resonance imaging and pathological study. Journal of Neurology 245, 116–122. doi:10.1007/s004150050189.

Borgefors, G., 1986. Distance Transformations in Digital Images. Computer vision, graphics, and image processing 34, 344–371. doi:10.1016/1049-9660(91)90070-6.

Bouvy, W.H., Zwanenburg, J.J.J., Reinink, R., Wisse, L.E.E., Luijten, P.R., Kappelle, L.J., Geerlings, M.I., Biessels, G.J., 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 and Metabolism 36, 1708–1717. doi:10.1177/0271678X16648970.

Braffman, B.H., Zimmerman, R.A., Trojanowski, J.Q., Gonatas, N.K., Hickey, W.F., Schlaepfer, W.W., 1988. Brain MR: Pathologic correlation with gross and histopathology. 1. Lacunar infarction and Virchow-Robin spaces. American Journal of Roentgenology 151, 551–558. doi:10.2214/ajr.151.3.551.

Cai, K., Tain, R., Das, S., Damen, F.C., Sui, Y., Valyi-Nagy, T., Elliott, M.A., Zhou, X.J., 2015. The feasibility of quantitative MRI of perivascular spaces at 7T. Journal of Neuroscience Methods 256, 151–156. doi:10.1016/j.jneumeth.2015.09.001, arXiv:15334406.

Cárdenes, R., Alberola-López, C., Ruiz-Alzola, J., 2010. Fast and Accurate Geodesic Distance Transform by Ordered Propagation. Image and Vision Computing 28, 307–316.

Cerrolaza, J.J., Oktay, O., Gomez, A., Matthew, J., Knight, C., Kainz, B., Rueckert, D., 2017. Fetal Skull Segmentation in 3D Ultrasound via Structured Geodesic Random Forest, in: Fetal, Infant and Ophthalmic Medical Image Analysis, Springer International Publishing. pp. 25–32. doi:10.1007/978-3-319-67561-9_3.

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. Journal of Neurology, Neurosurgery and Psychiatry 84, 624–629. doi:10.1136/jnnp-2012-304434.

Chen, W., Song, X., Zhang, Y., 2011. Assessment of the virchow-robin spaces in Alzheimer disease, mild cognitive impairment, and normal aging, using high-field MR imaging. American Journal of Neuroradiology 32, 1490–1495. doi:10.3174/ajnr.A2541.

Chollet, F., Others, 2015. Keras. \url{https://keras.io}.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). Frontiers in Artificial Intelligence and Applications 285, 1760–1761. doi:10.3233/978-1-61499-672-9-1760, arXiv:1511.07289.

Criminisi, A., Sharp, T., Blake, A., 2008. GeoS: Geodesic image segmentation, in: Proceedings of the 10th European Conference on Computer Vision: Part I, Springer-Verlag. pp. 99–112. doi:10.1007/978-3-540-88682-2_9.

Cserr, H.F., Knopf, P.M., 1992. Cervical lymphatics, the blood-brain barrier and the immunoreactivity of the brain: a new view. doi:10.1016/0167-5699(92)90027-5.

Cuisenaire, O., 1999. Distance transformations: fast algorithms and applications to medical image processing .

Ćurić, V., Landström, A., Thurley, M.J., Luengo Hendriks, C.L., 2014. Adaptive mathematical morphology - A survey of the field. Pattern Recognition Letters 47, 18–28. doi:10.1016/j.patrec.2014.02.022.

Danielsson, P.E., 1980. Euclidean distance mapping. Computer Graphics and Image Processing 14, 227–248. doi:10.1016/0146-664X(80)90054-4.

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., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980. doi:10.1016/J.NEUROIMAGE.2006.01.021.

Doubal, F.N., MacLullich, A.M.J., Ferguson, K.J., Dennis, M.S., Wardlaw, J.M., 2010. Enlarged Perivascular Spaces on MRI Are a Feature of Cerebral Small Vessel Disease. Stroke 41, 450–454. doi:10.1161/STROKEAHA.109.564914.

Dubost, F., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., de Bruijne, M., 2018a. 3D Regression Neural Network for the Quantification of Enlarged Perivascular Spaces in Brain MRI doi:10.1080/10937404.2015.1051611.INHALATION, arXiv:1802.05914.

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, pp. 214–221. doi:10.1007/978-3-319-66179-7_25, arXiv:1705.07999.

Dubost, F., Yilmaz, P., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., de Bruijne, M., 2018b. Enlarged perivascular spaces in brain MRI: Automated quantification in four regions. NeuroImage doi:10.1016/j.neuroimage.2018.10.026.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Monographs on statistics and applied probability 57, 1–436.

Esiri, M.M., Gay, D., 1990. Immunological and neuropathological significance of the Virchow-Robin space. doi:10.1016/0022-510X(90)90004-7.

Etemadifar, M., Hekmatnia, A., Tayari, N., Kazemi, M., Ghazavi, A., Akbari, M., Maghzi, A.H.H., 2011. Features of Virchow-Robin spaces in newly diagnosed multiple sclerosis patients. European Journal of Radiology 80, e104–e108. doi:10.1016/j.ejrad.2010.05.018.

Fabbri, R., Costa, L.D.F., Torelli, J.C., Bruno, O.M., 2008. 2D Euclidean distance transform algorithms: A comparative survey. ACM Computing Surveys 40, 1–44. doi:10.1145/1322432.1322434.

Faghih, M.M., Sharp, M.K., 2018. Is bulk flow plausible in perivascular, paravascular and paravenous channels? Fluids and Barriers of the CNS 15, 17. doi:10.1186/s12987-018-0103-8.

Fanous, R., Midia, M., 2007. Perivascular spaces: Normal and giant. doi:10.1017/S0317167100005722.

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. doi:10.1016/j.seizure.2017.11.004.

Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A., 1998. Multiscale vessel enhancement filtering, in: Wells, W.M., Colchester, A., Delp, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention — MICCAI'98, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 130–137. doi:10.1007/BFb0056195.

Gaonkar, B., Macyszyn, L., Bilello, M., Sadaghiani, M.S., Akbari, H., Atthiah, M.A., Ali, Z.S., Da, X., Zhan, Y., Rourke, D.O., Grady, S.M., Davatzikos, C., 2015. Automated tumor volumetry using computer-aided image segmentation. Academic Radiology 22, 653–661. doi:10.1016/j.acra.2015.01.005.

González-Castro, V., Hernández, M.d.C.V., Armitage, P.A., Wardlaw, J.M., 2016a. Texture-based Classification for the Automatic Rating of the Perivascular Spaces in Brain MRI. Procedia Computer Science 90, 9–14. doi:10.1016/J.PROCS.2016.07.003.

González-Castro, V., Valdés Hernández, M.d.C., Armitage, P.A., Wardlaw, J.M., 2016b. Automatic rating of perivascular spaces in brain MRI using bag of visual words, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Cham. pp. 642–649. doi:10.1007/978-3-319-41501-7_72.

González-Castro, V., Valdés Hernández, M.d.C., 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. doi:10.1042/CS20170051.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Grazzini, J., Soille, P., Bielski, C., 2007. On the use of geodesic distances for spatial interpolation, in: Proceedings of the 9th International Conference on GeoComputation.

Grevera, G.J., 2007. Distance Transform Algorithms And Their Implementation And Evaluation. Springer New York, New York, NY. chapter 2. pp. 33–60. doi:10.1007/978-0-387-68413-0_2.

Groeschel, S., Chong, W.K., Surtees, R., Hanefeld, F., 2006. Virchow-Robin spaces on magnetic resonance images: Normative data, their dilatation, and a review of the literature. Neuroradiology 48, 745–754. doi:10.1007/s00234-006-0112-1.

Heier, L.A., Bauer, C.J., Schwartz, L., Zimmerman, R.D., Morgello, S., Deck, M.D., 1989. Large Virchow-Robin spaces: MR-clinical correlation. AJNR. American journal of neuroradiology 10, 929–36.

Holuša, M., Sojka, E., 2015. A k-max Geodesic Distance and Its Application in Image Segmentation, in: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 618–629. doi:10.1007/978-3-319-23192-1_52.

Hou, Y., Park, S.H., Wang, Q., Zhang, J., Zong, X., Lin, W., Shen, D., 2017. Enhancement of Perivascular Spaces in 7 T MR Image using Haar Transform of Non-local Cubes and Block-matching Filtering. Scientific Reports 7, 8569. doi:10.1038/s41598-017-09336-5.

