Deep learning for segmentation of brain MRI - validation on the ventricular system and white matter lesions

Hans Emil Atlason

Faculty of Electrical and Computer Engineering
University of Iceland
2021

Deep learning for segmentation of brain MRI - validation on the ventricular system and white matter lesions

Hans Emil Atlason

Dissertation submitted in partial fulfillment of a
Philosophiae Doctor degree in Electrical and Computer Engineering

Advisor
Dr. Lotta María Ellingsen

PhD Committee
Dr. Lotta María Ellingsen
Dr. Magnús Örn Úlfarsson
Dr. Vilmundur Guðnason

Opponents
Dr. Bennett Allan Landman
Dr. Pierre-Louis Bazin

Faculty of Electrical and Computer Engineering
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, December 2021

Deep learning for segmentation of brain MRI - validation on the ventricular system and white matter lesions
Deep learning for segmentation of brain MRI - validation on ventricles and WMHs
Dissertation submitted in partial fulfillment of a *Philosophiae Doctor* degree in Electrical and Computer Engineering

Faculty of Electrical and Computer Engineering
School of Engineering and Natural Sciences
University of Iceland
Hjarðarhaga 6
107, Reykjavik
Iceland

Telephone: 525-4000

Bibliographic information:
Hans Emil Atlason, 2021, *Deep learning for segmentation of brain MRI - validation on the ventricular system and white matter lesions*, PhD dissertation, Faculty of Electrical and Computer Engineering, University of Iceland, 90 pp.

ISBN 978-9935-9647-1-7

Printing: Háskólaprent
Reykjavik, Iceland, December 2021

Abstract

Magnetic resonance images (MRIs) enable neuroradiologists to investigate the human brain to look for possible causes of disease. The clinical interpretation of these images is, however, mostly limited to subjective assessment or a rough measurement of the area or volume of brain structures and presence of lesions. Multiple automatic methods have been developed to label different brain structures from MRIs, from which size, shape, and location of these structures and lesions can be extracted. These types of measurements enable researchers to perform comparisons in large scale studies. Multiple conventional whole-brain segmentation methods are based on finding a geometric transformation from MRIs with manually delineated brain structures to a target MRI that will consequently have the corresponding brain structures automatically labeled. These methods have worked very well for transforming labels between subjects and they have been extensively used in brain MRI studies. However, their main disadvantages are: 1) They are very slow (6+ hours for labelling one image), 2) their results are often inaccurate when the brain is deformed, e.g., due to atrophy, and 3) it is not possible to transform brain lesions from one subject to another, since their placement in the brain is variable. Current state-of-the-art brain segmentation methods are often based on deep neural networks (DNNs). DNNs can learn to approximate any function between an input and output given enough training data. After training, the DNNs can be used to analyse images fast and accurately. However, it can be very expensive and time consuming to generate enough training data for DNNs. A DNN trained on one MRI data set often does not work adequately well on another data set where different MRI parameters or scanners are used. Therefore, it would be beneficial to develop DNN methods that minimize the need for manually delineated training data. We have developed novel, automatic methods to label the ventricular system and WM lesions in the brain. Ventricular enlargement and WM lesions are associated with neurodegenerative diseases, e.g. Alzheimer's disease, vascular dementia, and adult hydrocephalus. Our method, called SegAE, is the first unsupervised convolutional neural network for simultaneous segmentation of tissues and WM lesions from brain MRIs. SegAE‘s output are images that show the proportion of tissues and WM lesions in each voxel, but labelling WM lesions automatically has thus far been a challenging problem to solve, e.g. due to the variability of lesion load and location, and the inhomogeneous nature of MRI signal intensities within tissues. Furthermore, we use the output from SegAE to make images that have the same contrast irrespective of scanner type and parameters. That is advantageous because one major challenge in automatic medical image analysis is the lack of consistency of results using different data sets. This way we can make use of a DNN trained on manually labelled images from one data set and use it on another where the DNN input is a standardized image of the materials that cause the signal intensities in the MRI sequences. We have validated our methods on various data sets, including the AGES-Reykjavik study, that includes thousands of brain MRIs with a large variability of ventricular volumes and WM lesions. The methods have been compared to state-of-the-art methods and manual delineations by neuroradiologists. Our results indicate that the methods are accurate and robust to different scanners, and variability in brain structure, as well as being significantly faster than conventional methods.

Útdráttur

Segulómmyndir gera taugaröntgenlæknum kleift að líta inn í heila mannsins í leit að orsökum sjúkdóma. Túlkun myndanna í klíník er hins vegar að mestu leyti takmörkuð við huglægt mat eða grófa mælingu á stærð og umfangi heilasvæða og vefjaskemmda. Fjöldi sjálfvirkra aðferða hefur verið þróaður til að merkja mismunandi heilasvæði út frá segulómmyndum. Með sjálfvirkum merkingum fást mælingar á stærð, lögun og staðsetningu heilasvæða og vefjaskemmda. Slíkar mælingar gera rannsakendum kleift að framkvæma stórar samanburðarrannsóknir. Hefðbundnar merkingaraðferðir eru meðal annars byggðar á því að finna rúmfræðilega vörpun frá nokkrum handmerktum myndum yfir í þá mynd sem við höfum áhuga á að fá nýjar merkingar fyrir. Þessar aðferðir hafa gefist mjög vel til þess að yfirfæra merkingar á heilasvæðum og þær eru mikið notaðar í rannsóknum. Ókostir þeirra eru hins vegar að: 1) Þær eru mjög hægar (6+ klst að merkja eina mynd), 2) niðurstöður eru oft ónákvæmar þegar heilinn er mikið aflagaður svo sem vegna rýrnunar, 3) ekki er hægt að varpa staðsetningu vefjaskemmda frá einum einstaklingi til annars, þar sem staðsetning þeirra í heilanum getur verið breytileg. Nákvæmustu og hröðustu aðferðir sem til eru í dag byggja á djúpum tauganetum. Tauganet geta lært hvaða fall sem er milli inntaks og úttaks ef næg þjálfunargögn eru fyrir hendi. Eftir þjálfun geta tauganetin greint myndir mjög hratt og nákvæmlega. Hins vegar er mjög dýrt og tímafrekt að útbúa næg þjálfunargögn fyrir tauganet. Oft virkar tauganet sem þjálfað er á einu gagnasafni ekki jafn vel á öðru gagnasafni þar sem öðruvísi púlsaraðir (þ.e. stillingar á segulómtækinu) eru notaðar við myndatöku. Því er til mikils að vinna við þróun tauganeta sem lágmarka þörf á handgerðum þjálfunarmyndum. Við höfum þróað nýstárlegar, sjálfvirkar aðferðir til að merkja heilahólf og hvítavefsbreytingar í heilanum, en þekkt er að stækkun heilahólfa og umfang hvítavefsbreytinga eru tengd heilahrörnunarsjúkdómum, m.a. Alzheimerssjúkdómi, æðaheilabilun og fullorðinsvatnshöfði. Aðferðin okkar kallast SegAE, en hún er fyrsta földunartauganetið (e. Convolutional Neural Network) sem þjálfað er á óstýrðan hátt (e. unsupervised) til þess að finna vefi og hvítavefsbreytingar í heila út frá segulómmyndum. Úttak SegAE eru myndir sem sýna hlutfall vefja og hvítavefsbreytinga í hverjum myndpunkti en að merkja hvítavefsbreytingar sjálfvirkt hefur hingað til reynst erfitt vandamál að leysa m.a. vegna þess að stærð og staðsetning hvítavefsbreytinga er mjög breytileg milli einstaklinga og birtustig vefja er ekki staðlað í segulómmyndum. Auk þess notum við úttak SegAE til þess að búa til staðlaða mynd af heilanum óháð gerð og stillingum segulómtækja. Þetta er mikill kostur þar sem ein helsta áskorun læknisfræðilegrar myndgreiningar í dag er ósamræmi niðurstaðna á milli tækja. Þannig getum við tekið tauganet sem þjálfað er á handgerðum myndum úr einu gagnasafni og notað það á öðru þar sem inntakið í tauganetið eru staðlaðar myndir af þeim efnum sem orsaka birtuskil myndaraða í segulómun. Við höfum prófað aðferðirnar á mismunandi gagnasöfnum, þ.á.m. Öldrunarrannsókn Hjartaverndar, sem inniheldur þúsundir heilamynda af einstaklingum með mikinn breytileika m.t.t. heilahólfa og hvítavefsbreytinga. Aðferðirnar hafa verið bornar saman við aðferðir í fremstu röð auk handmerktra mynda, sem merktar hafa verið af taugaröntgenlæknum. Niðurstöður okkar benda til þess að aðferðirnar séu afar nákvæmar og stöðugar gagnvart breytileika milli segulómtækja og breytileika í byggingu heilans og eru þær þar að auki hraðvirkari en fyrri aðferðir.

