SegAE: Unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder

Hans E. Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, Lotta M. Ellingsen

PII: S2213-1582(19)30432-2
DOI: https://doi.org/10.1016/j.nicl.2019.102085
Reference: YNICL 102085

To appear in: NeuroImage: Clinical

Received date: 5 July 2019
Revised date: 30 September 2019
Accepted date: 7 November 2019

Please cite this article as: Hans E. Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, Lotta M. Ellingsen, SegAE: Unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder, NeuroImage: Clinical (2019), doi: https://doi.org/10.1016/j.nicl.2019.102085

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)
Highlights

- A convolutional neural network designed for unsupervised segmentation of tissue and white matter hyperintensities (WMHs) in brain MRIs.

- Evaluations were conducted on two distinct datasets comprising data from six different scanners, all with ground truth manual WMH label masks.

- The method compares favorably to existing WHM segmentation methods and is fast and robust to highly variable WMH lesion load and atrophy in the brains of elderly subjects.
SegAE: Unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder

Hans E. Atlason\textsuperscript{a}, Askell Love\textsuperscript{b,c}, Sigurdur Sigurdsson\textsuperscript{d}, Vilmundur Gudnason\textsuperscript{b,d}, Lotta M. Ellingsen\textsuperscript{a,c,*}

\textsuperscript{a}Dept. of Electrical and Computer Engineering, University of Iceland, Reykjavik, Iceland
\textsuperscript{b}Dept. of Medicine, University of Iceland, Reykjavik, Iceland
\textsuperscript{c}Dept. of Radiology, Landspitali - University Hospital, Reykjavik, Iceland
\textsuperscript{d}The Icelandic Heart Association, Kopavogur, Iceland
\textsuperscript{e}Dept. of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA

Abstract

White matter hyperintensities (WMHs) of presumed vascular origin are frequently observed in magnetic resonance images (MRIs) of the elderly. Detection and quantification of WMHs is important to help doctors make diagnoses and evaluate prognosis of their elderly patients, and once quantified, these can act as biomarkers in clinical research studies. Manual delineation of WMHs can be both time-consuming and inconsistent, hence, automatic segmentation methods are often preferred. However, fully automatic methods can be challenging to construct due to the variability in lesion load, placement of lesions, and voxel intensities. Several state-of-the-art lesion segmentation methods based on supervised Convolutional Neural Networks (CNNs) have been proposed. These approaches require manually delineated lesions for training the parameters of the network. Here we present a novel approach for WMH segmentation using a CNN trained in an unsupervised manner, by reconstructing multiple MRI sequences as weighted sums of segmentations of WMHs and tissues present in the images. After training, our method can be used to segment new images that are not part of the training set to provide fast and robust segmentation of WMHs in a matter of seconds per subject. Comparisons with state-of-the-art WMH segmentation methods evaluated

\textsuperscript{*}Corresponding author: Lotta M. Ellingsen

Email address: lotta@hi.is (Lotta M. Ellingsen)
on ground truth manual labels from two distinct data sets and six different scanners indicate that the proposed method works well at generating accurate WMH segmentations without the need for manual delineations.

Keywords: CNN, segmentation, white matter hyperintensity, brain, autoencoder, deep learning

1. Introduction

White matter lesions that appear hyperintense in T2-weighted (T2-w) and Fluid Attenuated Inversion Recovery (FLAIR) images, and can appear hypointense in T1-weighted (T1-w) images, are frequently observed in magnetic resonance images (MRIs) of the elderly. They are often attributed to cerebral small vessel disease [1, 2] and termed white matter hyperintensities (WMHs) of presumed vascular origin [2]. WMHs of presumed vascular origin are generally associated with cognitive decline and dementia, such as Alzheimer’s disease and vascular dementia, or a mixture thereof [3, 4].

Detection and quantification of WMHs could have clinical usefulness in diagnostic workup of patients with mild dementia, and to evaluate the patient’s prognosis. Such measurements could also be used as a biomarker for neurodegenerative diseases in clinical research studies [5, 1]. In any case, failure to account for WMHs in automatic segmentation methods can interfere with the segmentation of other brain structures, and thus, it is critical to be able to robustly identify these features [6].

The currently accepted gold standard in WMH segmentation is manual delineation by an expert in neuroanatomy. However, human raters can have great intra- and inter-rater variability [7] and acquiring such delineations is both time-consuming and expensive, making it impractical for analysis in large-scale studies. Automatic segmentation methods are necessary for such studies and they can be broadly classified into supervised and unsupervised methods. Supervised methods, most notably based on Convolutional Neural Networks (CNNs), have recently achieved state-of-the-art results on various datasets [7, 8, 9, 10]. A frequently used CNN architecture is the U-net [11], which is a fully convolutional network with skip connections between the downsampling and upsampling paths. Supervised CNNs usually need a considerable amount of manually delineated lesion labels for training to capture the possible lesion variation in unseen images. This can be a major drawback in these methods, since new manually delineated segmentations
are often needed to segment new data sets from different sites due to pulse sequence or scanner differences, and such delineations may be impractical to acquire. Attempts to reduce the number of manually delineated masks needed for training include transfer learning [12], and generative adversarial networks (GANs) [13, 14].

Unsupervised methods typically involve modeling of MRI brain tissue intensities. These include methods that obtain WMH lesions as outliers of tissue segmentation [15] and approaches that use specific features of lesions, such as voxel intensity and appearance [15, 16, 17]. Clustering or unmixing methods could potentially be used on a per image basis if a given image has enough WMH lesion load [18]. One cluster may then correspond to WMHs in the brain. However, the number of WMH lesions and their location can vary greatly between subjects, and in the case of an image with no lesions, no cluster would correspond to the lesion class. Furthermore, modelling tissue intensities can be challenging because tissue intensities of MRIs are not always consistent within the image, e.g., due to inhomogeneity artifacts and partial volume effects. FLAIR images are the structural sequence from which WMHs are usually most easily distinguished [1], however, various artifacts or poor skull-stripping can lead to high-intensity regions in FLAIR images [19] that could potentially be incorrectly classified as WMHs. Another unsupervised approach that has been proposed in the literature is to detect WMH lesions as outliers of pseudo-healthy synthesized images [20, 21]. A training data set with healthy brains (no lesions) is required to model normality in these approaches, such that lesions can be detected either as outliers or as results of large reconstruction errors [20, 21]. This is usually not the case when analyzing brain MRIs of subjects older than 65 years old, where around 95% of the population will be expected to have WMHs [22].