Hurford, R., Charidimou, A., Fox, Z., Cipolotti, L., Jager, R., Werring, D.J., 2014. MRI-visible perivascular spaces: Relationship to cognition and small vessel disease MRI markers in ischaemic stroke and TIA. Journal of Neurology, Neurosurgery and Psychiatry 85, 522–525. doi:10.1136/jnnp-2013-305815.

Hutchings, M., Weller, R., 1986. Anatomical relationships of the pia mater to cerebral blood vessels in man. Journal of neurosurgery 65, 316–325. doi:10.3171/jns.1986.65.3.0316.

Ikonen, L., 2007. Priority pixel queue algorithm for geodesic distance transforms. Image and Vision Computing 25, 1520–1529. doi:10.1016/j.imavis.2006.06.016.

Ikonen, L., Toivanen, P., 2007. Distance and nearest neighbor transforms on gray-level surfaces. Pattern Recognition Letters 28, 604–612. doi:10.1016/j.patrec.2006.10.010.

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., Tiemeier, H., Uitterlinden, A.G., Vernooij, M.W., Hofman, A., 2017. The Rotterdam Study: 2018 update on objectives, design and main results. European Journal of Epidemiology 32, 807–850. doi:10.1007/s10654-017-0321-4.

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. doi:10.1007/s10654-015-0105-7.

Iliff, J.J., Wang, M., Zeppenfeld, D.M., Venkataraman, A., Plog, B.A., Liao, Y., Deane, R., Nedergaard, M., 2013. Cerebral Arterial Pulsation Drives Paravascular CSF-Interstitial Fluid Exchange in the Murine Brain. Journal of Neuroscience 33, 18190–18199. doi:10.1523/JNEUROSCI.1592-13.2013.

Jain, R., Kasturi, R., Schunck, B.G., 1995. Binary Image Processing, in: Machine Vision. McGraw-Hill, Inc.. chapter 2, pp. 25–72.

Jang, Y., Jung, H.Y., Hong, Y., Cho, I., Shim, H., Chang, H.J., 2016. Geodesic distance algorithm for extracting the ascending aorta from 3D CT images. Computational and Mathematical Methods in Medicine 2016. doi:10.1155/2016/4561979.

Jung, E., Zong, X., Lin, W., Shen, D., Park, S.H., 2018. Enhancement of Perivascular Spaces Using a Very Deep 3D Dense Network, in: International Workshop on PRedictive Intelligence In MEdicine. Springer International Publishing. volume 2, pp. 18–25. doi:10.1007/978-3-030-00320-3.

Jungreis, C.A., Kanal, E., Hirsch, W.L., Martinez, A.J., Moossy, J., 1988. Normal perivascular spaces mimicking lacunar infarction: MR imaging. Radiology 169, 101–104. doi:10.1148/radiology.169.1.3420242.

Kainz, P., Urschler, M., Schulter, S., Wohlhart, P., Lepetit, V., 2015. You should use regression to detect Cells, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 276–283. doi:10.1007/978-3-319-24574-4_33, arXiv:1505.04597.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36, 61–78. doi:10.1016/j.media.2016.10.004, arXiv:1603.05959.

Kilsdonk, I.D., Steenwijk, M.D., Pouwels, P.J.W., Zwanenburg, J.J.M., Visser, F., Luijten, P.R., Geurts, J.J.G., Barkhof, F., Wattjes, M.P., 2015. Perivascular spaces in MS patients at 7 Tesla MRI: A marker of neurodegeneration? Multiple Sclerosis Journal 21, 155–162. doi:10.1177/1352458514540358.

Kimmel, R., Sethian, J.A., 1998. Computing geodesic paths on manifolds. Proceedings of the National Academy of Sciences 95, 8431–8435. doi:10.1073/pnas.95.15.8431, arXiv:arXiv:1011.1669v3.

Kontschieder, P., Kohli, P., Shotton, J., Criminisi, A., 2013. GeoF: Geodesic forests for learning coupled predictors, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 65–72. doi:10.1109/CVPR.2013.16.

Krähenbühl, P., Koltun, V., 2014. Geodesic object proposals, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 725–739. doi:10.1007/978-3-319-10602-1_47.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly 2, 83–97. doi:10.1002/nav.3800020109.

Kwee, R.M., Kwee, T.C., 2007. Virchow-Robin Spaces at MR Imaging. RadioGraphics 27, 1071–1086. doi:10.1148/rg.274065722, arXiv:arXiv:1011.1669v3.

Lantuejoul, C., Beucher, S., 1981. On the use of the geodesic metric in image analysis. Journal of Microscopy 121, 39–49. doi:10.1111/j.1365-2818.1981.tb01197.x.

Levi, G., Montanari, U., 1970. A grey-weighted skeleton. Information and Control 17, 62–91. doi:10.1016/S0019-9958(70)80006-7.

Lian, C., Zhang, J., Liu, M., Zong, X., Hung, S.C.C., Lin, W., Shen, D., 2018. Multi-channel multi-scale fully convolutional network for 3D perivascular spaces segmentation in 7T MR images. Medical Image Analysis 46, 106–117. doi:10.1016/j.media.2018.02.009.

Liang, Y., Chan, Y.L., Deng, M., Chen, Y.K., Mok, V., Wang, D.F., Ungvari, G.S., Chu, C.w.W.w.W., Tang, W.K., 2018. Enlarged perivascular spaces in the centrum semiovale are associated with poststroke depression: A 3-month prospective study. Journal of Affective Disorders 228, 166–172. doi:10.1016/j.jad.2017.11.080.

Litjens, G., Kooi, T., Bejnordi, B.E., Arindra, A., Setio, A.A.A.A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M.W.M., van Ginneken, B., Sánchez, C.I., 2017. A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis 42, 60–88. doi:10.1016/j.media.2017.07.005, arXiv:arXiv:1702.05747v2.

Long, J., Shelhamer, E., Darrell, T., Long, J., Darrell, T., 2017. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 640–651. doi:10.1109/TPAMI.2016.2572683, arXiv:1411.4038.

Maclullich, A.M.J., Wardlaw, J.M., Ferguson, K.J., Starr, J.M., Seckl, J.R., Deary, I.J., 2004. Enlarged perivascular spaces are associated with cognitive function in healthy elderly men. Journal of Neurology, Neurosurgery and Psychiatry 75, 1519–1523. doi:10.1136/jnnp.2003.030858.

Öztürk, M.H., Aydingöz, Ü., 2002. Comparison of MR signal intensities of cerebral perivascular (Virchow-Robin) and subarachnoid spaces. Journal of Computer Assisted Tomography 26, 902–904. doi:10.1097/00004728-200211000-00008.

Paglieroni, D.W., 1992. Distance transforms: Properties and machine vision applications. CVGIP: Graphical Models and Image Processing 54, 56–74. doi:10.1016/1049-9652(92)90034-U.

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. doi:10.1016/j.neuroimage.2016.03.076.

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.

Pinto, A., Alves, V., Silva, C.A., 2016. Brain Tumor Segmentation using Convolutional Neural Networks in MRI Images. IEEE Transactions on Medical Imaging 35, 1240–1251. doi:10.1109/TMI.2016.2538465, arXiv:arXiv:1502.02445v2.

Potter, G.M., 2011. Neuroimaging of cerebral small vessel disease (PhD Thesis). Ph.D. thesis. University of Edinburgh.

Potter, G.M., Chappell, F.M., Morris, Z., Wardlaw, J.M., 2015. Cerebral perivascular spaces visible on magnetic resonance imaging: Development of a qualitative rating scale and its observer reliability. Cerebrovascular Diseases 39, 224–231. doi:10.1159/000375153.

Ramirez, J., Berezuk, C., McNeely, A.A., Gao, F., McLaurin, J.A., Black, S.E., 2016. Imaging the Perivascular Space as a Potential Biomarker of Neurovascular and Neurodegenerative Diseases. Cellular and Molecular Neurobiology 36, 289–299. doi:10.1007/s10571-016-0343-6.

Ramirez, J., Berezuk, C., McNeely, A.A., Scott, C.J.M., 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 : JAD 43, 415–424. doi:10.3233/JAD-132528.

Raza, S.E.A., AbdulJabbar, K., Jamal-Hanjani, M., Veeriah, S., Quesne, J.L., Swanton, C., Yuan, Y., 2018. Deconvolving convolution neural network for cell detection arXiv:1806.06970.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, Cham. pp. 234–241. doi:10.1007/978-3-319-24574-4_28, arXiv:1505.04597.

Rosenfeld, A., Pfaltz, J.L., 1966. Sequential Operations in Digital Picture Processing. Journal of the ACM 13, 471–494. doi:10.1145/321356.321357.