Table of Contents

List of Figures

List of Tables

List of Original Papers

1: Hans E Atlason, Muhan Shao, Vidar Robertsson, Sigurdur Sigurdsson, Vilmundur Gudnason, Jerry L Prince, and Lotta M Ellingsen. Large-scale parcellation of the ventricular system using convolutional neural networks. In SPIE Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, volume 10953, page 109530N, 2019

2: Hans E Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, and Lotta M Ellingsen.Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder. In SPIE Medical Imaging 2019: Image Processing, volume 10949, page 109491H, 2019

3: Hans E Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, and Lotta M Ellingsen. Segae: Unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder. NeuroImage: Clinical, page 102085, 2019

4: Hans E Atlason, Askell Love, Vidar Robertsson, Ari M Blitz, Sigurdur Sigurdsson, Vilmundur Gudnason, and Lotta M Ellingsen. A joint ventricle and WMH segmentation from MRI for evaluation of age-related changes in the brain. Submitted to Artificial Intelligence in Medicine in 2021

Abbreviations

AC-PC	Anterior commissure - posterior commissure
ADNI	Alzheimer’s Disease Neuroimaging Initiative
AGES	Age, gene/environment susceptibility
AMS	Amsterdam
ANNC	Artificial neural network classifier
APOE	Apolipoprotein E
AVD	Absolute volume difference
BMI	Body mass index
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
DNN	Deep neural network
DSC	Dice similarity coefficient
DWMH	Deep white matter hyperintensities
FA	Flip angle
FLAIR	Fluid attenuated inversion recovery
FN	False negative
FOV	Field of view
FP	False positive
GAN	Generative adversarial networks
GE	General Electric
GM	Grey matter
GPU	Graphical processing unit
ICV	Intracranial volume
ID	Identity
LGA	Lesion growth algorithm
LLV	Left lateral ventricle
LPA	Lesion prediction algorithm
LST	Lesion segmentation tool
LVR	Log volume ratio
MAS	Multi atlas segmentation
MICCAI	Medical Image Computing and Computer Assisted Intervention Society
MNI	Montreal Neurological Institute
MONSTR	Multi-cONtrast brain STRipping
MPRAGE	Magnetization-prepared rapid acquisition with gradient echo
MR	Magnetic resonance
MRI	Magnetic resonance imaging
MS	Multiple sclerosis

MSE	Mean square error
NPH	Normal pressure hydrocephalus
NUHS	National university health system
OASIS	The Open Access Series of Imaging Studies
PD	Proton density
PET	Positron emission tomography
PVH	Periventricular hyperintensities
RLV	Right lateral ventricle
RMSE	Root mean square error
RUDOLPH	RobUst DictiOnary-learning and Label Propagation Hybrid
SD	Standard deviation
SegAE	Segmentation autoencoder
SVD	Small vessel disease
TA	Time of acquisition
TBI	Traumatic brain injury
TE	Time of echo
TI	Time of inversion
TP	True positive
TPR	True positive rate
TR	Time of repetition
UMC	University Medical Center
VU	Vrije Universiteit
WM	White matter
WMH	White matter hyperintensity
V_CNN	Ventricle CNN

Acknowledgments

I would like to thank my supervisor, professor Lotta María Ellingsen, for her guidance and support for the duration of this work. Thanks for giving me the opportunity to develop new research ideas. I would also like to thank Dr. Magnús Örn Úlfarsson and Dr. Vilmundur Guðnason for their valuable comments and for accepting to be part of my PhD committee. My thanks to the opponents, professor Bennett Landman, and dr. Pierre-Louis Bazin for their constructive assessment and participation in the defense. I am very greatful to my friends at VR-II who have worked beside me, and inspired me through daily discussions; Burkni, Frosti, Jakob, Sveinn, Magnús, Bin, Han, and others. Last but not least, I would like to thank my family for their invaluable support in life, and my girlfriend; who is working to finish her PhD at the same time, and makes every day more colorful with her distractions.

1 Introduction

1.1 Clinical background and significance

As we age, the brain undergoes progressive brain atrophy and the risk of neurodegenerative diseases and cognitive decline increases [6]. Alzheimer's disease and cerebrovascular diseases [7] are two of the most common causes of dementia, although there are many other causes [8]. Many of these diseases cause changes in the brain that may be visible long before onset of dementia [9], such as region specific atrophy and lesions that are visible in structural magnetic resonance images (MRIs) [10].

1.1.1 Brain imaging

The human brain consists mainly of grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) (see Figure 1.1). The histological terms GM and WM are used to distinguish regions consisting mostly of neural cell bodies, dendrites, and synapses (GM) and of myelinated axons (WM) [11]. The color difference is mainly caused by the myelin sheaths surrounding the axons which insulate them and greatly increase the conduction velocity of the axons [11]. The CSF is a transparent fluid derived from blood plasma. It is found in the ventricles where it is produced and in the subarachnoid space covering the brain and spinal cord.

MRI can be used for non-invasive imaging of the brain by utilizing the macroscopic polarization of the hydrogen atoms in the brain when it is positioned in external magnetic fields; by measuring the radio frequency signal caused by hydrogen atoms that have been excited at specific frequencies [12]. This signal decays with time, with a time

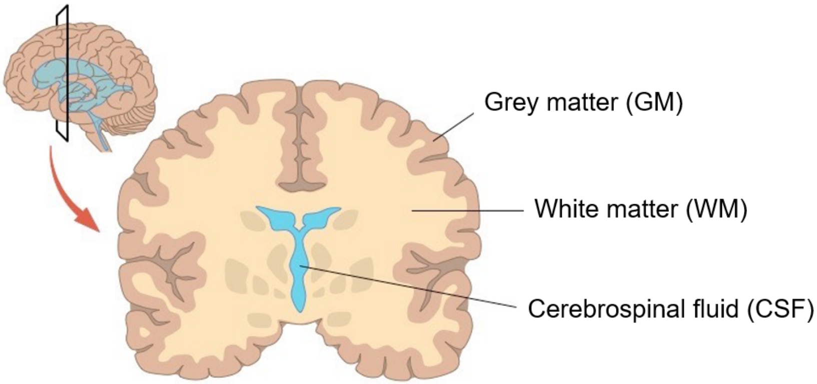


Figure 1.1. A coronal slice of the brain showing the gray matter, white matter, and cerebrospinal fluid within the brain. Image modified from BioNinja.com.au.

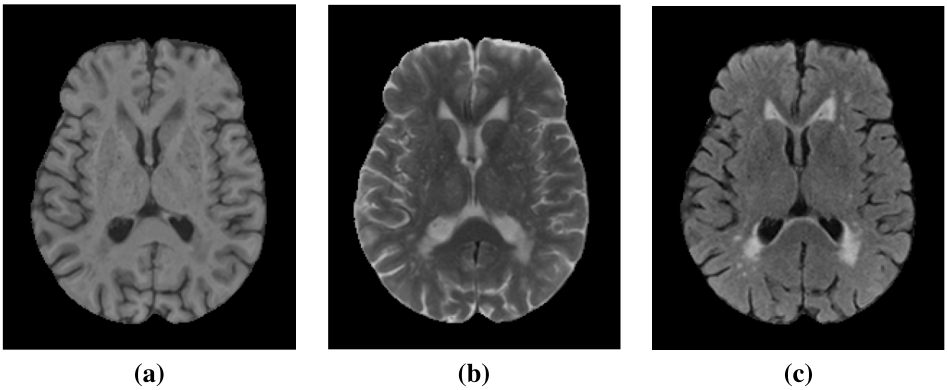


Figure 1.2. Images ***(a)****,*(***b***)*, and* ***(c)*** *show T1-w, T2-w, and FLAIR images, respectively, after spatial normalization to a standardized coordinate system and skull removal.*

constant that is different depending on the material the hydrogen atoms are a part of, which gives rise to tissue contrast [12]; e.g. the myelin is a fat-rich substance which influences MRI signal strength. Various methods exist for controlling the tissue contrast. MRI pulse-sequences that emphasize signal differences during T2 and T1 relaxation times are named T2-weighted (T2-w) and T1-weighted (T1-w) images, respectively. By minimizing the impact of the T1 and T2 differences, proton density weighted (PD-w) images can be produced. Furthermore, the fluid attenuated inversion recovery (FLAIR) sequence is set to null fluids, which brings out the white matter lesions discussed in Section 1.1.3. T1-w MRIs show higher signal intensity for white matter than grey matter, and low intensity for CSF. T2-w images show the CSF with the strongest signal, while in FLAIR images CSF is completely attenuated (see Figure 1.2).