A combination of linear unmixing and a neural network autoencoder has been proposed in hyperspectral unmixing of remote sensing images [23]. The purpose of these methods is to simultaneously find the amount of materials (such as water, grass, soil, etc.) in every pixel of the image and its contribution to the image intensity. By viewing various MRI sequences as "multispectral data" and individual brain tissues as different materials [such as WMHs, white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF)], one can adopt such strategies into medical imaging. In our proposed segmentation method we model the intensities of multiple MRI sequences as weighted sums of the segmentations of materials present in the MRIs, as estimated by a convolutional autoencoder from the corresponding MRI.
In hyperspectral unmixing, the number of image channels is usually much higher than the number of materials to be estimated, however, in the case of MRI, fewer MR sequences — or MRI modalities — are available to restrict this ill-posed inverse problem; hence, a regularization is needed. Our proposed CNN has a U-net like architecture, but with an additional linear layer and parameter constraints to perform linear unmixing. This allows the network to generalize the unmixing of materials from a set of training data. The network is trained using a scale-invariant cost function with regularization to determine the materials from which to reconstruct the MRIs. The training images are inhomogeneity corrected during the training phase, such that the CNN learns to segment new images in presence of inhomogeneity artifacts. After training the CNN autoencoder on a training set with a sufficient lesion load, it can be used to directly segment images that were not part of the training set. The segmentations are consistent for new images regardless of lesion load and location. We will hereafter refer to the proposed method as the Segmentation Auto-Encoder (SegAE).

A preliminary version of SegAE was recently published in conference format [24]. Here we present substantial improvements to this prior work by means of: 1) A scale-invariant loss function and a regularizer, 2) more MR sequences contributing to the calculation of the loss function, 3) an inhomogeneity correction performed during the training phase; and 4) a more extensive evaluation of the method on two data sets from 6 distinct scanners, all with ground truth manual lesion labels. Furthermore, a comparison with the preliminary version is presented as Supplementary materials.

2. Materials

Two data sets were used for the evaluation of SegAE; MRIs from the AGES-Reykjavik study [25], and the WMH challenge [8] initiated at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017. We note that the MRIs in the WMH challenge originate from 5 different scanners.

2.1. The AGES-Reykjavik data set

The Age, Gene/Environment Susceptibility-Reykjavik Study (AGES-Reykjavik) was initiated in 2002 and was designed to examine risk factors, including
genetic susceptibility and gene/environment interaction, in relation to dis-
ease and disability in old age [26]. The AGES-Reykjavik Study cohort com-
prises 5764 participants (female and male, age 66-93 at first visit), 4811 of which underwent brain MRI [25]. The MRIs were acquired using a ded-
icated General Electrics 1.5-Tesla Signa Twinspeed EXCITE system with a multi-channel phased array head cap coil. T1-w three dimensional (3D) spoiled gradient echo sequence (time to echo (TE): 8 ms, time repetition (TR): 21 ms, flip angle (FA): 30°, field of view (FOV): 240 mm; 256×256 matrix) with 0.94×0.94×1.5 mm³ voxel size and 110 slices; Proton Density (PD)/T2-w fast spin echo sequence (TE1: 22 ms, TE2: 90 ms, TR: 3220 ms, echo train length: 8, FA: 90°, FOV: 220 mm²; 256×256 matrix); and FLAIR sequence (TE: 100 ms, TR: 8000ms, time from inversion (TI): 2000 ms, FA: 90°, FOV: 220 mm; 256×256 matrix) with 0.86×0.86×3.0 mm³ voxel size and 54 slices.

For developmental purposes, we randomly selected 60 subjects from the cohort; 30 subjects for training, 5 for validation of model parameters, and 25 for testing. The developmental set consists of images from a second visit acquired 5 years later than the first visit on average. The WMHs in the test images were manually annotated by an experienced neuroradiologist to be used as ground truth data. The images used for validation were used to determine model architecture and hyperparameters based on visual inspection.

2.2. The WMH challenge data

We submitted our method to the WMH challenge [8], initiated at MICCAI 2017. This challenge aims to provide a benchmark for automatic segmen-
tation of WMHs of presumed vascular origin and remains open and ongoing1. The publicly available training set includes 60 cases from 3 different scanners, while the challenge organizers keep 110 cases from 5 different scanners hidden for evaluation. The WMH challenge only provides T1-w and FLAIR sequences. Table 1 shows an overview of how the data set is separated into training and test sets. Table 2 shows scanning parameters for the 5 scanners.