Rosenfeld, A., Pfaltz, J.L., 1968. Distance Functions on Digital Pictures. Pattern Recognition 1, 33–61. doi:10.1016/0031-3203(68)90013-7.

Saito, T., Toriwaki, J.I., 1994. New algorithms for euclidean distance transformation of an n-dimensional digitized picture with applications. Pattern Recognition 27, 1551–1565. doi:10.1016/0031-3203(94)90133-3.

Shen, D., Wu, G., Suk, H.I., 2017. Deep Learning in Medical Image Analysis. Annu. Rev. Biomed. Eng 19, 221–48. doi:10.1146/annurev-bioeng-071516, arXiv:15334406.

Sironi, A., Lepetit, V., Fua, P., 2014. Multiscale centerline detection by learning a scale-space distance transform, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2697–2704.

Strand, R., Normand, N., 2012. Distance transform computation for digital distance functions. Theoretical Computer Science 448, 80–93. doi:10.1016/j.tcs.2012.05.010.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S.S.S., Cardoso, M.J., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 240–248. doi:10.1007/978-3-319-67558-9_28, arXiv:1707.03237.

Toivanen, P.J., 1996. New geodosic distance transforms for gray-scale images. Pattern Recognition Letters 17, 437–450. doi:10.1016/0167-8655(96)00010-4.

Trebeschi, S., Van Griethuysen, J.J., Lambregts, D.M., Lahaye, M.J., Parmer, C., Bakers, F.C., Peters, N.H., Beets-Tan, R.G., Aerts, H.J., 2017. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. Scientific Reports 7, 5301. doi:10.1038/s41598-017-05728-9.

Uchiyama, Y., Kunieda, T., Asano, T., Kato, H., Hara, T., Kanematsu, M., Iwama, T., Hoshi, H., Kinosada, Y., Fujita, H., 2008. Computer-aided diagnosis scheme for classification of lacunar infarcts and enlarged Virchow-Robin spaces in brain MR images. Conference proceedings : 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society doi:10.1109/IEMBS.2008.4650064.

Valdés Hernández, M.d.C., Piper, R.J., Wang, X., Deary, I.J., Wardlaw, J.M., 2013. Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: A systematic review. Journal of Magnetic Resonance Imaging 38, 774–785. doi:10.1002/jmri.24047.

Vincent, L., 1993. Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algorithms. IEEE Transactions on Image Processing 2, 176–201. doi:10.1109/83.217222.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T., 2018. DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence doi:10.1109/TPAMI.2018.2840695, arXiv:1707.00652.

Wang, J., Tan, Y., 2013. Efficient Euclidean distance transform algorithm of binary images in arbitrary dimensions. Pattern Recognition 46, 230–242. doi:10.1016/j.patcog.2012.07.030.

Wang, P., Olbricht, W.L., 2011. Fluid mechanics in the perivascular space. Journal of Theoretical Biology 274, 52–57. doi:10.1016/j.jtbi.2011.01.014.

Wang, X., Chappell, F.M., Valdes Hernandez, M., Lowe, G., Rumley, A., Shuler, K., Doubal, F., Wardlaw, J.M., 2016a. Endothelial Function, Inflammation, Thrombosis, and Basal Ganglia Perivascular Spaces in Patients with Stroke. Journal of Stroke and Cerebrovascular Diseases 25, 2925–2931. doi:10.1016/j.jstrokecerebrovasdis.2016.08.007.

Wang, X., Valdés Hernández, M.d.C., Doubal, F., Chappell, F.M., Piper, R.J., Deary, I.J., Wardlaw, J.M., 2016b. Development and initial evaluation of a semi-automatic approach to assess perivascular spaces on conventional magnetic resonance images. Journal of Neuroscience Methods 257, 34–44. doi:10.1016/J.JNEUMETH.2015.09.010.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., O'Brien, J.T., Barkhof, F., Benavente, O.R., Black, S.E., Brayne, C., Breteler, M., Chabriat, H., DeCarli, C., de Leeuw, F.E., Doubal, F., Duering, M., Fox, N.C., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., van Oostenbrugge, R., Pantoni, L., Speck, O., Stephan, B.C., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P.B., Dichgans, M., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. doi:10.1016/S1474-4422(13)70124-8.

Wei, Y., Wen, F., Zhu, W., Sun, J., 2012. Geodesic saliency using

background priors, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Berlin, Heidelberg. pp. 29–42. doi:10.1007/978-3-642-33712-3_3.

Wuerfel, J., Haertle, M., Waiczies, H., Tysiak, E., Bechmann, I., Wernecke, K.D., Zipp, F., Paul, F., 2008. Perivascular spaces - MRI marker of inflammatory activity in the brain? Brain 131, 2332–2340. doi:10.1093/brain/awn171.

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 and Visualization 6, 283–292. doi:10.1080/21681163.2016.1149104.

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. doi:10.1016/j.media.2017.07.003.

Yatziv, L., Bartesaghi, A., Sapiro, G., 2006. O(N) implementation of the fast marching algorithm. doi:10.1016/j.jcp.2005.08.005.

Zeiler, M.D., 2012. ADADELTA: An Adaptive Learning Rate Method doi:http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503, arXiv:1212.5701.

Zhang, E.T., Inman, C.B., Weller, R.O., 1990. Interrelationships of the pia mater and the perivascular (Virchow-Robin) spaces in the human cerebrum. Journal of anatomy 170, 111–23.

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: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Cham. pp. 61–68. doi:10.1007/978-3-319-47157-0_8.

Zhang, J., Gao, Y., Park, S.H., Zong, X., Lin, W., Shen, D., 2017. Structured Learning for 3-D Perivascular Space Segmentation Using Vascular Features. IEEE Transactions on Biomedical Engineering 64, 2803–2812. doi:10.1109/TBME.2016.2638918.

Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R., 2015. Minimum barrier salient object detection at 80 FPS, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1404–1412. doi:10.1109/ICCV.2015.165.

Zhou, X., Ito, T., Takayama, R., Wang, S., Hara, T., Fujita, H., 2016. Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 111–120. doi:10.1007/978-3-319-46976-8_12, arXiv:1608.04117.

Zhu, Y.C., Dufouil, C., Mazoyer, B., Soumaré, A., Ricolfi, F., Tzourio, C., Chabriat, H., 2011. Frequency and location of dilated Virchow-Robin spaces in elderly people: A population-based 3D MR imaging study. American Journal of Neuroradiology 32, 709–713. doi:10.3174/ajnr.A2366.

Zhu, Y.C., Dufouil, C., Soumaré, A., Mazoyer, B., Chabriat, H., Tzourio, C., 2010a. High degree of dilated virchow-robin spaces on MRI is associated with increased risk of dementia. Journal of Alzheimer's Disease 22, 663–672. doi:10.3233/JAD-2010-100378.

Zhu, Y.C., Tzourio, C., Soumaré, A., Mazoyer, B., Dufouil, C., Chabriat, H., 2010b. Severity of dilated virchow-robin spaces is associated with age, blood pressure, and MRI markers of small vessel disease: A population-based study. Stroke 41, 2483–2490. doi:10.1161/STROKEAHA.110.591586.

Zong, X., Park, S.H., Shen, D., Lin, W., 2016. Visualization of perivascular spaces in the human brain at 7T: Sequence optimization and morphology characterization. NeuroImage 125, 895–902. doi:10.1016/j.neuroimage.2015.10.078.

Appendix A. Distance Map

Appendix A.1. Distance Transform

A distance map is an image that shows for every pixel how far away it is from a chosen subset of pixels and was first presented by Rosenfeld and Pfaltz (1968) (Cárdenes et al., 2010; Saito and Toriwaki, 1994). A distance map is computed from a binary image using an operation called a distance transform. The binary image distinguishes between pixels belonging to the background (0) and foreground (1). The distance transform calculates for every pixel the closest distance to the specified foreground.The output is an image with pixels that have a value corresponding to their distance to the chosen subset defined by the binary image. Essentially a distance map is a composition of distance isocontours, each contour containing all pixels that are a certain distance from the foreground (Paglieroni, 1992; Borgefors, 1986; Rosenfeld and Pfaltz, 1966; Grevera, 2007; Wang and Tan, 2013).

To acquire an exact distance map, the distance transform would have to be a global operator. This is an operator that applies the same operation to every pixel independent of its location using (almost) all other pixels in the image. As you can imagine this would be by far too costly. Instead an approximated distance map is computed by using only a pixel's local neighborhood to compute the distance for that pixel. This is based on the assumption that the global distance can be approximated by propagating local distances (Wang and Tan, 2013; Borgefors, 1986; Rosenfeld and Pfaltz, 1966). Various algorithms using different local distance measures have been proposed to approximate the distance map as accurately as possible while also optimizing speed (Rosenfeld and Pfaltz, 1966; Yatziv et al., 2006; Danielsson, 1980).