1.1.2 Ventricle enlargement

The human brain contains four interconnected ventricles, i.e. the left and right lateral ventricles, and the third and fourth ventricles, in which CSF is produced (see Figure 1.3). Enlargement of the ventricles may occur due to atrophy or impaired CSF circulation [13]. Ventricular enlargement due to atrophy can happen for reasons such as cortical atrophy, traumatic brain injury (TBI), or a cerebral vascular incident and impaired CSF circulation may be due to impaired outflow or absorption of CSF from the ventricles [14]. The foramina of Monro (which connect both lateral ventricles to the third ventricle) may be congenitally malformed, or obstructed by infection, hemorrhage, or tumor. In neonatal patients, ventricular enlargement is found in around 1-2 per 1000 pregnancies [14]. Furthermore, increase in ventricular volume has been associated with normal aging, schizophrenia, bipolar disorder, multiple sclerosis (MS), as well as several age related brain diseases such as Alzheimer's disease and adult hydrocephalus [14–16]. However, ventricular enlargement alone is not diagnostic of any clinical process [14]. A neurodegenerative disease that can cause extreme ventricle enlargement in the aging brain is idiopathic normal pressure hydrocephalus (NPH). The prevalence of probable NPH was estimated to be at least 21.9 per 100,000 in a Norwegian population study

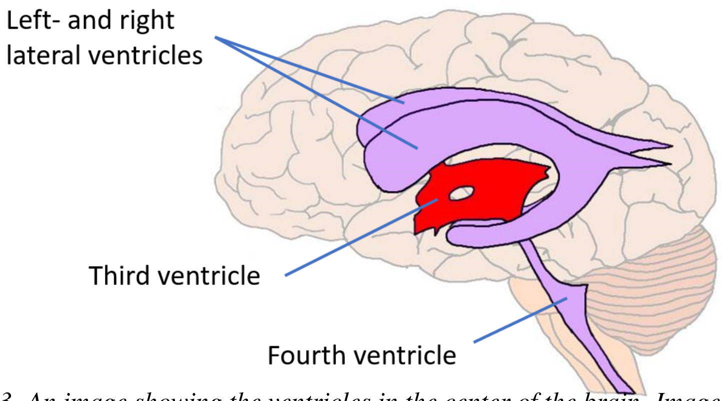


Figure 1.3. An image showing the ventricles in the center of the brain. Image modified from neuroscientificallychallenged.com .

and the prevalence was found to be increasing with age [17]. NPH is treatable with shunt-surgery or endoscopic third ventriculostomy and is potentially reversible in properly selected patients [18]. NPH is characterized by the classic clinical triad: Gait and balance impairment, cognitive decline, and urinary incontinence [19]. Ventricular enlargement and the characteristics of the triad are features that overlap with normal aging and other neurodegenerative diseases [20, 21]. Furthermore, NPH frequently co-occurs with Alzheimer's disease and cerebrovascular disease [18]. Therefore, the disease is difficult to diagnose and differentiate from other causes of dementia. Structural MRI has been used for diagnosis using traditional measurements such as the callosal angle [22] (see Figure 1.4), and the Evan's index [23] which gives a rough estimate of ventricular volume (see Figure 1.5). More accurate measurements using automatic segmentation methods that are robust to extreme ventricular enlargement could replace these measurements [23], and elucidate new biomarkers of NPH. Other types of MRI measurements, such as diffusion weighted imaging and phase contrast MRI, can also aid in diagnosing causes of ventricular enlargement [24, 25].

1.1.3 White matter lesions

White matter lesions that appear hyperintense in T2-w and FLAIR images, and can appear hypointense in T1-w images, are frequently observed in MRIs of the elderly (see Figure 1.6). In cohort studies of the elderly, they are often attributed to cerebral small vessel disease [26, 27] and termed white matter hyperintensities (WMHs) of presumed vascular origin [27] (also termed leukoaraiosis). Despite the commonly used terms WMHs and WM lesions (due to how frequently they are seen in the periventricular and deep white matter) they have also been identified to occur in the deep gray matter and can be referred to as subcortical hyperintensities in those cases [26].

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 [28, 29]. Lesions visible in brain images of subjects with a common form of vascular cognitive impairment are collectively known as small vessel disease (SVD) [26]. These lesions include WMHs of presumed vascular origin, lacunar infarcts, and

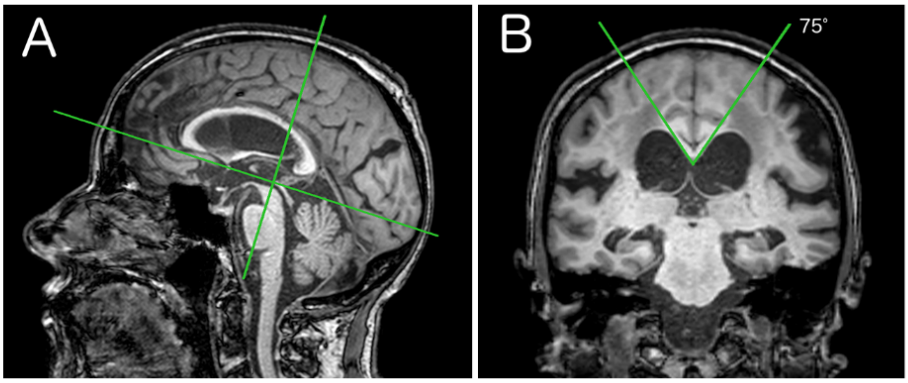


Figure 1.4. Callosal angle measured on a T1-w image. A sagittal image (A) is used to select the coronal slice (B) for measurement of the callosal angle perpendicular to the AC-PC plane [1]. Image modified from [1].

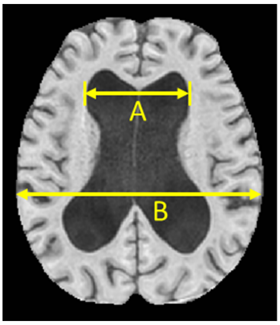


Figure 1.5. Evan's index is a clinical marker for ventricular volume; calculated as the ratio of the maximum width of the frontal horns of the lateral ventricles (width A) and the maximal internal diameter of the skull (width B) in the same axial slice [2].

enlarged perivascular spaces [26]. The effects of these SVD lesions on cognition are cumulative [26]. The prevalence of WMHs increases dramatically with age. Only 4,4% of subjects older than 65 years old were free of any findings in one study of 3301 people 65 years or older, with 80% of subjects falling into categories 1, 2 and 3 of the Fazekas scale [5, 30] (see Table 1.2). However, WMH load is highly variable between individuals [31], and is associated with an increased risk of stroke, dementia, and death [28].

Post-mortem immunohistochemical and gene expression microarray studies indicate a role for hypoxia/ischemia in the development of the disease, and a contributing role of immune activation, blood-brain barrier dysfunction, altered cell metabolic pathways and glial injury [32]. Varying signal strength of WMHs in MRIs can represent different amount of tissue damage, such as in "dirty white matter", and structural and vascular changes might extend further than visible WMHs [26]. The increase in mean diffusivity in MRIs of normal appearing white matter, even in subjects with the mildest Fazekas

Table 1.2. A definition of the Fazekas scale [5] for periventricular hyperintensities (PVH) and deep white matter hyperintensities (DWMH).