2.3. Preprocessing

AGES-Reykjavik: Images were preprocessed using standard prepro-
cessing procedures: Resampling to 0.8×0.8×0.8 mm³ voxel size, rigid regis-

1https://wmh.isi.uu.nl/
Table 1: Overview of the WMH challenge data set, showing how the 170 cases from 5 scanners are separated into training (Tr.) and test (Te.) sets.

| Institute          | Scanner                        | Tr. | Te. |
|--------------------|--------------------------------|-----|-----|
| UMC Utrecht        | 3 T Philips Achieva            | 20  | 30  |
| NUHS Singapore     | 3 T Siemens TrioTim           | 20  | 30  |
| VU Amsterdam       | 3 T GE Signa HDxt              | 20  | 30  |
|                    | 1.5 T GE Signa HDxt            | 0   | 10  |
|                    | 3 T Philips Ingenuity (PET/MR) | 0   | 10  |

Table 2: Scanning parameters for the WMH challenge data set, comprising data from 3 sites and 5 different scanners.

| Scanner   | Sequence | TR[ms] | TE[ms] | TI[ms] | Voxel size[mm$^3$] | slices |
|-----------|----------|--------|--------|--------|--------------------|--------|
| Utrecht   | 3D T1-w  | 7.9    | 4.5    | -      | 1.00 × 1.00 × 1.00 | 192    |
|           | 2D FLAIR | 11.000 | 125    | 2,800  | 0.96 × 0.95 × 3.00 | 48     |
| Singapore | 3D T1-w  | 2,300  | 1.9    | 900    | 1.00 × 1.00 × 1.00 | N/A    |
|           | 2D FLAIR | 9,000  | 82     | 2,500  | 1.00 × 1.00 × 3.00 | N/A    |
| AMS GE3T  | 3D T1-w  | 7.8    | 3.0    | -      | 0.94 × 0.94 × 1.00 | 176    |
|           | 3D FLAIR | 8,000  | 126    | 2,340  | 0.98 × 0.98 × 1.20 | 132    |
| AMS GE1.5T| 3D T1-w  | 12.3   | 5.2    | -      | 0.98 × 0.98 × 1.50 | 172    |
|           | 3D FLAIR | 6,500  | 117    | 1,987  | 1.21 × 1.21 × 1.30 | 128    |
| AMS PETMR | 3D T1-w  | 9.9    | 4.6    | -      | 0.87 × 0.87 × 1.00 | 180    |
|           | 3D FLAIR | 4,800  | 279    | 1,650  | 1.04 × 1.04 × 0.56 | 321    |

For improved inhomogeneity correction in presence of WMHs and enlarged ventricles, the inhomogeneity correction was integrated into the method, as discussed in detail in Sections 3.3 and 3.4.

**WMH challenge**: Resampling of the WMH challenge data to 3 mm in the transversal direction and alignment of the 3D T1-w images to the FLAIR images was performed by the challenge organizers as described in [8]. Since the resolution of the training data and the manually delineated test data needs to be the same, we did not alter the resolution of any of the WMH challenge data. We performed skull removal of the training data set with MONSTR, however, for skull removal of unseen images in the testing phase (performed by the WMH challenge team), we developed a skullstripping U-
net that was trained on the MONSTR brainmasks derived from the training
set (see Supplementary materials). As for the AGES-Reykjavik data set,
inhomogeneity correction was integrated into the segmentation method (see
Section 3.3).

3. Methods

3.1. CNN architecture

The proposed method, SegAE, is an autoencoder with fully convolutional
layers on three resolution scales. The input into SegAE consists of large
three-dimensional (3D) patches of MRI sequences, such as FLAIR, T1-w,
and T2-w images (see Section 3.5 for details on the training procedure). The
autoencoder is constrained to reconstruct the corresponding image patches
with a linear unmixing model,

$$\hat{Y}_c = \sum_{i=1}^{M} w_{i,c} S_i,$$

where $\hat{Y}_c$ is one channel of the output, $w_{i,c} \in \mathbb{R}_{\geq 0}$ are the weights, $S_i$ is
the soft segmentation of materials (such as WMHs, WM, GM, CSF and
meninges), $M$ is the number of materials to be estimated, $S_i \geq 0$ and
$\sum_{i=1}^{M} S_i = B$, where $B$ is a binary brainmask (1 for voxels on the brain,
0 for voxels outside the brain).

The non-negativity constraint and the sum-to-one constraint of $S$ are
enforced with a Softmax activation function. A patch-wise brainmask ob-
tained by binarizing the input patches is applied after the Softmax function.
The weighted sum is implemented with a 1x1x1 convolutional layer that is
constrained to have non-negative weights and zero bias. With appropriate
regularization (see Section 3.2) , the Softmax-layer outputs a soft segmenta-
tion of the materials present in the images.

The autoencoder consists of 3D convolutional layers followed by leaky
rectified linear units (LReLU) activation functions and batch normalization
layers. Downsampling is performed with $2 \times 2 \times 2$ strided convolutions, and
$2 \times 2 \times 2$ upsampling is performed to obtain an output of the same size
as the input. Skip connections are added between activations of the same
spatial resolution from the downsampling to the upsampling paths. The
CNN architecture is demonstrated in Figure 1.
Figure 1: The proposed convolutional autoencoder architecture. The input comprises large 3D patches from different MRI sequences (FLAIR, T1-w, and T2-w are shown here). The final convolutional layer is restricted to have non-negative weights and zero bias for the reconstruction of the output patches $\hat{Y}$ to be a weighted sum of the Softmax outputs $S$. The number of output channels (one for each MRI sequence used) is denoted with $C$ ($C = 3$ in this case), and the number of materials to be estimated from the images is denoted with $M$ ($M = 5$ in this case).

3.2. Loss and regularization

The Cosine proximity function,

$$f(y, \hat{y}) = \frac{y \cdot \hat{y}}{|y||\hat{y}|},$$

(2)

is used to construct a scale invariant loss function between the true patches $Y$ and the predicted patches $\hat{Y}$:

$$L(Y, \hat{Y}) = -\frac{1}{C} \sum_{c=1}^{C} (f(\text{vec}(Y_c), \text{vec}(\hat{Y}_c)) + f(\text{vec}(K * Y_c), \text{vec}(K * \hat{Y}_c))),$$

(3)
where $C$ is the number of channels in $Y$ and $\hat{Y}$, $K$ is the 3D discrete Laplace operator

$$K_{1,3} = \begin{bmatrix}
0 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix},
K_2 = \begin{bmatrix}
0 & 1 & 0 \\
1 & -6 & 1 \\
0 & 1 & 0
\end{bmatrix},$$

and $*$ denotes a convolution. Using the differential operator $K$ in the loss function was found to improve robustness to the slowly varying tissue inhomogeneity.