Appendix A.2. Distance Measures

The definition of the distance in a distance transform greatly effects the resulting distance map. The distance in general is defined as the shortest path between two pixels. Which path is shortest is greatly dependent on the chosen distance measure and corresponding assumed pixel connectivity. Well-known distance measures are the city block distance (also known as Manhattan distance), the chessboard distance and the Euclidean distance (Rosenfeld and Pfaltz, 1966; Vincent, 1993; Fabbri et al., 2008; Jain et al., 1995). These first two measures are solely dependent on the number of pixels that connect one pixel to another. The definition of how pixels are connected vary between these two methods. City block distance assumes 4-connectivity (N_4), meaning that it allows only horizontal and vertical steps between pixels. Pixels at a diagonal angle are therefore defined as a distance of 2 away from each other. For chessboard distance this distance would be 1, because 8-connectivity (N_8) is assumed which also allows diagonal steps (Jain et al., 1995; Rosenfeld and Pfaltz, 1966; Borgefors, 1986). These two distance measures are a relatively rough approximation of global Euclidean distance, however they are less costly in terms of computation (Fabbri et al., 2008; Wang and Tan, 2013). A better approximation is given by using local Euclidean distance defined on the Cartesian discrete plane, meaning horizontal and vertical neighbors are at a distance of 1 and diagonal neighbors are at a distance of $\sqrt{2}$ (Danielsson, 1980; Wang and Tan, 2013).

All of these distance measures are for binary images, they take only the spatial difference into account (Wang and Tan, 2013; Borgefors, 1986; Paglieroni, 1992; Ćurić et al., 2014). To calculate the distance in gray-scale images however it is important to take the intensity values into account as well. Several intensity-weighted distance measures for gray-scale images have been proposed with varying ways of combining the spatial and

intensity values (Strand and Normand, 2012; Levi and Montanari, 1970). Besides weighting, combining the spatial and intensity information by using the intensity as an extra dimension is also possible, resulting for 2D images in the image being viewed as a height map. In other words the image is seen as a curved space defined by (two) spatial coordinates and one intensity coordinate. The shortest path on curved space is referred to as the geodesic distance, as the path is restricted to the top surface of this height map. Intuitively distances between pixels that are connected by flat terrain are shorter than pixels that have hills and valleys in the height map between them (Grazzini et al., 2007; Toivanen, 1996).

Toivanen (1996) proposed two (geodesic) distance measures using this idea, the distance on curved space (DOCS) and the weighted distance on curved space (WDOCS). The spatial distance used in DOCS is chessboard distance. In WDOCS this is local Euclidean distance, which explains why WDOCS is also referred to as Euclidean distance on curved space (EDOCS) (see Figure 2). The corresponding geodesic distance transforms (for gray-scale images) to compute a geodesic map are referred to as distance transform on curved space (DTOCS) and the weighted distance transform on curved space (WDTOCS). Both transforms require a binary image F(x) defining the foreground as well as a gray-scale image G(x). The DOCS between pixel e and its neighboring pixel $x_i \in (N_8(e)$, with intensities $G(e)$ and $G(x_i)$ respectively, is defined as

$$d(e, x_i) = \sqrt{\Big(G(e) - G(x_i)\Big)^2 + 1} \tag{A.1}$$

For WDOCS this is defined as

$$d(e, x_i) = \begin{cases} \sqrt{\Big(G(e) - G(x_i)\Big)^2 + 1} & \text{if } x_i \in N_4(e) \\ \sqrt{\Big(G(e) - G(x_i)\Big)^2 + 2} & \text{if } x_i \in (N_8(e) \setminus N_4(e)) \end{cases} \tag{A.2}$$

Appendix A.3. Computation

Many algorithms for distance transforms of binary images have been proposed. How these methods search for the closest foreground pixel to each background pixel can be roughly summarized in three general approaches. Firstly, ordered propagation handles this by starting at the foreground pixels and iteratively propagating a front with a velocity over the image until all pixels have gotten a value (Yatziv et al., 2006; Wang and Tan, 2013). Secondly independent scanning computes the distance map by dimensional reduction. The distance transform is initially done separately on every column or row of the image, after which the result is combined in various ways to compute the eventual distance map (Paglieroni, 1992; Saito and Toriwaki, 1994; Wang and Tan, 2013). Lastly raster scanning is an approach that passes a mask over the image multiple times sequentially computing and updating the distance for every pixel (Wang and Tan, 2013; Rosenfeld and Pfaltz, 1966; Danielsson, 1980).

Originally DTOCS and WDTOCS proposed by Toivanen (1996) were implemented by adapting the raster scan method proposed by Rosenfeld and Pfaltz (1966). This iterative algorithm consists of two passes over the image. The minimal geodesic distance per pixel is updated sequentially by passing a mask operation over the image first from the left upper corner to the lower right corner (forward pass using the kernel in Figure A.12b) and the second pass goes over the image in reversed order (backward pass using the kernel in Figure A.12c). All background pixels are initiated with the maximal integer number and the foreground pixels with 0. As the mask moves over the image, the new distance value $F^*(e)$ for pixel e is computed using the neighbors defined in the mask. This is done by first calculating the distance $d(e, x_i)$ (as defined in equation A.1 for DTOCS and equation A.2 for WDTOCS) for all neighbors x_i in the mask. Per neighboring pixel this distance $d(e, x_i)$ is added to the already calculated value for this pixel $F^*(x_i)$. The new distance value $F^*(e)$ for pixel e is the minimum value of these calculated distances and the initial value of pixel e $F(e)$

$$F^*(e) = min(F(e), min\{d(e, x_i) + F^*(x_i) \mid i \in mask\}) \tag{A.3}$$

After several iterations the algorithm converges and the optimal approximation of the geodesic distance map is reached (Toivanen, 1996).

Several other methods have been proposed to improve the approximation of geodesic distance maps and the computational efficiency, e.g. using wave-front propagation (Kimmel and Sethian, 1998; Yatziv et al., 2006; Ikonen, 2007; Ikonen and Toivanen, 2007; Cárdenes et al., 2010). Surprisingly the original (or a slightly adapted version of the) algorithm using raster scan as described by (Toivanen, 1996) in 1996 is still often used in methods for medical image analysis (Wang et al., 2018; Criminisi et al., 2008; Kontschieder et al., 2013; Jang et al., 2016; Wei et al., 2012; Zhang et al., 2015; Cerrolaza et al., 2017). Various papers state the method is efficient with an optimal complexity of O(N), relatively accurate and straightforward to implement (Zhang et al., 2015; Wei et al., 2012). Wang et al. (2018) propose a method for interactive segmentation based on this algorithm, clearly showing the algorithm is quite fast.

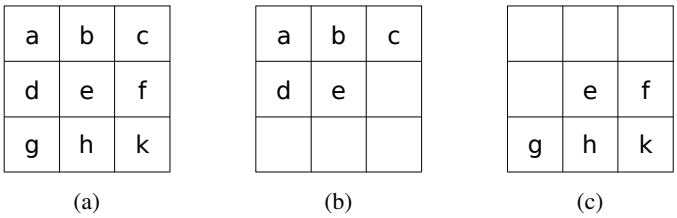


Figure A.12: **3 by 3 kernel for raster scan.** Full kernel (a), split kernel for forward pass (b), split kernel for backward pass (c) (adapted from Toivanen (1996))

Appendix A.4. Application

Many image processing fields like pattern recognition and image analysis use distance maps. Useful applications for distance maps are for instance shape analysis, clustering, k nearest neighbor classification, level set segmentation, handwritten character recognition and connected component analysis (Cuisenaire, 1999; Wang and Tan, 2013; Lantuejoul and Beucher, 1981;

Cárdenes et al., 2010; Holuša and Sojka, 2015). Which distance measure is most suited is dependent on the application.