Grade	PVH	DWMH
0	Absence	Absence
1	"caps"	Punctate foci
2	Smooth "halo"	Beginning confluence of foci
3	Irregular PVH extending into DWM	Large confluent areas

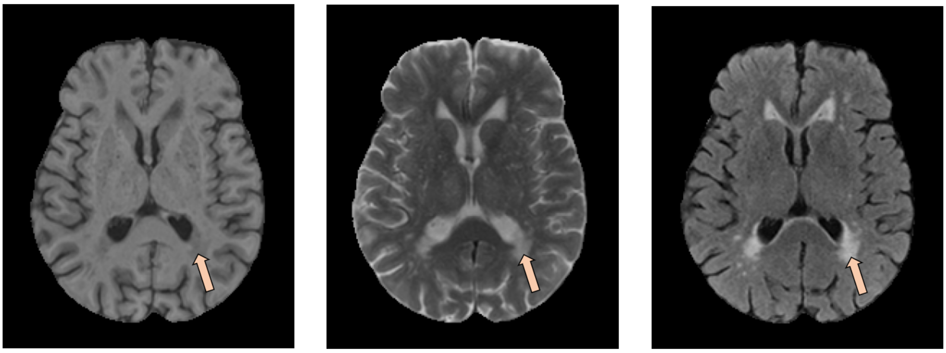


*Figure 1.6. White matter hyperintensities in a single subject. Images **(a)**,**(b)**, and **(c)** show the appearance of WMHs in T1-w, T2-w, and FLAIR images, respectively.*

score, suggests that altered water mobility may be an early feature of white matter pathology in the aging brain, before changes in myelin, axonal integrity or total water content [26, 33]. MRI is thus highly suitable for early detection of SVD onset.

1.1.4 WMHs and the ventricular system as biomarkers

Neurodegenerative diseases and normal aging can both cause WMHs and enlarged ventricles. Both WMHs and enlarged ventricles are biomarkers for numerous conditions, e.g., genetic diseases [34] and autoimmune diseases, such as MS [35]. Early detection of neurodegenerative diseases by use of neuroimaging biomarkers, such as WMH load or ventricle volume, is important to aid in understanding the pathogenesis of these diseases, and make strides towards therapeutics development. Robust detection at early stages enables investigators to start testing possible therapeutic strategies and select presymptomatic patients for clinical trials [36].

To investigate causes of dementia using brain MRI, various structural biomarkers must be analysed, including volumes, shapes, and location in the brain. Biomarkers should not be looked at in isolation when there may be other causes of similar cognitive or physical impairment that present with different biomarkers in the same subject [37]. Furthermore, it may be difficult to distinguish abnormal size of structures, such as the ventricles, because the size may also depend on factors that are not caused by disease, such as age, sex, and intracranial volume (ICV) [38]. Information from large data sets of brain MRIs can help elucidate biomarkers that better predict abnormality.

Studies have found associations between WMH load and region specific atrophy [39], which are both biomarkers of small vessel disease [27]. The WMH load and CSF volume increase with age while the GM and WM volumes decrease in the elderly population [31]. Disproportionate ventricular dilation is associated with WMH load, which may relate ventricular dilation to small vessel disease [40]. The association of WMH load and ventricular volume has also been shown to be independent of demographics, vascular burden and Apolipoprotein E genotype (APOE) [41] (the gene's alleles carry different risk of dementia, Alzheimer's disease and cardiovascular disease [42]). Further analysis of underlying pathologies is made possible as data acquired using brain segmentation methods and large scale biomedical databases is made more informative and reliable.

1.2 Brain MRI segmentation

The use of robust, accurate, and automated brain segmentation methods is crucial when using specific brain structures as biomarkers, especially when analysing large data sets of MRIs. The alternative being: 1) Inaccurate approximations such as the Evan's Index [23] to measure ventricle size (see Figure 1.5) and the Fazekas scale for measuring WMH load [43] (see Figure 1.7), and 2) counting voxels manually, which is highly impractical for data sets of large three-dimensional images. Segmentation methods classify each pixel/voxel in a given image with a label that represents an object (e.g. 0 for background and 1 for a lesion). The currently accepted gold standard in brain segmentation is manual delineation by an expert in neuroanatomy. However, human raters can have great intra- and inter-rater variability [44] and acquiring such delineations is both time-consuming and expensive, making it impractical for analysis in large-scale studies.

A typical brain segmentation pipeline involves registration to MNI-space [45,46] to standardize spatial coordinates between MRIs, bias field correction [47,48] to remove the low frequency spatial intensity variations due to magnetic field inhomogeneity, brain extraction (also known as skull-stripping) to remove non-brain tissue (such as the skull, tongue, and eyes) from the MRIs, and automatic brain segmentation for labelling and parcellation of structures of interest; using, for example, registration or machine learning based methods.

1.2.1 Atlas-based segmentation

MRI provides great contrast between brain tissue types, although automatic segmentation is challenging due to image artifacts such as partial volume effects, image noise, and the bias field. Furthermore, many brain regions cannot be uniquely identified by tissue contrast, so a priori anatomical information is often used for segmenting brain structures. That information is usually provided using an atlas, which consists of an intensity image (such as a T1-w MRI), and a labelled image [49]. The labels of the

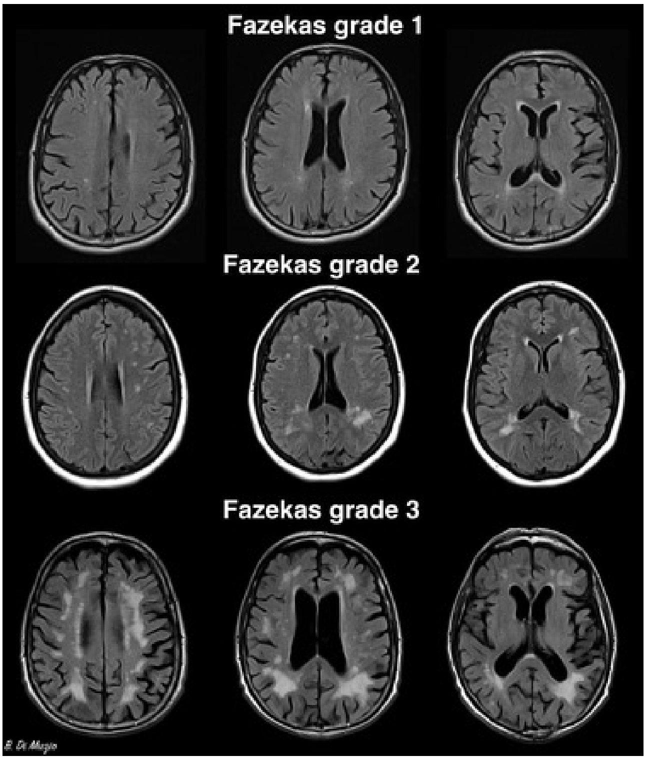


Figure 1.7. Three subjects with a Fazekas grade from 1 to 3 (from top to bottom). Image from Radiopedia.

atlas can be transformed to an image of a subject of interest (a target image) by finding a transformation from the intensity image of the atlas to the target image [49]. The task of finding the spatial correspondence between images is termed registration, and it involves warping an image to optimize spatial alignment restricted to physically plausible deformations [50].

For intra-subject applications such as longitudinal studies and multi-modal registration, rigid and affine transformations usually suffice. However, the anatomical variation of inter-subject applications can only be captured with non-linear algorithms [49], e.g. due to the variation between sulci and gyri patterns in different individuals. A volumetric registration of brain images is often performed by first doing a rigid or affine transformation for an initial alignment, and then a non-linear registration to model local deformation at a higher computational cost [49]. Inter-subject registration of a single atlas to a target subject is error-prone due to anatomical variability, which can be mitigated to a large extent with multi-atlas segmentation (MAS) [49]. After multi-atlas label propagation, the atlas labels are often combined using voting rules (such as majority voting) [49]. The basic steps of MAS consist of registration, label propagation and fusion (see Figure 1.8). Since the introduction of MAS, more advanced methods have been developed based on MAS with additional processing steps and sophisticated