Reconstructing the MRI sequences as weighted sums of the materials present in the images is an ill-posed inverse problem, since we have fewer MRI sequences than materials of interest, and hence, a regularization is needed. For this we add an activity regularization term to the loss function that penalizes the sum of Cosine proximity between the Softmax outputs,

$$\Omega(S) = \frac{\alpha}{M} \sum_{i=1}^{M} \sum_{j=1}^{M} f(\text{vec}(S_i), \text{vec}(S_j)),$$

(4)

where $\alpha$ is the regularization parameter.

### 3.3. Inhomogeneity correction

A disturbance of the field homogeneity in MR scanners leads to low frequency signal artifacts in MRIs, which can make intensities of the brain tissues and WMHs overlap substantially. A widely used state-of-the-art method for inhomogeneity correction is the N4 bias correction method [29]. We observed that when N4 was directly applied to the FLAIR images (using 125 mm spline distance), it caused a substantial degradation of the lesion contrast in FLAIR images with a large lesion load [see Figure 2 (c) and a more detailed comparison in Supplementary materials]. Hence, to avoid this degradation, we alternated between using N4 bias correction and tissue segmentation to obtain “pure-tissue” probability masks, as suggested in [30]. This improved the N4 bias correction, which in turn improved the next iteration of tissue segmentation. This iterative inhomogeneity correction was performed as follows:

We used SegAE to obtain a soft segmentation of tissues and WMHs, and created a pure-tissue probability mask using Softmax outputs that correspond to CSF, GM, and WM (excluding WMHs and meninges). Then we applied N4 bias correction using the pure-tissue probability mask so regions
Figure 2: The figure shows the effect of N4 bias correction on a FLAIR image with a large lesion load. (a) The original FLAIR image before skull-stripping; (b) after N4 bias correction (with skull); (c) After N4 bias correction (without skull); and (d) After skull-stripping and bias correction using pure-tissue probability mask.

containing WMHs and partial volume effects would have minimal contribution to the inhomogeneity correction itself, leading to improved contrast between the WMH lesions and surrounding tissue. After the bias correction, SegAE was trained again using the original images as input, but now the bias corrected images were used for evaluation of the cost function during training. This way, SegAE learned to segment the original images without the need for intermediate inhomogeneity correction when evaluating new images, which were not in the training set.

3.4. Image enhancement

Presumed inhomogeneity artifacts within the CSF in T2-w and PD-w images were substantial in subjects with enlarged ventricles in the AGES-Reykjavik data set [see Figure 3 (a) and (b)]. N4 bias correction using a pure-tissue probability mask was not sufficient to eliminate these artifacts (see Figure 3 (c), yellow arrows). We observed that inhomogeneity artifacts in the T2-w images and the PD-w images that were acquired simultaneously for each subject were highly correlated and the PD-w images had much lower contrast between the signals of interest. We synthesised enhanced images by multiplying the T1-w and T2-w images with the corresponding intensity transformed PD-w images (see Figure 3 (d) and Figure 4),

\[ I_{new} = I_{orig} \odot (\text{Max}(I_{PD})J - I_{PD}), \]

where \( I_{new} \) is the enhanced image, \( I_{orig} \) is the original T1-w or T2-w image, \( I_{PD} \) is the original PD-w image, \( J \) is a matrix of ones of the same size as the PD-w image, and \( \odot \) denotes an element-wise multiplication. Multiplying the
Figure 3: Image enhancement of a T2-w image using a PD image. (a) and (b) show the original PD and T2-w images, respectively; (c) shows the T2-w image after N4 bias correction with a pure-tissue probability mask; and (d) shows an enhanced image (T2\textsubscript{PD}).

Figure 4: Image enhancement of a T1-w image using a PD image. (a) shows the original T1-w image and (b) shows an enhanced image (T1\textsubscript{PD}).

Intensity transformed PD-w image with a T2-w image results in an image with a slightly degraded contrast of GM and WM compared to the original T2-w image, however, a contrast enhanced image can be acquired by multiplying it with the T1-w image (see Figure 4). We will refer to the enhanced T1-w and T2-w images using PD-w images as T1\textsubscript{PD} and T2\textsubscript{PD}, respectively.

3.5. Training

Two SegAE networks were constructed; one for the AGES-Reykjavik data set and one for the WMH challenge data set, since the AGES-Reykjavik data set comprises T1-w, T2-w, PD-w, and FLAIR images, while the WMH challenge data set only contains T1-w and FLAIR images. Table 3 gives an overview of the training data for each network. The number of Softmax output volumes in both models was 5, one for each material (WMH, WM, GM, CSF, and meninges). The regularization coefficient $\alpha$ was 0.0075 for the AGES-Reykjavik model and 0.02 for the WMH challenge model. Input images were intensity normalized by dividing by the 99th percentile of the
non-zero elements of the image. The training images were cropped to the smallest cuboid containing the brain and patches from the images were acquired with a stride of 40 voxels. Only 50% of the extracted patches, which had the fewest background voxels, were used for training.