Examples of applications for distance maps in medical image analysis are centerline detection (Sironi et al., 2014), nerve morphometry prediction (Cuisenaire, 1999) and cell detection (Xie et al., 2018a,b). Specifically geodesic distance maps seem to be useful for a lot of applications in medical image analysis because it takes image context into account. At edges in the intensity image the geodesic distance map will show large differences in values (Wang et al., 2018; Criminisi et al., 2008). This property is especially useful for semi-automatic methods for segmentation (Wang et al., 2018; Criminisi et al., 2008; Bai and Sapiro, 2007). A clear example of this is the method proposed by Wang et al. (2018). This method first computes an initial segmentation using a neural network. The user can refine this segmentation by drawing doodles indicating either background or foreground. These doodles are transformed to geodesic distance maps and together with the input image given to a different neural network. This network refines the segmentation based on this and outputs a new segmentation proposal. This continues until the user is satisfied with the result (Wang et al., 2018). Furthermore (Jang et al., 2016) use a combination of Hough transform and geodesic distance maps for aorta segmentation. Kontschieder et al. (2013) tackle semantic segmentation with a forest-based model using geodesic distance to compute connectivity features. Gaonkar et al. (2015) propose a semi-automatic approach using geodesic distance maps combined with thresholding for tumor volume segmentation (see Figure A.13). Krähenbühl and Koltun (2014) present a method that produces objects proposals in images using critical level sets in geodesic distance maps computed using seeds placed by trained classifiers. An important disadvantage of geodesic distance transform for image segmentation is that is sensitive to noise in the image. Holuša and Sojka (2015) discuss this problem and propose a new geodesic distance that is more robust to noise. Park et al. (2016) mention as a recommendation for enlarged perivascular spaces segmentation that geodesic distance might be useful for capturing the complex patterns of enlarged perivascular spaces.

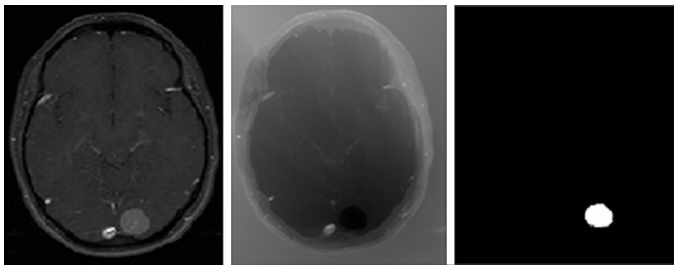


Figure A.13: **Example of segmentation using geodesic distance map.** The original intensity image (a), the geodesic distance map computed with a seed located in the tumor (b) and the resulting segmentation of the tumor after thresholding the geodesic distance map. In the geodesic distance map the tumor is more clearly defined than in the original images, enabling a better segmentation (adapted from Gaonkar et al. (2015))

Appendix B. Perivascular Spaces

Appendix B.1. Anatomy

PVS, also known as Virchow-Robin spaces, surround arteries, arterioles, veins and venules as they enter and emerge from the brain (see Figure 1). The brain is enveloped by three membranes, the dura mater, arachnoid mater and the pia mater. Between the arachnoid mater and the pia mater which covers the cerebral cortex lies the subarachnoid space. Vessels entering the brain from the subarachnoid space or emerging from the brain into the subarachnoid space are enveloped by pia mater. The space between the pia mater and the vessel is referred to as PVS (Zhang et al., 1990; Braffman et al., 1988; Barkhof, 2004; Kwee and Kwee, 2007; Valdés Hernández et al., 2013). In the subarachnoid space vessels are also covered by pia mater. Consequently PVS are separated from the subarachnoid space by the pia mater. Differences exist in structure of PVS and its surroundings in different brain regions and with different types of vessels (Bakker et al., 2016; Hutchings and Weller, 1986).

Controversy exists as to whether PVS are filled with cerebrospinal fluid (CSF) (Ramirez et al., 2016; Potter et al., 2015) and/or interstitial fluid (ISF) (Fanous and Midia, 2007; Öztürk and Aydingöz, 2002). The signal intensity of PVS has been compared to the signal intensity of CSF on T2-weighted (T2w) MR sequences. Visually on all pulse sequences these signal intensities seem to be similar if not identical. However quantitative analysis shows that the signal intensity of PVS is significantly lower than CSF-filled structures in and surrounding the brain. This could indicate that PVS are filled with ISF instead of CSF. However this could also be caused by partial volume effects (Öztürk and Aydingöz, 2002; Kwee and Kwee, 2007). The pia mater has a selective permeability and could affect the composition of CSF that passes through it to perivascular spaces (Bakker et al., 2016).

Appendix B.2. Physiology

PVS are believed to be involved in the clearance of ISF, CSF, metabolic waste and solutes in the brain. The exact (fluid) mechanics of this process are not yet fully understood, research on this is ongoing (Wang and Olbricht, 2011; Zhang et al., 1990; Bacyinski et al., 2017; Fanous and Midia, 2007; Valdés Hernández et al., 2013; Faghih and Sharp, 2018; Ramirez et al., 2016; Cserr and Knopf, 1992). Knowing the exact mechanics could help understanding the pathology of neurodegenerative diseases as well as for research in drug delivery in the brain (Valdés Hernández et al., 2013; Bakker et al., 2016; Bacyinski et al., 2017).

Studies using tracer injections attempt to shed light on where and how the drainiage of CSF and ISF happens. Various pathways and explanations for transport of fluid and solutes have been proposed (Wang and Olbricht, 2011; Faghih and Sharp, 2018; Fanous and Midia, 2007). In general PVS are assumed to be in contact with the cervical lymphatics (Esiri and Gay, 1990; Ramirez et al., 2016). Peristaltic motions of the blood vessel walls are thought to aid in the transport of fluid and solutes in PVS. The observation that fluid drainage stops post-mortem supports this theory (Wang and Olbricht, 2011; Faghih and Sharp, 2018; Iliff et al., 2013; Bakker et al., 2016).

Furthermore PVS are thought to play an important role in immunological and inflammatory responses in the brain (Esiri and Gay, 1990; Etemadifar et al., 2011; Fanous and Midia, 2007; Ramirez et al., 2016; Zhang et al., 1990). PVS appear to be constantly monitored by macrophages in the blood (Groeschel et al., 2006; Bechmann et al., 2001). Moreover PVS supply an important site of interaction between macrophages and lymphocytes as well as for accumulation and migration into the brain (Esiri and Gay, 1990; Fanous and Midia, 2007; Wang et al., 2016a,b).

Appendix B.3. Pathology

Formerly the enlargement of PVS was assumed to be benign (Zhu et al., 2010b; Adams et al., 2015). Recent studies however support the contrary and an increasing amount of research is being done on this emerging neuroimaging marker. PVS have been associated with worse cognition, hypertension, as well as with markers of cerebral small vessel disease namely white matter hyperintensities and lacunar infarctions (Maclullich et al., 2004; Zhu et al., 2010a; Potter et al., 2015; Chen et al., 2011; Zhu et al., 2010b; Charidimou et al., 2013). PVS are seen in individuals of all ages and in the elderly population these lesions are highly prevalent (Zhu et al., 2011, 2010b). With increasing age the number and size of PVS in the brain has been shown to increase (Doubal et al., 2010; Kwee and Kwee, 2007; Dubost et al., 2018b). Additionally a high number of PVS has been associated with many neurological conditions including cerebral small vessel disease, cerebral arteriosclerosis, traumatic brain injury, poststroke depression, Parkinson's disease, incident dementia and Alzheimer's disease (Zhu et al., 2010a; Maclullich et al., 2004; Zhu et al., 2011; Ramirez et al., 2016; Zhu et al., 2010b; Chen et al., 2011; Hurford et al., 2014; Potter et al., 2015; Liang et al., 2018; Cai et al., 2015; Doubal et al., 2010).

The cause and mechanism of the enlarging of PVS is not clear yet. Atrophy of the brain, tissue fibrosis, arterial wall permeability, microvascular or lymphatic obstruction, perivascular demyelination, hypertension and inflammation have been proposed as mechanisms that may contribute to the enlargement of PVS (Chen et al., 2011; Adams et al., 2015; Groeschel et al., 2006).

Understanding the function of PVS and why these spaces become enlarged could help improve treatment of diseases that are associated with PVS (Charidimou et al., 2013; Chen et al., 2011; Ramirez et al., 2016).