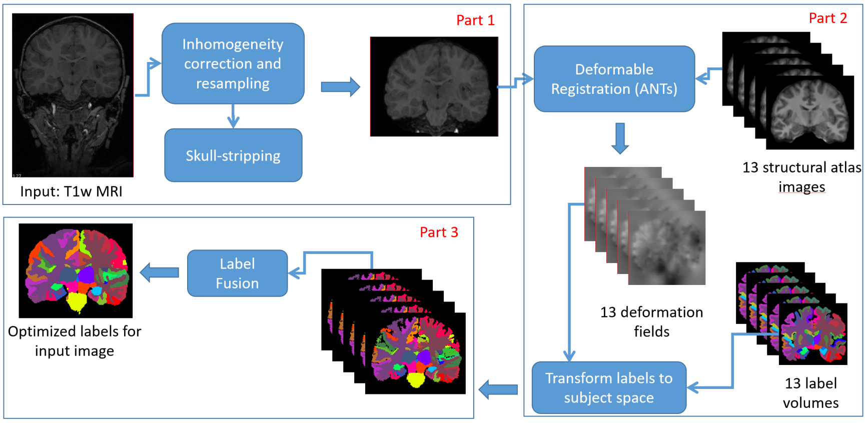

Part 1
Input: T1w MRI
Inhomogeneity correction and resampling
Skull-stripping
Part 2
Deformable Registration (ANTs)
13 structural atlas images
13 deformation fields
13 label volumes
Transform labels to subject space
Part 3
Label Fusion
Optimized labels for input image

Figure 1.8. Multi-atlas segmentation pipeline. Image from Ellingsen et al., SPIE Medical Imaging presentation, 2016.

optimization procedures [50]. Conventional whole brain segmentation methods include atlas based methods [51–53], such as MAS methods that use deformable registration of multiple annotated atlas images to the subject at hand. MAS methods work very well for registering a large amount of labels to subject brain MRIs that resemble the atlas images.

1.2.2 Segmentation in aging and neurodegenerative diseases

Automatic segmentation of brain MRIs of elderly subjects is challenging due to normal brain atrophy and increased prevalence of neurodegenerative diseases. Most freely available atlases are constructed using brain MRIs of relatively young people that may introduce bias in the study of elderly anatomy. Therefore, atlases for specifically studying the aging brain have been introduced [54]. The aging effect is also present in brains containing abnormal regions [49]. Some work has been done on atlas construction using patients with specific diseases to allow quantitative examination of the progression of a disease [49].

A key challenge when using MAS in relation to WMH segmentation is that the size and location of WMHs varies greatly between subjects and hence, they cannot be accurately registered from one subject to another [55–57]. 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 [58]. Also, most MAS methods for whole brain segmentation solely rely on T1-w images, which do not provide as good WMH lesion contrast as FLAIR images. T2-w images show better WMH to WM contrast compared to T1-w images, however, the boundaries between WMHs and CSF spaces can be hard to distinguish. Furthermore, T2-w images are less sensitive to subtle WMHs compared to FLAIR images.

While FLAIR images provide good WMH contrast, several types of artifacts complicate the automatic segmentation of WMHs: Hyperintensities surrounding the third and fourth ventricle, fornix, aqueduct, and cisterns ventral to the mesencephalon; uniform hyperintensity along the external capsule; hyperintense signal within the ventricles due to the choroid plexus; hyperintensities in the corticospinal tract pathway; motion artifacts; image reconstruction artifacts (Gibbs ringing); small linear hyperintensities near the sinuses and carotid arteries; and magnetic susceptibility of several structures [26]. Hyperintense pulsation artifacts can appear within the ventricles in FLAIR images. They are often more severe in the third and fourth ventricles than in the lateral ventricles, and they are associated with ventricular size and increasing age [59].

Other effects can complicate WMH segmentation in MRIs: The bias field may be removed with bias field correction tools [47, 48], however, these tools may degrade the WMH signal intensities if WMHs are not accounted for in the bias field correction process [60]. The presence of subtle hyperintense white matter known as "dirty white matter" may be an indicator of emerging WM lesions [26]. They are ill-defined and may influence manual and automatic delineation of WMHs. Furthermore, random noise can effect detection of subtle changes in normal appearing white matter. Hyperintense lesions in FLAIR images can appear for reasons other than WMHs of presumed vascular origin; such as stroke, MS and TBI. If not accounted for, stroke lesions will distort the measurement of WMH volume and progression, as well as study outcomes [61].

Finally, MAS can fail when presented with severely enlarged ventricles [62]. The variability of ventricle sizes in elderly cohorts or in patient populations with neurodegenerative diseases such as NPH is much larger than that of healthy subjects. Since individuals with risk of ventricle deformations are often subjects of interest in clinical research studies, it is advantageous for brain segmentation methods to be robust to these changes.

Current state-of-the-art methods in most brain segmentation tasks are based on convolutional neural networks (CNNs) [63–66]; a deep learning method discussed in more detail in Section 1.4. These methods have successfully been used for ventricle segmentation [63] and WMH segmentation [64–66] separately. Many earlier WMH segmentation methods obtained WMH lesions as outliers of tissue segmentation [67]. The best results of the 2008 Multiple sclerosis challenge used an expectation-maximization algorithm initialized with atlas probabilities [68]. Five of nine submitted methods used an atlas-based strategy [49]. In contrast, the best result of the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) 2017 WMH segmentation challenge was based on an ensemble of deep neural networks, and 14 of the 20 submitted results employed some form of deep neural networks [64]. CNNs often generate results in a fraction of the time of the MAS methods for brain segmentation [55], which is important when analysing big data sets and for use in clinical settings.

1.3 Brain MRI data sets

A plethora of brain MRIs have been acquired in clinical settings; for diagnosing neurodegenerative diseases and evaluating patient's prognosis, and in research settings to gain knowledge about the anatomy, function, and diseases of the brain. In recent years an emphasis has been on acquiring large biomedical databases to accelerate scientific discoveries that can improve health.

The AGES-Reykjavik study at the Icelandic Heart Association was initiated in 2002 and was designed to examine risk factors, including genetic susceptibility and gene/environment interaction, in relation to disease and disability in old age [69]. The AGES-Reykjavik study cohort comprises 5764 participants (female and male, age 66-93 at first visit), 4811 of which underwent brain MRI [54]. The methodological development and analysis conducted for this thesis is mainly centered on data acquired in the AGES-Reykjavik study. *The UK Biobank* is a large scale biomedical database with health information from half a million UK participants, thereof 100,000 participants will undergo MRI [70]. Other large brain imaging data sets include the *Open Access Series of Imaging Studies (OASIS)* [71] and the *Alzheimer's Disease Neuroimaging Initiative (ADNI)* [72]. The OASIS-3 data set consists of 1098 participants [73] and the ADNI3 study is ongoing [74].

1.4 Deep learning for brain MRI segmentation

Neural networks are universal function approximators between an input $\boldsymbol{X}$ and output $\boldsymbol{Y}$ [75]. The building block of a neural network is the node (or neuron) which consist of weights, w_i, and a bias, b, followed by a non-linear activation function g:

$$y = g(\sum_i w_i x_i + b), \tag{1}$$

where the activation function is often one of sigmoid, the hyperbolic tangent, rectified linear unit (ReLu), a modified version of ReLu such as Leaky ReLu (LReLU), or the softmax function for multi-class classification. These nodes can be ordered into multiple layers with multiple nodes in each layer to form what is commonly called a fully connected neural network. A neural network with numerous hidden layers is called a deep neural network (DNN) (see Figure 1.9). A special case of DNNs is the CNN which is usually applied in computer vision tasks [76] (see Figure 1.10). In CNNs, the weights are filters of a certain size, which is equivalent to weight sharing in fully connected neural networks, leading to shift- and translation invariance which is beneficial for analysing visual information.

1.4.1 Neural network architectures

A neural network architecture is specified by design choices such as the number of layers, types of layers (convolutional, fully connected, pooling, upsampling, etc.), number of filters in each convolutional layer, kernel sizes of the filters, and types of activation functions. Many factors influence these design choices, such as; the number of parameters needed for the DNN to learn an acceptable solution, hardware limitations such as graphics processing unit (GPU) memory size, processing time, compressed latent space representation using a bottleneck, and the use of skip connections to transfer lower level features directly to layers after the next layer [77, 78]. Notable architectural improvements of CNNs for use in medical image segmentation include the change from

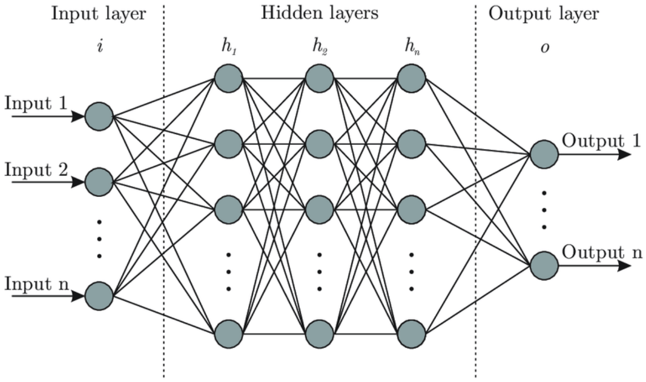


Figure 1.9. An artificial neural network architecture with multiple hidden layers. Image from [3].