Table 3: Overview of the data used to train the two SegAE models for the AGES-Reykjavik (AGES-R.) and WMH challenge (WMH chall.) data sets.

| SegAE model   | Patch size | Modalities          | Reconstruction    |
|---------------|------------|---------------------|-------------------|
| AGES-R.       | 80x80x80x3 | T1, T2, FLAIR       | T1_{PD}, T2_{PD}, FLAIR_{N4} |
| WMH chall.    | 80x80x40x2 | T1, FLAIR           | T1_{N4}, FLAIR_{N4} |

A GTX1080 Ti GPU was used to train the network for 80 epochs with a learning rate of 0.001 using the Adam optimizer [31], with Nesterov momentum [32], with $\beta_1 = 0.9$, $\beta_2 = 0.999$, schedule decay of 0.004, and a batch size of one. During training, Gaussian noise with a standard deviation of 0.05 and zero mean was added to the input patches, and different scalar values drawn from a Gaussian distribution with a mean value of 1 and standard deviation of 0.5 were multiplied with each channel of the input patches to improve the invariance of the network to possibly inconsistent normalization of unseen images. All weights of the convolutional network were initialized using Glorot uniform initialization [33] and biases were initialized as zero. LReLU activation functions had a slope of 0.1 for the negative part. Hyperparameters where chosen by trial and error. The regularization coefficient alpha was the main hyperparameter that needed to be estimated. The 5 validation images were visually inspected and alpha was determined based on the mixture between the estimated materials. Alpha was increased if there was too much mixture between segmentations and decreased if the segmentations were too coarse. The hyperparameters of the optimizer were set to default Tensorflow [34] values.

3.6. Prediction and post-processing

After training, the 5 Softmax output volumes ($\mathbf{S}$ in Figure 1) were used for prediction, while the reconstructed images ($\hat{\mathbf{Y}}$ in Figure 1) were discarded. Prediction was performed with a stride of 40, and patches were assembled using the average of overlapping voxels. The assembled Softmax outputs from SegAE of a subject from the AGES-Reykjavik validation set revealed
the segmentation of WMHs, GM, WM, CSF, and the meninges that remain in the image after skullstripping (see Figure 5).

In this article we focus on automated segmentation of WMH lesions, and hence, only the output volume corresponding to the WMH segmentation is used in our evaluation of the method. The WMH segmentation for the AGES-Reykjavik model was binarized with a threshold of 0.5 and the WMH segmentation from the WMH challenge model was binarized with a threshold of 0.87, as determined with Bayesian optimization for maximizing the average Dice Similarity Coefficient (DSC) [35] on the training data [36], and structures smaller than 3 voxels were removed from the segmentation results from the WMH challenge data due to noise in the cerebellum.

3.7. Evaluation metrics

For each test subject the following similarity metrics were computed to quantify the performance of SegAE and the competing methods compared to manually delineated lesions in the test cases:

- **Absolute Volume Difference (AVD)**
  The absolute difference in volumes divided by the true volume. Defined as $\frac{|V_T - V_P|}{V_T}$, where $V_T$ and $V_P$ denote the volumes of the manually delineated masks and predicted masks, respectively. Lower AVD indicates a more accurate prediction of WMH lesion volume.

- **Dice Similarity Coefficient (DSC) [35]**
  A measure of overlap between the ground truth and predicted segmentations. Using the true positives (TP), false positives (FP), and false negatives (FN) from the confusion matrix, DSC is defined as $\frac{2TP}{2TP + FP + FN}$, and takes values in the range $[0, 1]$. A DSC of 1 indicates a perfect overlap.
• **Modified Hausdorff distance (H95)**

Hausdorff distance measures the longest distance one has to travel from a point in one set to a point in the other set, defined as:

\[
d_H(X, Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\},
\]

where \(d(x, y)\) denotes the distance between \(x\) and \(y\), \(\sup\) denotes the supremum and \(\inf\) the infimum. Here the 95th percentile is used instead of the maximum distance, since the Hausdorff distance is sensitive to outliers. Lower H95 scores indicate better performance.

• **Lesion-wise true positive rate (L-TPR)**

Let \(N_T\) be the number of individual WMH lesions in the ground truth mask \((T)\), and \(N_P\) be the number of correctly detected lesions after comparing the overlap of the predicted mask \((P)\) to \(T\). An individual lesion is defined as a 3D connected component. Then the lesion-wise true positive rate \((L-TPR)\) is defined as \(\frac{N_P}{N_T}\). Higher L-TPR indicates better performance.

• **Lesion-wise F1-score (L-F1)**

Let \(N_P\) be the number of correctly detected lesions after comparing \(P\) to \(T\). \(N_F\) is the number of incorrectly detected lesions in \(P\). An individual lesion is defined as a 3D connected component, and L-F1 is defined as \(\frac{N_P}{N_P + N_F}\). Higher L-F1 indicates better performance.

Finally, for the AGES-Reykjavik test set, a best linear fit was identified between the predicted and manually delineated volumes and the Pearson’s correlation coefficient \((r)\) was used for comparison.

3.8. **Comparison segmentations for the AGES-Reykjavik data set**

The WMHs in a total of 25 subjects were manually delineated by a neuroradiologist to be used as ground truth lesion segmentations for evaluation of the proposed method. We compared the proposed method with three state-of-the-art methods; two publicly available WMH segmentation methods, i.e., the Lesion Growth Algorithm (LGA) [37] and the Lesion Prediction Algorithm (LPA) [38] as implemented in the LST toolbox\(^2\) version 2.0.15, and

\(^2\)www.statisticalmodelling.de/lst.html
one method developed previously for the AGES-Reykjavik data set based on an artificial neural network classifier (ANNC) [39]:

- **LGA** segments WMHs from T1-w and FLAIR images. A CSF, GM and WM segmentation is first obtained from the T1-w image and combined with FLAIR image intensities for calculation of WMH belief maps. The belief maps are thresholded by a pre-chosen threshold ($\kappa$) for an initial binary map, which is grown to include voxels that appear hyperintense in the FLAIR image for a final lesion probability map [37]. We used $\kappa = 0.1$ as determined by the result on our 5 validation images.

- **LPA** segments WMHs from a FLAIR image. LPA includes a logistic regression model trained on MRIs of 53 MS patients with severe lesion patterns obtained at the Department of Neurology, Technische Universität München, Munich, Germany. As covariates for this model a similar lesion belief map as for LGA was used as well as a spatial covariate that takes into account voxel specific changes in lesion probability. This model provides an estimated lesion probability map that can be thresholded for a WMH segmentation [38].