Appendix B.4. Visualization

Normal PVS are too small to be noticed on MRI scans at clinical field strengths, however when PVS increase in size they become more visible and quantifiable (Ramirez et al., 2016; Kwee and Kwee, 2007; Doubal et al., 2010; Wardlaw et al., 2013). In other words, the PVS that are visible on MRI are larger than normal PVS, which explains why often any PVS visible on MRI is referred to as an enlarged perivascular space or dilated Virchow-Robin space. However it is not clear yet when PVS are enlarged enough to be clinically significant and the visibility of PVS is dependent on the MRI sequence parameters used for acquisition, which vary per study. Basing the definition

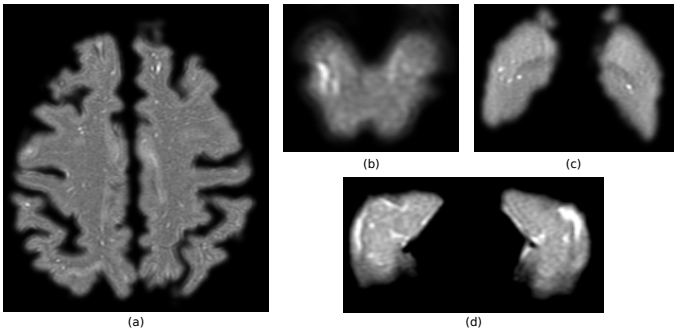


Figure B.14: **Perivascular spaces (PVS) in different brain regions.** T2-contrast axial slices showing the centrum semiovale (a), the mesencephalon (b), the basal ganglia (c) and the hippocampi (d). Hyperintensities are either enlarged perivascular spaces or structures that look similar (adapted from Dubost et al. (2018b))

of the term enlarged perivascular spaces on visibility is therefore not very robust (Wardlaw et al., 2013). A clear example of the inconvenience this entails, is the observation that studies using 3 T MRI scans report a prevalence of 100% of enlarged PVS, while a lot lower prevalences are reported by studies using 1.5 T MRI scans. This is logical as scans acquired at 1.5 T have a lower spatial resolution than scans acquired at 3 T. If all visible PVS without a lower limit in size are seen as enlarged PVS then the spatial resolution of the MRI scans will determine the lower bound of PVS that are taken into account in the study. Clearly these sets will have different lower thresholds and will count different numbers of PVS. For this reason referring to all PVS (visible on MRI or not) as PVS and specifying the sizes of PVS that are examined in the study is recommended (Wardlaw et al., 2013; Adams et al., 2015, 2013)

As PVS follow the course of the vessel they surround, they appear as elongated structures on 3D MRI scans. In a 2D image slice PVS can be round, ovoid or linear dependent on what the orientation of the PVS is with respect to the image slice (Wardlaw et al., 2013). The minimum diameter of PVS is assumed to be 1 mm. However there are different opinions on the maximum diameter of PVS. The most general assumption seems to be 3 mm (Adams et al., 2015; Zhu et al., 2011; Ramirez et al., 2016; Valdés Hernández et al., 2013). PVS appear mainly in the basal ganglia, centrum semiovale, hippocampus and in the mesencephalon (see Figure B.14) (Kwee and Kwee, 2007; Barkhof, 2004; Adams et al., 2015). PVS have a similar intensity to CSF (see Figure B.15) (Ramirez et al., 2016; Kwee and Kwee, 2007). Although PVS can be distinguished on scans acquired at field strengths of 1.5 and 3 Tesla (T) which is used clinically, visibility of PVS is noticeably better on scans acquired at a field strength of 7 T which have a higher spatial resolution and contrast (Bouvy et al., 2016; Lian et al., 2018; Feldman et al., 2018). PVS that are barely enlarged can occasionally be seen on MRI scans at clinical field strengths. 7 T MRI scans due to the increased spatial resolution appear to be able to even visualize PVS that are barely or not enlarged (with a diameter smaller than 1 mm) (Bouvy et al., 2016).

The appearance of PVS on MRI scans bears most resemblance to lacunar infarcts, lacunes and small punctual white

matter hyperintensities (WMH) (Potter et al., 2015; Valdés Hernández et al., 2013; Kwee and Kwee, 2007; Bokura et al., 1998). As all three of these are defined to be larger than 3 mm (in diameter), and PVS are mainly defined with a diameter between between 1 and 3 mm, size seems to be a helpful measure in distinguishing PVS. Furthermore PVS are described as linearly elongated slit-like structures, whereas lacunes and lacunar infarcts seem to be more ovoid or spherical in shape and WMH are said to be flame or cotton wool-like shaped. Lacunar infarcts, lacunes and especially WMH often have more irregular, vague edges in comparison to PVS which have sharp edges. All four types of lesions are hypointense on T1-weighted (T1w) sequences and hyperintense on T2w sequences. However on fluid-attenuated inversion recovery (FLAIR) sequences PVS and lacunes are hypointense as opposed to WMH which are hyperintense on FLAIR. Acute lacunar infarcts are hyperintense on FLAIR or can have a hyperintense rim, older lacunar infarcts can evolve into a lacune. As the intensity range of PVS on T2w overlaps with the other three lesion types this is not a useful measure for identifying PVS (Valdés Hernández et al., 2013; Dubost et al., 2018b; Chen et al., 2011; Jungreis et al., 1988). Especially the shape, location, size, spatial distribution and appearance on different MRI sequences are thought to be important descriptors for distinguishing PVS from mimicks (Valdés Hernández et al., 2013; Boespflug et al., 2018)

Determining whether a faintly enlarged PVS is large enough to count as an PVS is difficult as well, especially on 7 T MRI scans due to the improved visibility of not or barely enlarged PVS (Bouvy et al., 2016; Potter et al., 2015). Furthermore motion artifacts can also look similar to PVS in MRI scans (Dubost et al., 2018b; Park et al., 2016).

Appendix B.5. Assessment

MRI is without a doubt an invaluable tool for research on PVS. However to be able to study PVS associations, a way of measuring is needed to compare PVS burden in the brain. Until recently most studies used a visual scoring system, categorizing PVS burden in an image into in general 4 to 6 burden levels (Doubal et al., 2010; Maclullich et al., 2004; Hurford et al., 2014; Chen et al., 2011; Potter et al., 2015). Almost every study had its own method of assessing PVS burden, making it difficult to compare studies. Efforts have been made to establish a more general and robust way of evaluating burden of PVS (Ikram et al., 2017; Adams et al., 2015; Wardlaw et al., 2013; Doubal et al., 2010; Potter et al., 2015; Valdés Hernández et al., 2013; Ballerini et al., 2018).

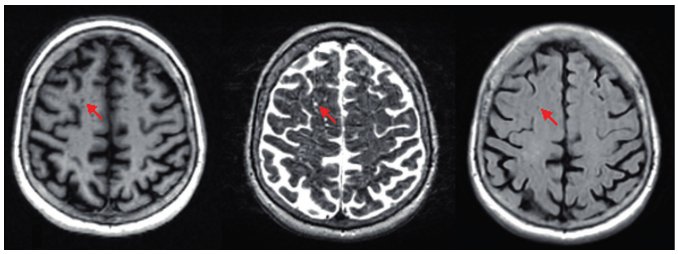


Figure B.15: **Appearance of perivascular spaces (PVS) on different MR sequences.** Axial slices showing the centrum semiovale on a T1-weighted (left) T2-weighted (middle) and fluid-attenuated inversion recovery (FLAIR) MRI scan (right). An example of an PVS is indicated by the red arrow (adapted from Valdés Hernández et al. (2013))

Visual scoring systems are a fast way of PVS assessment. However these scales are based on subjective classification. Furthermore clustering the PVS burden into few categories results in floor and ceiling effects. Evidently these scales do not directly indicate any information on location, morphology or volume of PVS (Wang et al., 2016b; Ballerini et al., 2018; Boespflug et al., 2018; Ramirez et al., 2015). Established visual scoring systems that currently appear to be most used are the Patankar scale and the Potter scale (Patankar et al., 2005; Ramirez et al., 2015; Potter, 2011; Wang et al., 2016b; González-Castro et al., 2017).

Other proposed measures of PVS burden are counting the number of PVS per slice or per full brain region, pixel-wise binary labels with either a dot per PVS or segmentations of the PVS (in order of containing increasing information about PVS). Besides containing more information about PVS, these measures also pose a more objective way of assessing PVS burden. However, as PVS are difficult to distinguish from other structures like lacunes (see section Appendix B.4), these measures still do suffer from some subjectivity (Adams et al., 2015; Dubost et al., 2018b; Valdés Hernández et al., 2013; Wang et al., 2016b). An important disadvantage of these measures is that obtaining them manually is very time-consuming, especially for the basal ganglia and the centrum semiovale which are large brain regions. This could be somewhat alleviated by the observation that for these two regions the number of PVS found in one slice of is highly correlated with the number of PVS in the whole brain region (Adams et al., 2015). However this correlation was computed on a set of only 40 scans, so additional studies that further examine this correlation would be favorable. Also further research on which slice best represents the full brain region would be beneficial. As these are large brain regions, annotating only one slice to describe PVS burden in the full volume would be very useful and considerably decrease the annotation time (Adams et al., 2015).