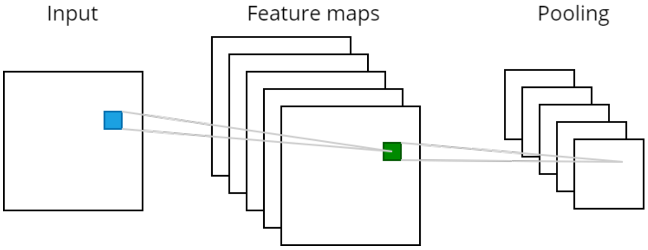


Figure 1.10. Convolutional neural networks use a number of filters of certain size, biases, and an activation function to create feature maps. The filter weights and the bias are updated using backpropagation during training. Downsampling of the feature maps can be done using strided convolutions or pooling layers.

two-dimensional (2D) to three-dimensional (3D) CNNs, which resulted in increased accuracy by incorporating 3D context [66, 79]. Furthermore, classical CNNs often consisted of multiple convolutional layers of decreasing resolution, until the convolutional layers are succeeded by one or more fully connected layers for classification or regression. Later this architecture was mostly replaced by fully convolutional networks [80], with an encoder path of decreasing resolution and a subsequent decoder path of increasing resolution until the output is of the same size as the input. Thus, the fully convolutional network makes a prediction for every voxel of the input image, with each channel representing a segmentation of different label. A widely used fully convolutional network architecture for image segmentation is the U-net [77]. It has skip connections between the downsampling and upsampling paths, which preserves details from higher resolution stages of the encoder and decreases training time.

1.4.2 Training neural networks

Training a neural network involves minimizing a cost function with respect to the parameters of the network. The cost function, $C(f(\boldsymbol{x}),\boldsymbol{y})$, consists of the loss function, $L(f(\boldsymbol{x}),\boldsymbol{y})$, that measures error between the predicted output, $f(\boldsymbol{x})$, and the true image (e.g. a ground-truth segmentation), $\boldsymbol{y}$, and sometimes a regularization term, $\Omega(f)$, which can prevent over-fitting or help to solve ill-posed problems by penalizing undesirable properties of the model f:

$$C(f(\boldsymbol{x}),\boldsymbol{y}) = L(f(\boldsymbol{x}),\boldsymbol{y}) + \alpha\Omega(f),$$

where the regularization coefficient α controls the weighting of the regularization term relative to the loss function. Examples for the use of regularization include penalties on the neural network parameters to reduce overfitting (such as the L2 norm), and sparcity penalties.

Training is an iterative minimization of the cost function performed with backpropagation [81]: The gradients of the cost function with respect to each weight and bias are computed with the chain rule, iterating backwards from the last layer. The learning rate determines the step size that the parameters are updated in the direction of the gradient at each iteration. The batch size is the number of training samples the model evaluates before updating the network parameters, and the number of epochs is the number of passes through the whole training set during training. Learning rate, number of epochs, batch size, and the regularization coefficient are examples of hyperparameters. An optimal set of hyperparameters can be found by comparing prediction results on a validation set from multiple training runs, using a different set of values for the hyperparameters each time.

The development data set is usually divided into: A training set, for training the parameters of the network; a validation set, for determining hyperparameters; and a test set, to measure the performance of the network on new data. This three-way split helps to analyse the bias-variance trade-off and find the optimal hyperparameters without over-fitting on the test set.

1.4.3 Supervised learning

Supervised neural networks are trained using the input $\boldsymbol{X}$ and the corresponding ground truth output $\boldsymbol{Y}$ to create a prediction $\hat{\boldsymbol{Y}}$ of the output. In the case of CNNs for image segmentation, the input is the image that is to be segmented and the network is trained to predict the ground truth segmentation. A fundamental issue in supervised learning is the lack of ground truth labels. In brain MRI segmentation a manual delineation by an expert in neuroanatomy is still the gold standard for ground truth segmentations. Obtaining manually segmented images is laborious and slow, and hence often impractical for generating new training data for CNNs. Furthermore, supervised CNNs may not produce as accurate results when applied to different data sets from different MRI scanners or populations [82].

Attempts to reduce the number of manually delineated masks needed for training include data augmentation [83], training on external data sets, weight regularization, and generative adversarial networks (GANs) [84–86]. Data augmentation such as translation, rotation, and scaling is used to generate new labeled images from available images. External labeled data sets can be used for training to improve generalization if annotated data is scarce. Transfer learning involves using a model trained on a large external data set and fine-tune it on the target data set [87, 88]. A major obstacle in medical imaging is the difference in image features between available labeled data and new data encountered in practice. These differences can arise due to different acquisition protocols, scanner types, or study population differences. Domain adaptation techniques reduce this distribution difference between data sets by learning a common latent representation or by translating images between different domains [87]. CycleGANs are frequently used for domain translation [89]. Furthermore, it is possible to simply incorporate multiple data sets to train a segmentation model, which has shown superior performance compared to those trained on individual data sets [87].

A common problem in medical image segmentation is class-imbalance. That is when the number of class samples in a training set is not balanced, such as when a much fewer number of voxels in the training set of images represent one class than another. If the loss function treats all training samples equally in a training set with class-imbalance, the trained model can end up biased towards predicting the correct outcome accurately for the majority classes at the expense of the minority class. Some methods to improve training in a training set with class-imbalance include: 1) Using a weighted-crosscategorical entropy loss function or Dice loss [90], 2) sampling more patches with classes of low prevalence and less with high prevalence [91], and 3) using data augmentation to create artificial new instances of the low prevalence classes.

1.4.4 Unsupervised learning

Unsupervised methods require no ground truth segmentations for training. They typically involve modeling of MRI brain tissue intensities. However, many brain segmentation tasks depend on human-made naming conventions so manually created atlases are needed to label these structures, e.g. the parcellation of the ventricular system into its four main compartments, i.e., the left and right lateral ventricles, and the 3rd and 4th ventricles. Nevertheless, tissue and lesion segmentation with unsupervised methods can provide enormous benefit by eliminating the need for manual delineations, which are often impractical or impossible to obtain with acceptable accuracy. A number of unsupervised methods have been proposed for WMH segmentation. These include methods that obtain WMH lesions as outliers of tissue segmentation [67] and approaches that use specific features of lesions, such as voxel intensity and appearance [67, 92, 93]. Clustering or unmixing methods could potentially be used on a per image basis if a given image has enough WMH lesion load [94]. 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. It would be beneficial for tissue and lesion segmentation methods to have standardized output for every subject, e.g. if the soft segmentations represent the true proportion of a certain tissue or lesion in each voxel.

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 [26], however, various artifacts or poor skull-stripping can lead to high-intensity regions in FLAIR images [95] 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 [96–98]. 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 [96, 97]. 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 [30].

Autoencoders are neural networks that simultaneously learn an encoding of the input and a reconstruction of the input from the encoding. The loss function penalizes differences in the reconstruction and the input image. That is to say, autoencoders learn in an unsupervised manner without the need for training labels. Autoencoders are often trained with loss functions such as mean square error (MSE), root mean square error (RMSE), and other functions comparing the absolute values of the differences between the input image and the reconstructed image. However, in many image analysis tasks, such as reconstruction of MRIs, the scaling of an image $\boldsymbol{X}$ with a constant a should result in maximal similarity between $\boldsymbol{X}$ and $a\boldsymbol{X}$. This can be achieved using scale-invariant loss functions, such as cosine proximity. Autoencoders can be used for compression, noise reduction, feature learning, generative modelling (e.g. variational autoencoders), and anomaly detection; such as by modelling healthy brain tissue to detect lesions in brain MRIs as mentioned above [98]. The encoder path in these applications is usually made up of gradually smaller layers (lower spatial resolution in the case of CNNs), that can be regarded as a compressed representation of the input, and the decoder path, which learns to upsample the representations and reconstruct the input. Another application of interest is to incorporate into an autoencoder a model of the causal mechanisms that give rise to the pixel values in images. This has been done in hyperspectral unmixing of remote sensing images [99, 100]. A linear unmixing model can be denoted as follows:

$$\hat{\boldsymbol{Y}}_c = \sum_{i=1}^{M} w_{i,c}\boldsymbol{S}_i, \tag{2}$$

where $\hat{\boldsymbol{Y}}_c$ is one channel of the output, $w_{i,c} \in \mathbb{R}_{\geq 0}$ are the weights, $\boldsymbol{S}_i$ are images showing the pixel-wise proportion of each material, M is the number of materials to be estimated, $\boldsymbol{S}_i \geq 0$ and $\sum_{i=1}^{M} \boldsymbol{S}_i = \boldsymbol{J}$ where $\boldsymbol{J}$ is the matrix of ones. Estimating the weights $w_{i,c}$ and proportions $\boldsymbol{S}_i$ is an ill-posed inverse problem that requires appropriate regularization. This model can be implemented in a neural network with restrictions on the convolutional layers and the representations. The non-negativity constraint and the sum-to-one constraint of $\boldsymbol{S}$ can be enforced with an activation function such as the Softmax function. The weighted sum can be implemented with a 1x1x1 convolutional layer that is constrained to have non-negative weights and zero bias. In this thesis, the first effort to implement a linear mixture model with an autoencoder for brain MRI segmentation will be presented.