- **ANNC** is an artificial neural network classifier in the four dimensional intensity space defined by the four sequences (FLAIR, T1-w, PD-w, and T2-w) that was previously developed to obtain WMHs, GM, WM, and CSF segmentation for the AGES-Reykjavik MRIs. The input is the voxelwise intensities of FLAIR, T1-w, T2-w, and PD-w images and the classifier was trained on 11 manually annotated subjects [39].

4. Results

4.1. Evaluation on the AGES-Reykjavik data set

Figure 6 visually demonstrates the performance of the methods on four test images; two with the largest and second largest lesion load (1st and 2nd row), one with a medium lesion load (3rd row), and one with the smallest lesion load (4th row).

Table 4 shows the mean and standard deviation of the DSC, H95, AVD, L-TPR, and L-F1 for each of the four methods. We used a paired Wilcoxon signed-rank test to obtain the p-values for determining statistical significance.

We computed the total WMH volume estimated by the four methods and compared with the volume of the manual masks (see Figure 7, top), as well
Figure 6: Visual comparison of the four methods with a manual rater for four different subjects, two with the largest and second largest lesion load (1st and 2nd row), one with a medium lesion load (3rd row), and one with the smallest lesion load (4th row).

as corresponding DSC of the four methods against the manual masks (see Figure 7, bottom). The total WMH volume and DSC for every test subject is ordered by the volume of the manual masks (small lesion load on the left and large lesion load on the right side of the figure) for a direct comparison of DSC for different WMH lesion loads.

Scatter plots showing predicted lesion volumes versus manual lesion volumes for the four methods, as well as the best linear fit and correlation coefficient, can be seen in Figure 8. ANNC and SegAE achieve $r = 0.98$, while LGA and LPA have $r = 0.78$ and $r = 0.73$, respectively.

4.2. Evaluation on the WMH challenge data set

Figure 9 shows a visual comparison between the WMH segmentation of SegAE and the manually delineated masks for 3 subjects in the WMH
Table 4: AGES-Reykjavik results. The mean and standard deviation for each of the
evaluation metrics. Asterisk (*) denotes values that are significantly different from SegAE
\((p < 0.01)\), and bold figures denote the best result for each metric.

| Method | DSC     | H95     | AVD     | L-TPR   | L-F1    |
|--------|---------|---------|---------|---------|---------|
| ANNC   | 0.62 \((\pm 0.13)\)* | 10.16 \((\pm 10.40)\) | 60.49 \((\pm 29.75)\)* | 0.44 \((\pm 0.12)\)* | 0.39 \((\pm 0.10)\) |
| LGA    | 0.66 \((\pm 0.15)\)* | 15.22 \((\pm 9.93)\)     | 26.50 \((\pm 23.58)\) | 0.29 \((\pm 0.12)\) | 0.36 \((\pm 0.11)\) |
| LPA    | 0.66 \((\pm 0.19)\) | 9.20 \((\pm 6.56)\)      | 62.28 \((\pm 73.75)\)  | 0.53 \((\pm 0.27)\)   | 0.40 \((\pm 0.20)\)  |
| SegAE  | 0.77 \((\pm 0.11)\) | 10.97 \((\pm 11.45)\)    | 33.31 \((\pm 36.30)\)  | 0.64 \((\pm 0.19)\)   | 0.47 \((\pm 0.09)\)  |

Table 5: WMH challenge results. The average performance of SegAE on each of the
metrics of the WMH challenge on test data from each scanner, and the weighted average
of the scores achieved on images from each scanner type for each metric.

| Scanner | DSC     | H95     | AVD     | L-TPR   | L-F1    |
|---------|---------|---------|---------|---------|---------|
| Utrecht \((n=30)\) | 0.57 | 31.57 | 79.90 | 0.35 | 0.30 |
| Singapore \((n=30)\) | 0.67 | 17.70 | 16.61 | 0.25 | 0.32 |
| AMS GE3T \((n=30)\) | 0.65 | 16.56 | 22.41 | 0.39 | 0.48 |
| AMS GE1.5T \((n=10)\) | 0.64 | 17.04 | 17.76 | 0.31 | 0.44 |
| AMS PETMR \((n=10)\) | 0.53 | 54.87 | 111.59 | 0.40 | 0.23 |
| Weighted average | 0.62 | 24.49 | 44.19 | 0.33 | 0.36 |

challenge training set. Table 5 shows the average AVD, DSC, Hausdorff
distance, L-TPR, and L-F1 of SegAE on one test data from each of the five
scanners, and a weighted average of the scores achieved for each scanner type
as reported by the WMH challenge website\(^3\). Furthermore, the website shows
boxplots for all 5 metrics comparing the results obtained for each scanner.

5. Discussion

\(^3\)https://wmh.isi.uu.nl/results/himinn/
Figure 7: The top graph shows the overall WMH volume for the manual masks (red) and masks generated by ANNC (purple), LPA (orange), LGA (green), SegAE (blue, dotted), ordered by the volume of the manual masks. The bottom graph shows the DSC for the same methods compared with the manual masks.

Given a training set of brain MRIs, SegAE learns the segmentation of the predominant materials that make up these images. Whether a material is predominant depends on the contrast and abundance of the material in the image. In our case, it was sufficient to randomly sample brain MRIs from the population of elderly subjects to get WMHs as one of those materials (see the lesion load of our training and test data in Figure 10). After training, the segmentations of WMHs, GM, WM and CSF generated by SegAE were visually validated, and if the training was successful, SegAE could be used to directly generate segmentations for new images that were not in the training set. We trained and evaluated SegAE on brain images from a population study with a highly variable WMH lesion load, from almost no WMHs to a very high WMH lesion load. The segmentation results indicate the robustness of our method regardless of lesion load and location.