Appendix B.6. Computational Methods

Various studies have suggested that computational methods could improve reliability and generalization of PVS measures while decreasing the time this takes considerably (Valdés Hernández et al., 2013; Adams et al., 2015; Park et al., 2016; Boespflug et al., 2018; Dubost et al., 2018b). In line with this idea various (semi-)automated methods have been proposed to compute these measures of PVS burden. Three important distinctions can be made in these methods. Firstly some methods are semi-automated and others automated, with the former requiring some user interactions and the latter requiring none (Ballerini et al., 2018). Secondly methods differ in being supervised or unsupervised, with unsupervised meaning methods that require no labels and supervised meaning methods that require a dataset with ground truth labels to learn from (Goodfellow et al., 2016). Lastly, the PVS measures computed by these methods vary between classification, quantification (number of PVS per slice

or brain region), detection (location of PVS) and segmentation (location and morphology). Additionally methods have been proposed that aim to enhance MRI scans to improve visibility of PVS and with that facilitate assessment of PVS burden. The most recent and promising methods will be discussed per PVS measure.

Most of the currently proposed (semi-)automated methods focus on segmenting PVS. Obtaining the segmentations of PVS is most ideal in terms of information that it contains, as it includes information on quantity, location and morphology. Obtaining ground truth on the other hand for development and evaluation of these methods is time-consuming as it requires a pixel-wise ground truth. Various semi-automatic methods have been proposed using for instance Frangi's vessel enhancing filter (also referred to as vesselness filter) (Frangi et al., 1998) and adaptive thresholding (Ramirez et al., 2015; Wang et al., 2016b; Wuerfel et al., 2008; Zong et al., 2016). However these methods still require some user interaction, which can lead to inter-observer variations and especially for large datasets can become excessively time-consuming (Ballerini et al., 2018). (Ballerini et al., 2018) use ordered logit models to optimize parameters for vesselness filtering based on visual ratings of PVS burden of the images. Boespflug et al. (2018) propose a segmentation method that also evaluates morphologic features per PVS. They use voxel-wise regression based on the intensities in four MRI modalities, T1w, T2w, FLAIR and proton density-weighted (PDw). The fact that they need four different MR modalities restricts the applicability of this method however. (Lian et al., 2018) developed a method using a fully convolutional neural network. The network is given the original image and filtered image that is enhanced with respect to tabular structures and outputs a probability map. Resulting probability maps are recursively incorporated into the network to further refine the output of the model, which can be iterated until convergence. This method is developed on 7 T MRI scans. Zhang et al. (2017) and Park et al. (2016) also propose methods for 7 T MRI scans both using region proposal, random forest and vessel enhancing, the former using vascular features for classification of the regions and the latter using randomized Haar features. Cai et al. (2015) segment PVS using k-means clustering also on 7 T MRI scans. Although these methods are promising for PVS segmentation, MRI scans at 7 T have a higher spatial resolution than MRI scans at clinical field strength (1.5 T or 3 T), which limits the applicability of these methods for clinical use (González-Castro et al., 2017; Ballerini et al., 2018). Uchiyama et al. (2008) perform region proposal by combining morphological operations with intensity thresholding and subsequently classify the proposed regions as PVS or lacunar infarcts based on properties like size and location.

Furthermore several methods have been proposed for automated detection, quantification and classification of PVS (burden).

Dubost et al. (2017) uses a convolutional neural network that is trained to regress the number of PVS. At inference the network is used to predict the lesion count and to generate a heatmap. The heatmap is obtained by removing the last layer (global pooling), which changes the output of the network from the predicted count to a heatmap image that can be used for detection. The heatmap is thresholded in such manner that the number of connected components in the heatmap is equal to the predicted count. This method is promising because it can be trained on the number of PVS and it will output the location of the PVS, which would take much longer to annotate. This is also apparent in the data that is used, as the count labeled dataset for training is very large (over a thousand) and the detection labeled set used for testing is only 30 images. However all mentioned methods in this section up till now have been evaluated on relatively small sets, as pixel-wise annotations are so time-consuming to produce. Many methods are developed and evaluated on barely 20 images. Especially for learning based models this is a problem because if there are only few images available for training, the method will be prone to overfit. Using cross-validation and training on patches helps this somewhat. Testing these methods on larger datasets would be very useful to evaluate their true potential and robustness. The same authors of the PVS detection method also propose a different convolutional neural network for quantification of PVS in the basal ganglia Dubost et al. (2018a). Later they extend their work to other brain regions (Dubost et al., 2018b). González-Castro et al. (2016a) proposed a supervised method for binary classification of PVS burden (absent/mild PVS to moderate/severe PVS) using a support vector machine combined with texture descriptors (González-Castro et al., 2016a) and combined with bag of visual words based descriptors descriptors (González-Castro et al., 2016b). In a later paper comparing these methods the same authors show that the combination of a support vector machine with bag of visual words based descriptors performs best and its performance is close to human performance (González-Castro et al., 2017). As the ground truth needed for this binary classification problem is less time-consuming, the dataset used in this paper is also larger (264 images). However visual scoring of PVS is more prone to inter-observer variability (Ballerini et al., 2018).

To improve performance of identifying PVS, several methods have been proposed for enhancing PVS by highlighting thin tubular structures. Uchiyama et al. (2008) and Hou et al. (2017) show unsupervised methods can be used to improve visibility of PVS. Uchiyama et al. (2008) uses white top hat transformation to emphasize tubular structures. Hou et al. (2017) use Haar transform of non-local cubes to suppress noise in the image and to intensify details of PVS.

Appendix C. Previous Experiments

Appendix C.1. Varying Loss Functions

Optimizing the network to detect PVS was not straightforward. The loss that we started with, namely mean squared error (MSE), does not directly optimize the detection of the PVS. Instead it optimizes the voxelwise regression of a geodesic distance map (GDM).

The first networks trained using MSE seemed to show that mainly the background was being optimized instead of the detection of the PVS. In multiple experiments models with a higher

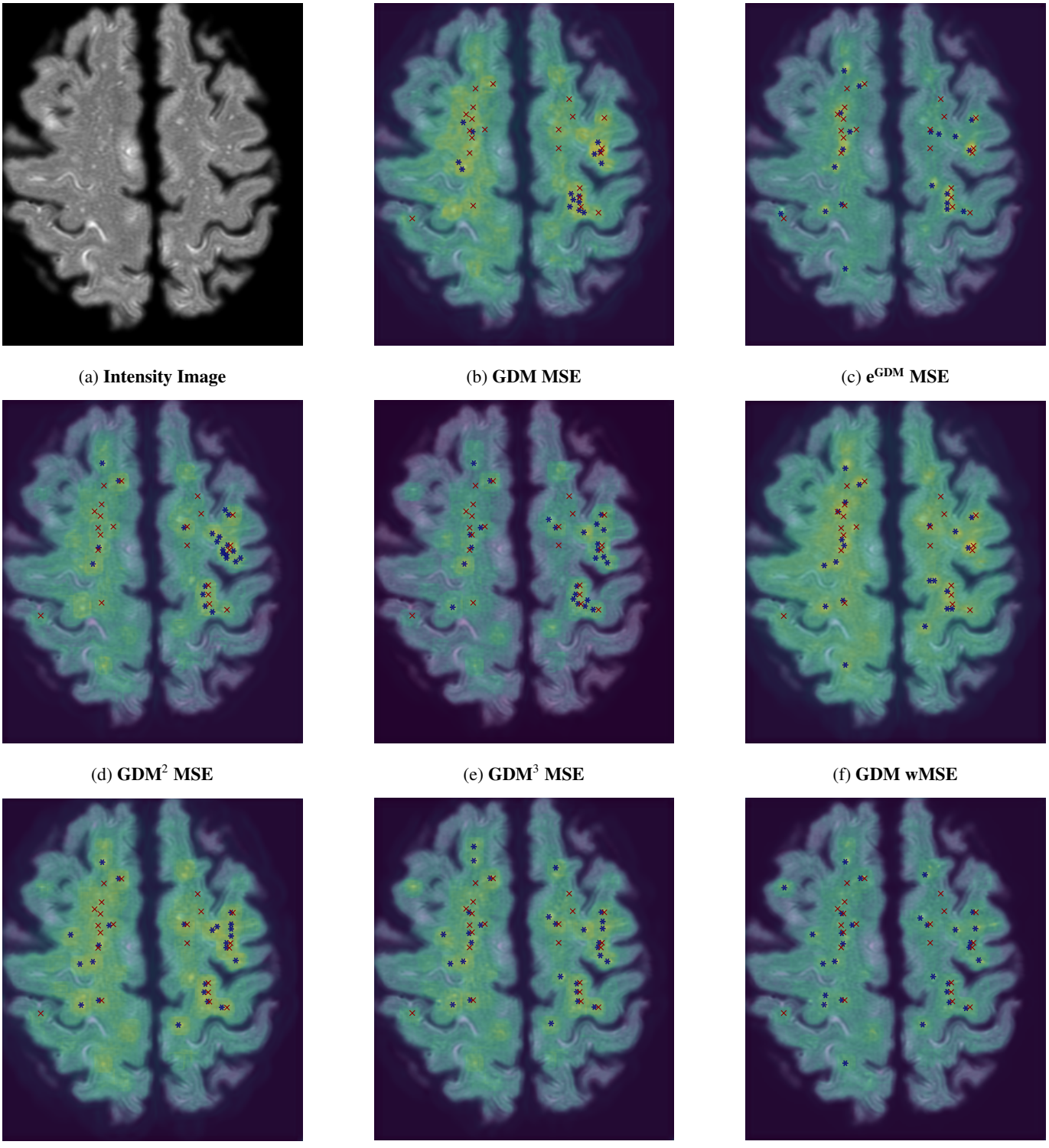
loss turned out to perform better than models with a lower loss on the detection objective (when looking at the Free-Response Receiver Operating Characteristic curve (FROC curve)). It was clear the optimization was not working. We experimented with different losses and ground truth images to shift the focus more to detecting the PVS. For instance clipping the GDM so only high values were kept, so only using GDM values surrounding the PVS and elsewhere zero. We also tried optimizing the CNN with multiple losses and label images, first the GDM then the segmentation etc. Most of these approaches did optimize the CNN to detect PVS, but the detection performance was often not following the same trend as the loss while training the network. Tensorboard was used to monitor training. Only looking at the loss did not paint the full picture of what was happening, what the network was focusing on. Writing the output of the network to Tensorboard every 10 epochs, made it possible to follow the progress of the network visually, making it possible to see if what the network was focusing on. Additionally every 30 epochs the model with the lowest loss on the validation set was saved. After the network was converged, all saved models were evaluated on FAUC. This way we could see how the performance of the model varied during optimization of the network.