1.5 Thesis contributions and organization

The main objective of this thesis is to develop robust methods for ventricle and WMH segmentation of MRI of elderly brains. The major challenges in automatic segmentation that we address in this work are 1) the lack of training data for CNN segmentation methods, 2) the segmentation failures of conventional multi-atlas registration based methods due to high variance in brain structures and abnormalities, 3) the long processing time of conventional methods, 4) inconsistent segmentation results of images acquired with different MRI parameters or MRI scanners, and 5) effective pre- and post-processing to adjust for MRI artifacts (such as inhomogeneity and pulsation artifacts).

The solution we have developed is a brain segmentation pipeline comprising skull-stripping, tissue and lesion segmentation, and ventricle segmentation with a parcellation of the ventricular system into its four main compartments, i.e., the left and right lateral ventricles, and the 3rd and 4th ventricles. The training data generation for each step was automatic and did not involve manual delineation of brain structures. Instead, a combination of unsupervised learning and multi-atlas segmentation was used to generate enough training data for the development of the pipeline.
The major contributions of this dissertation are the following:

In Chapter 2 we introduce **Unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder**: Unsupervised methods have previously been used for WMH and tissue segmentation. However, modelling tissue intensities can be challenging because tissue intensities of MRIs are not always consistent within the image, e.g., due to inhomogeneity artifacts. Clustering or unmixing methods can be used on a per image basis. 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. Here we propose the first unsupervised CNN autoencoder for simultaneous WMH and tissue segmentation. A linear mixture model is incorporated into the CNN architecture, and we introduce a novel way to regularize and train the CNN to provide a meaningful solution. Robustness to signal inhomogeneity is made inherent in the CNN, by introducing iterative tissue and inhomogeneity correction steps during training to better preserve WMHs. Finally, we show that the CNN generalizes well for unseen images after training.

In Chapter 3 we introduce **Large-scale parcellation of the ventricular system using convolutional neural networks**: Conventional multi-atlas registration methods can fail when presented with severely enlarged ventricles. Furthermore, they require excessive processing time, which is limiting when exploring a very large data set, such as the brain MRIs of the AGES-Reykjavik cohort. By generating training data using RUDOLPH [52], a multi-atlas segmentation method specifically designed for extreme ventricle enlargement, and selecting subjects that cover the entire spectrum of ventricle sizes in the AGES-Reykjavik cohort, we showed that a CNN can replicate the robustness of RUDOLPH to ventricle size variability while generating results 360x faster.

In Chapter 4 we introduce a **Skull-stripping U-net for brain MRIs**: Brain extraction is an important step in many brain image analysis pipelines. Accuracy is often not consistent between subjects due to atrophy, enlarged ventricles, TBI, or random errors. Generating manual delineations of intracranial matter for the development of a robust CNN is immensely time-consuming. Here we show qualitatively that by training a CNN on selected brainmasks generated with a multi-contrast atlas based skull-stripping method, we gain the following: 1) we eliminate the need for manual delineation of intracranial matter; 2) we increase the accuracy of results; 3) the CNN learns features that generalize better than the original atlas based method; and 4) the brain masks are generated much faster.

In Chapter 5 we introduce **A joint ventricle and WMH segmentation from MRI for evaluation of age-related changes in the brain**: The co-occurrence of ventricular enlargement and WMHs in neurodegenerative diseases and in aging brains often requires investigators to take both into account when studying the brain. Here we build upon our previous work and present a hybrid multi-atlas segmentation and convolutional autoencoder approach. We use an unsupervised convolutional autoencoder to generate a standardized image of grey matter, white matter, CSF, and WMHs, which, in conjunction with labels generated by a multi-atlas segmentation approach, is then fed into another CNN for parcellating the ventricular system. Hence, our approach does not depend on manually delineated training data for new data sets. Furthermore, we show that the proposed ventricle CNN is not as dependent on the type of MRI sequences used as input to the pipeline. The proposed method is trained and validated on healthy elderly subjects from the AGES-Reykjavik cohort. The method is further validated on NPH patients imaged at a different site with a different MRI scanner, using only unsupervised fine-tuning to generate the standardized images. Moreover, no fine-tuning is needed for the ventricle CNN to work on the NPH data set due to the use of standardized images, despite cases of considerably larger ventricle sizes than that of the initial training set coming from the AGES-Reykjavik cohort. Therefore, this work presents a solution to the problem of automatic segmentation methods failing when presented with images where the MRI scanners and MRI parameters used are different from the training set. We also note that the external validation set comprises MRIs of NPH patients with severe pathology, which further demonstrates the robustness of the segmentation pipeline. The method is used to segment brain MRIs of 2401 subjects in the AGES-Reykjavik cohort, who showed up for two MRI sessions, to explore the mean and standard deviation of ventricular volumes and WMH loads in relation to age. Ventricle volume and WMH load depend on multiple factors. Therefore, to explore the individual association between the ventricle size and WMH load we use multiple linear regression models that take several confounders into account.

Chapter 6 concludes the dissertation with a discussion about limitations of this work, future directions, and an overall conclusion.

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

2.1 Introduction

A fundamental issue in supervised learning is the lack of ground truth labels. In brain MRI segmentation a manual delineation by an expert in neuroanatomy is still the gold standard for ground truth segmentations. Obtaining manually segmented images is laborious and slow, and hence often impractical for generating new training data for CNNs. Tissue and lesion segmentation with unsupervised methods can provide enormous benefit by eliminating the need for manual delineations, which are often impractical or impossible to obtain with sufficient accuracy, for reasons such as their complex structure, thin lines, and partial volume effects. A combination of linear unmixing and a neural network autoencoder has been proposed in hyperspectral unmixing of remote sensing images [99, 100]. 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 sequences.

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 published in conference format at the SPIE Medical Imaging conference [4] and was followed by journal publication in NeuroImage:Clinical [60] with substantial improvements 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 in Appendix B.1.

2.2 Materials

Two data sets were used for the evaluation of SegAE; MRIs from the AGES-Reykjavik study [54] (see Appendix A.1 for details), and the WMH challenge [64] 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 (see Appendix A.3 for details).

For developmental purposes, we randomly selected 60 subjects from the AGES-Reykjavik cohort; thereof 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.

We submitted our method to the WMH challenge [64]. 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 1.11 in Appendix A.3 shows an overview of how the data set is separated into training and test sets. Table 1.12 in Appendix A.3 shows scanning parameters for the 5 scanners.

2.2.1 Preprocessing

AGES-Reykjavik: Images were preprocessed using standard preprocessing procedures: Resampling to $0.8\times0.8\times0.8$ mm^3 voxel size, rigid registration to the MNI-ICBM152 template [46], and skull removal using MONSTR [101]. 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 2.5 and 2.6.
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 [64]. 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 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 in [60]). As for the AGES-Reykjavik data set, inhomogeneity correction was integrated into the segmentation method (see Section 2.5).