An advantage of SegAE is that we do not need a large data set of training subjects because our unsupervised methodology is based on the intensity...
features that are shared between all the sequences used as training images. Then after training the method on images from 30 subjects from the AGES-Reykjavik data set, it can be used to segment the remaining subjects (4781 subjects) extremely fast. The average run time per scan in the AGES-Reykjavik test set was 19 seconds using a GTX1080 Ti GPU.

The DSC, AVD, H95, L-TPR, and L-F1 were used as evaluation metrics in the WMH challenge, and we used the same metrics to evaluate our results on the AGES-Reykjavik data set for consistency. On the AGES-Reykjavik test set, we compared the method with three alternative WMH segmentation methods, i.e., LPA, LGA, and ANNC. SegAE achieved the best average
Figure 9: Visual comparison between the WMH segmentation of SegAE and the manually delineated masks for subjects in the WMH challenge training set. The top row shows the first subject (ID: 0) from the Utrecht scanner, the middle row shows the first subject from the Singapore scanner (ID: 50) and the bottom row shows the first subject in the GE3T scanner (ID: 100).

DSC, L-TPR, and L-F1 scores, while LPA achieved the best average H95 score (cf. Table 4). WMHs are not an intact structure so the H95 score is not very informative, however, a high H95 might suggest skullstripping errors causing oversegmentation of WMHs at the brain boundary. LGA achieved the best average AVD score despite having a volume correlation of only 0.78, seemingly because the AVD score penalizes undersegmentations less than oversegmentations, as mentioned in [8]. SegAE and ANNC achieved the highest volume correlation ($r = 0.98$), however, ANNC seems to systematically overestimate the lesion volumes, as indicated in Figure 8, hence SegAE achieves a significantly better AVD ($p < 0.01$) than ANNC. A systematic
overestimation of WMHs can explain the higher AVD and high correlation in ANNC because the correlation coefficient is bias and scale invariant.

The DSC is more sensitive to errors in segmentation of small structures, so DSC was plotted with manual volumes as a reference in Figure 7. Bottom part of Figure 7 demonstrates the robustness of SegAE to a variety of WMH volumes and in Figure 6, bottom row, we visually verify that the segmentation where SegAE achieves the lowest DSC is not a failure.

The results on the MICCAI 2017 WMH segmentation challenge test set can be seen in Table 5. On the challenge website\(^4\), methods are ranked according to the average rank for all metrics, but methods can also be compared for each metric individually. SegAE is currently the best performing unsupervised method, using either the website’s ranking system or the average DSC. The method also compares favorably to some supervised methods.

Assuming that the true WMH segmentations from the WMH challenge and the AGES-Reykjavik data set come from the same distribution, then comparing average scores in Tables 4 and 5 shows that SegAE performs better on the AGES-Reykjavik test set than the WMH challenge test set. This is not surprising, since the FLAIR images in the AGES-Reykjavik data set have better contrast between WMH and GM, and T2-w and PD-w images are used in addition to the FLAIR and T1-w images for training the AGES-Reykjavik network. Figure 5 in Supplementary materials shows that using only FLAIR images or T1-w and FLAIR images for training the AGES-Reykjavik data set can increase susceptibility to artifacts. Visual inspection of the WMH challenge training images shows that some small, low intensity WMHs are not detected (see Figure 9, middle and bottom rows). This could explain the substantially lower L-TPR and L-F1 scores for the WMH challenge test set than the AGES-Reykjavik test set. Furthermore, during training of SegAE on the WMH challenge training set, data from three different scanners are used, while the method is tested on data from five different scanners. This could interfere with training if the image contrast in the different scanners differs, since SegAE reconstructs all training images by the same weighted sum of the segmentation of materials present in the images during training. We note that the meninges class did not appear in the WMH challenge model, possibly due to the absence of T2-w or PD-w images. Finally, it is unknown whether any WMH segmentation errors in the WMH challenge test

---

\(^4\)https://wmh.isi.uu.nl/
Although segmentation of WMHs of presumed vascular origin is the main focus of this paper, hyperintense lesions in FLAIR images can have other causes, such as multiple sclerosis (MS) and traumatic brain injury (TBI). Methods for unsupervised segmentation of FLAIR hyperintensities are often used interchangeably [20] and we believe that the proposed method should be able to segment any lesions with similar intensities in the MRI sequences that we use.

6. Conclusions

We have presented SegAE, a CNN architecture that can be trained in an unsupervised manner to segment WMHs in brain MRIs. We evaluated the WMH segmentation from the proposed method on two separate data sets acquired from six different scanners, i.e. the AGES-Reykjavik data set and the MICCAI 2017 WMH segmentation challenge data set, using ground truth manual WMH labels. For the AGES-Reykjavik test set the method was compared with three alternative WMH segmentation methods, i.e., LPA,
LGA, and ANNC. SegAE achieved the best average DSC, L-TPR, and L-F1 scores, while LPA achieved the best H95 score, and LGA the best AVD score. SegAE achieved a WMH lesion volume correlation of 0.98. The results on the MICCAI 2017 WMH segmentation challenge test set can be seen in Table 5. The scores can be compared with any method sent to the WMH segmentation challenge via the WMH challenge website5.

Acknowledgements

This work was supported by RANNIS (The Icelandic Centre for Research) through grant 173942-051. We thank Burkni Palsson for a valuable discussion about hyperspectral unmixing using a neural network autoencoder, and Nicholas J. Tustison for valuable insights on N4 bias correction on FLAIR images with WMHs.

References

[1] J. M. Wardlaw, M. C. Valdés Hernández, S. Muñoz-Maniega, What are white matter hyperintensities made of? Relevance to vascular cognitive impairment, Journal of the American Heart Association 4 (2015) e001140.

[2] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T O’Brien, F. Barkhof, O. R. Benavente, et al., Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration, The Lancet Neurology 12 (2013) 822–838.

[3] S. Debette, H. Markus, The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis, Bmj 341 (2010) c3666.