For the segmentation approach we tried first with ReLU and BCE and DICE. However the CNN failed to learn anything. When we tried ELU it did learn with BCE. ELU has a normalizing effect which helped the network to learn.

Based on the experiments on this subset of the training data we set up the experiments described in the paper.

Appendix D. Extra Visualizations of the Results

Figure D.16 shows the predictions for another image of the test set.

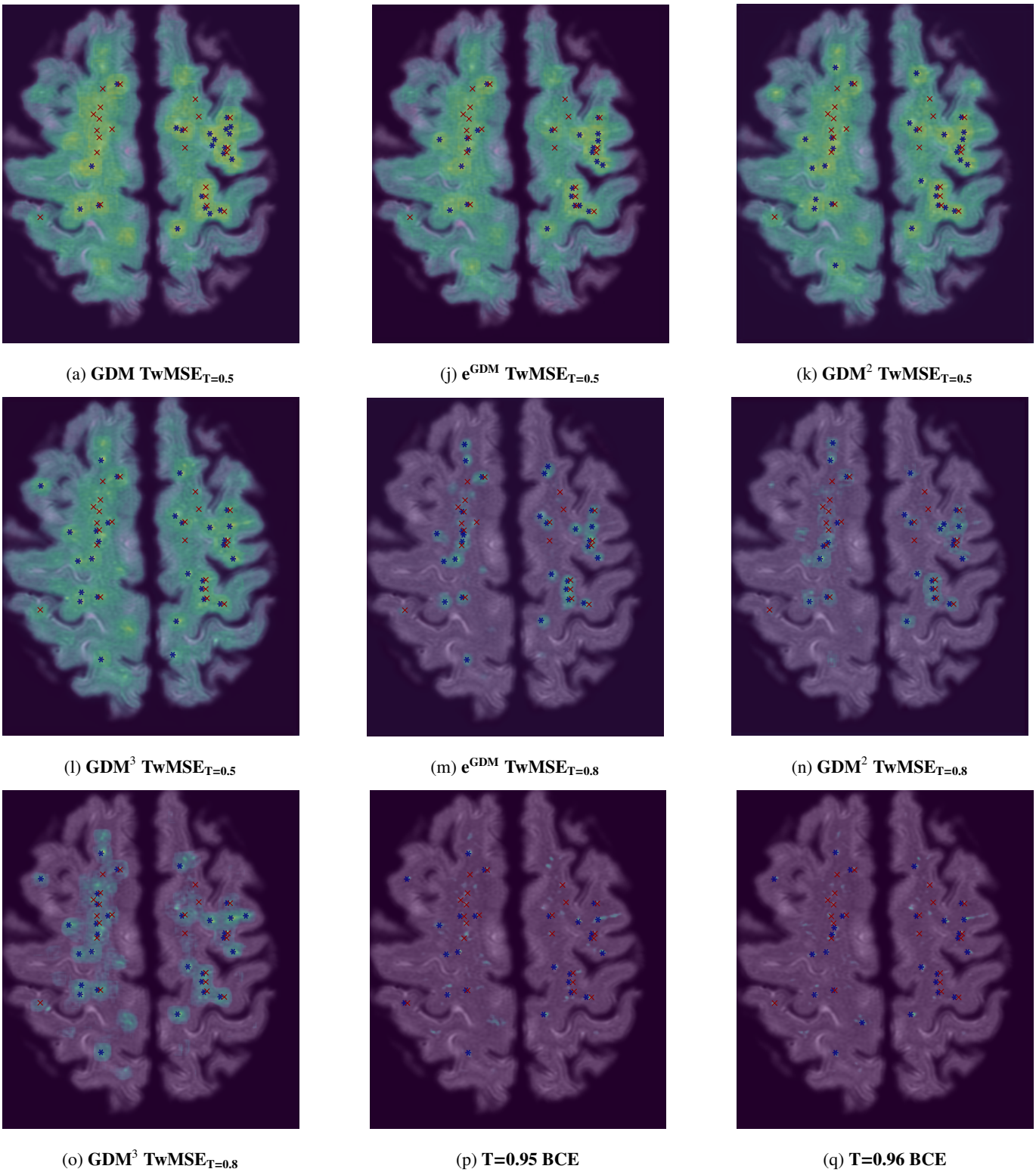


(a) **Intensity Image** (b) **GDM MSE** (c) **e^{GDM} MSE**

(d) **GDM^2 MSE** (e) **GDM^3 MSE** (f) **GDM wMSE**

(g) **e^{GDM} wMSE** (h) **GDM^2 wMSE** (i) **GDM^3 wMSE**

Figure D.16: **Predictions on another image of the test set.** For every method the predicted map is overlaid on the intensity image given as input to the model. Red crosses are the ground truth given by the expert rater (so no shifting of dots during inference). The blue asterisks indicate the proposed detections at the threshold closest to the performance of the expert rater on the test set. The overlaid prediction map is shown in a sequential perceptually uniform color scale that ranges from purple for the lowest value in the image to yellow for the highest value. The contrast in the predicted maps illustrate the spread of the values in the predicted maps. Note that the predicted maps are not scaled, so the range of values in predicted maps may vary. Learning failed for the two CNNs using DSC loss as well as for the CNN that was supposed to predict the GDM using $\text{TwMSE}_{T=0.8}$ loss. For this reason the visualizations for these CNNs are missing.



(a) **GDM TwMSE$_{T=0.5}$** (j) **e^{GDM} TwMSE$_{T=0.5}$** (k) **GDM^2 TwMSE$_{T=0.5}$**

(l) **GDM^3 TwMSE$_{T=0.5}$** (m) **e^{GDM} TwMSE$_{T=0.8}$** (n) **GDM^2 TwMSE$_{T=0.8}$**

(o) **GDM^3 TwMSE$_{T=0.8}$** (p) **T=0.95 BCE** (q) **T=0.96 BCE**

Figure D.16: **Predictions on another image of the test set (continued).** For every method the predicted map is overlaid on the intensity image given as input to the model. Red crosses are the ground truth given by the expert rater (so no shifting of dots during inference). The blue asterisks indicate the proposed detections at the threshold closest to the performance of the expert rater on the test set. The overlaid prediction map is shown in a sequential perceptually uniform color scale that ranges from purple for the lowest value in the image to yellow for the highest value. The contrast in the predicted maps illustrate the spread of the values in the predicted maps. Note that the predicted maps are not scaled, so the range of values in predicted maps may vary. Learning failed for the two CNNs using DSC loss as well as for the CNN that was supposed to predict the GDM using TwMSE$_{T=0.8}$ loss. For this reason the visualizations for these CNNs are missing.