[4] A. Wu, A. R. Sharrett, R. F. Gottesman, M. C. Power, T. H. Mosley, C. R. Jack, D. S. Knopman, B. G. Windham, A. L. Gross, J. Coresh, Association of Brain Magnetic Resonance Imaging Signs With Cognitive Outcomes in Persons With Nonimpaired Cognition and Mild Cognitive Impairment, JAMA Netw Open 2 (2019) e193359.

5https://wmh.isi.uu.nl/
[5] S. Lee, F. Viqar, M. E. Zimmerman, A. Narkhede, G. Tosto, T. L. Benzinger, D. S. Marcus, A. M. Fagan, A. Goate, N. C. Fox, et al., White matter hyperintensities are a core feature of Alzheimer’s disease: evidence from the dominantly inherited Alzheimer network, Annals of neurology 79 (2016) 929–939.

[6] S. González-Villà, A. Oliver, Y. Huo, X. Lladó, B. A. Landman, Brain structure segmentation in the presence of multiple sclerosis lesions, NeuroImage: Clinical 22 (2019) 101709.

[7] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, et al., Longitudinal multiple sclerosis lesion segmentation: resource and challenge, NeuroImage 148 (2017) 77–102.

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

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

[10] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation, Medical image analysis 36 (2017) 61–78.

[11] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention 2015, Springer, pp. 234–241.

[12] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken, et al., Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation, in: International Conference on
[13] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, D. Rueckert, GAN augmentation: augmenting training data using generative adversarial networks, arXiv preprint arXiv:1810.10863 (2018).

[14] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, Synthetic data augmentation using GAN for improved liver lesion classification, in: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, pp. 289–293.

[15] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, À. Rovira, Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches, Information Sciences 186 (2012) 164–185.

[16] X. Tomas-Fernandez, S. K. Warfield, A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation, IEEE transactions on medical imaging 34 (2015) 1349–1361.

[17] R. Khayati, M. Vafadust, F. Towhidkhah, M. Nabavi, Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model, Comput. Biol. Med. 38 (2008) 379–390.

[18] J.-W. Chai, C. Chi-Chang Chen, C.-M. Chiang, Y.-J. Ho, H.-M. Chen, Y.-C. Ouyang, C.-W. Yang, S.-K. Lee, C.-I. Chang, Quantitative analysis in clinical applications of brain mri using independent component analysis coupled with support vector machine, Journal of Magnetic Resonance Imaging 32 (2010) 24–34.

[19] K. Krupa, M. Bekiesińska-Figatowska, Artifacts in magnetic resonance imaging, Polish journal of radiology 80 (2015) 93.

[20] C. Bowles, C. Qin, R. Guerrero, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, D. Rueckert, Brain lesion segmentation through image synthesis and outlier detection, NeuroImage: Clinical 16 (2017) 643–658.
[21] C. Baur, B. Wiestler, S. Albarqouni, N. Navab, Fusing unsupervised and supervised deep learning for white matter lesion segmentation, in: International Conference on Medical Imaging with Deep Learning 2019, pp. 63–72.

[22] W. Longstreth, T. A. Manolio, A. Arnold, G. L. Burke, N. Bryan, C. A. Jungreis, P. L. Enright, D. O’Leary, L. Fried, Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people: the Cardiovascular Health Study, Stroke 27 (1996) 1274–1282.

[23] B. Palsson, J. Sigurdsson, J. R. Sveinsson, M. O. Ulfarsson, Hyperspectral unmixing using a neural network autoencoder, IEEE Access 6 (2018) 25646–25656.

[24] H. E. Atlason, A. Love, S. Sigurdsson, V. Gudnason, L. M. Ellingsen, Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder, in: Medical Imaging 2019: Image Processing, volume 10949, International Society for Optics and Photonics, p. 109491H.

[25] L. Forsberg, S. Sigurdsson, J. Fredriksson, A. Egilsdottir, B. Oskarsdottir, O. Kjartansson, M. A. van Buchem, L. J. Launer, V. Gudnason, A. Zijdenbos, The AGES-Reykjavik study atlases: Non-linear multispectral template and atlases for studies of the ageing brain, Medical image analysis 39 (2017) 133–144.

[26] T. B. Harris, L. J. Launer, G. Eiriksdottir, O. Kjartansson, P. V. Jonsson, G. Sigurdsson, G. Thorgerisson, T. Aspelund, M. E. Garcia, M. F. Cotch, H. J. Hoffman, V. Gudnason, Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics, Am. J. Epidemiol. 165 (2007) 1076–1087.

[27] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage (2009) S102.

[28] S. Roy, J. A. Butman, D. L. Pham, Robust skull stripping using multiple MR image contrasts insensitive to pathology, Neuroimage 146 (2017) 132–147.
[29] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee, N4ITK: improved N3 bias correction, IEEE transactions on medical imaging 29 (2010) 1310–1320.

[30] N. J. Tustison, P. A. Cook, A. Klein, G. Song, S. R. Das, J. T. Duda, B. M. Kandel, N. van Strien, J. R. Stone, J. C. Gee, et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, Neuroimage 99 (2014) 166–179.

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[32] T. Dozat, Incorporating nesterov momentum into adam, https://openreview.net/forum?id=OM0jvwB8jip57ZJjtNEZ, 2016.

[33] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics 2010, pp. 249–256.

[34] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[35] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302.

[36] J. Bergstra, D. Yamins, D. D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures (2013).

[37] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, et al., An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis, Neuroimage 59 (2012) 3774–3783.
[38] P. Schmidt, Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging, Ph.D. thesis, lmu, 2017.

[39] S. Sigurdsson, T. Aspelund, L. Forsberg, J. Fredriksson, O. Kjartansson, B. Oskarsdottir, P. V. Jonsson, G. Eiriksdottir, T. B. Harris, A. Zijdenbos, et al., Brain tissue volumes in the general population of the elderly: the AGES-Reykjavik study, Neuroimage 59 (2012) 3862–3870.