2.3 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 2.7 for details on the training procedure). The autoencoder is constrained to reconstruct the corresponding image patches with a linear unmixing model,

$$\hat{\boldsymbol{Y}}_c = \sum_{i=1}^{M} w_{i,c} \boldsymbol{S}_i, \tag{3}$$

where $\hat{\boldsymbol{Y}}_c$ is one channel of the output, $w_{i,c} \in \mathbb{R}_{\geq 0}$ are the weights, $\boldsymbol{S}_i$ is the soft segmentation of materials (such as WMHs, WM, GM, CSF and meninges), M is the number of materials to be estimated, $\boldsymbol{S}_i \geq 0$ and $\sum_{i=1}^{M} \boldsymbol{S}_i = \boldsymbol{B}$, where $\boldsymbol{B}$ 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 $\boldsymbol{S}$ are enforced with a Softmax activation function. A patch-wise brainmask obtained 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 2.4), the Softmax-layer outputs a soft segmentation of the materials present in the images.

The autoencoder consists of 3D convolutional layers followed by 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 2.11.

2.4 Loss and regularization

The Cosine proximity function,

$$f(\boldsymbol{y},\hat{\boldsymbol{y}}) = \frac{\boldsymbol{y} \cdot \hat{\boldsymbol{y}}}{||\boldsymbol{y}||_2 ||\hat{\boldsymbol{y}}||_2}, \tag{4}$$

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

$$L(\boldsymbol{Y},\hat{\boldsymbol{Y}}) = -\frac{1}{C} \sum_{c=1}^{C} \left(f(\mathrm{vec}(\boldsymbol{Y}_c), \mathrm{vec}(\hat{\boldsymbol{Y}}_c)) + f(\mathrm{vec}(K * \boldsymbol{Y}_c), \mathrm{vec}(K * \hat{\boldsymbol{Y}}_c)) \right), \tag{5}$$

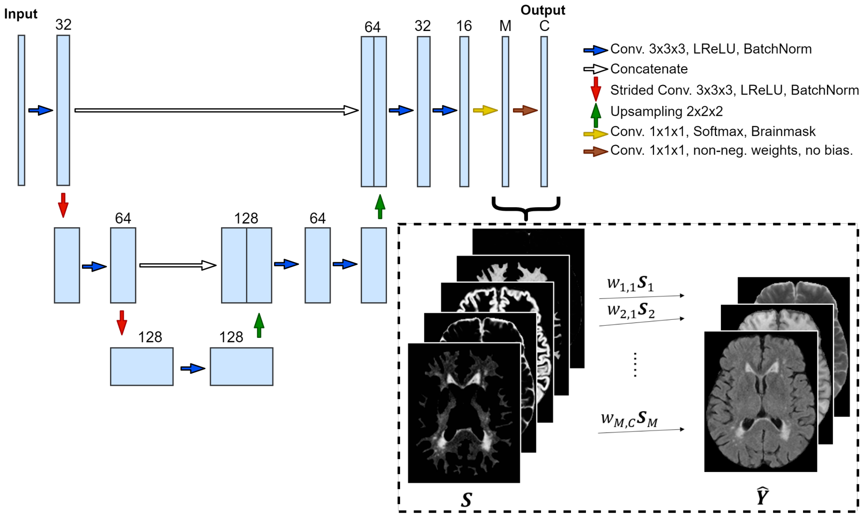

Input
32
64
32
16
M
Output
C
Conv. 3x3x3, LReLU, BatchNorm
Concatenate
Strided Conv. 3x3x3, LReLU, BatchNorm
Upsampling 2x2x2
Conv. 1x1x1, Softmax, Brainmask
Conv. 1x1x1, non-neg. weights, no bias.
64
128
64
128
128
$w_{1,1}S_1$
$w_{2,1}S_2$
$w_{M,C}S_M$
S
$\hat{Y}$

Figure 2.11. 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 $\hat{\boldsymbol{Y}}$ to be a weighted sum of the Softmax outputs $\boldsymbol{S}$. 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).

where C is the number of channels in $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$, 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,

$$\Omega(\boldsymbol{S}) = \frac{\alpha}{M} \sum_{i=1}^{M} \sum_{j=1}^{M} f(\text{vec}(\boldsymbol{S}_i), \text{vec}(\boldsymbol{S}_j)), \tag{6}$$

where α is the regularization parameter.

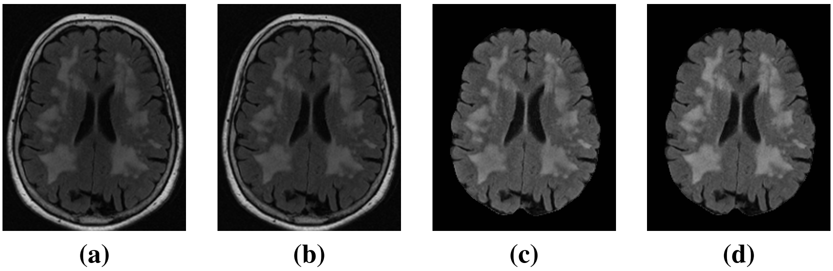


Figure 2.12. The figure shows the effect of N4 bias correction on a FLAIR image with a large lesion load. ***(a)*** *The original FLAIR image before skullstripping;* ***(b)*** *after N4 bias correction (with skull);* ***(c)*** *After N4 bias correction (without skull); and* ***(d)*** *After skull-stripping and bias correction using pure-tissue probability mask.*

2.5 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 [48]. 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 2.12 **(c)** and a more detailed comparison in Appendix B.2]. Hence, to avoid this degradation, we alternated between using N4 bias correction and tissue segmentation to obtain "pure-tissue" probability masks, as suggested in [102]. 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 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. These steps were repeated two times which resulted in accurate SegAE segmentations observed in the validation set. In general, the number of iterations for the iterative inhomogeneity correction can be determined by repetition until a metric such as the DSC stops improving.

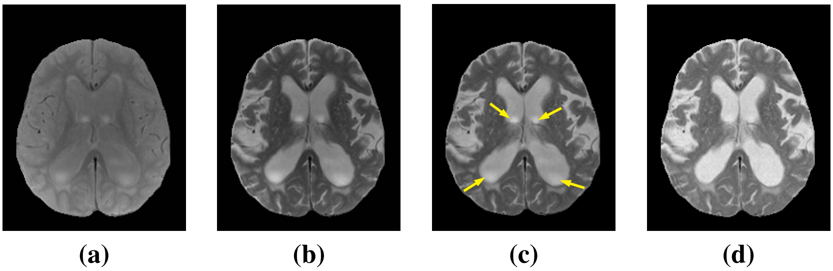


Figure 2.13. Image enhancement of a T2-w image using a PD image. ***(a)*** *and* ***(b)*** *show the original PD and T2-w images, respectively;* ***(c)*** *shows the T2-w image after N4 bias correction with a pure-tissue probability mask; and* ***(d)*** *shows an enhanced image (T2$_{PD}$).*

2.6 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 2.13 **(a)** and **(b)**]. N4 bias correction using a pure-tissue probability mask was not sufficient to eliminate these artifacts (see Figure 2.13 **(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 2.13 **(d)** and Figure 2.14),

$$\boldsymbol{I}_{new} = \boldsymbol{I}_{orig} \odot (\mathrm{Max}(\boldsymbol{I}_{PD})\boldsymbol{J} - \boldsymbol{I}_{PD}), \tag{7}$$

where $\boldsymbol{I}_{new}$ is the enhanced image, $\boldsymbol{I}_{orig}$ is the original T1-w or T2-w image, $\boldsymbol{I}_{PD}$ is the original PD-w image, $\boldsymbol{J}$ is a matrix of ones of the same size as the PD-w image, and $\odot$ denotes an element-wise multiplication. Multiplying the 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 2.14). We will refer to the enhanced T1-w and T2-w images using PD-w images as T1$_{PD}$ and T2$_{PD}$, respectively.

2.7 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 2.3 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 α was 0.0075

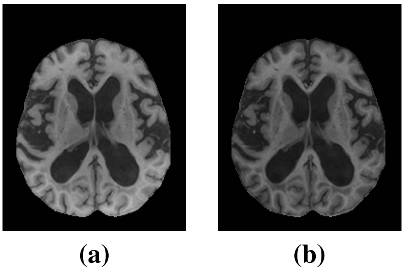


Figure 2.14. Image enhancement of a T1-w image using a PD image. ***(a)*** *shows the original T1-w image and* ***(b)*** *shows an enhanced image (T1$_{PD}$).*

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 2.3. 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 [103], with Nesterov momentum [104], 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 [105] 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 [106] values.

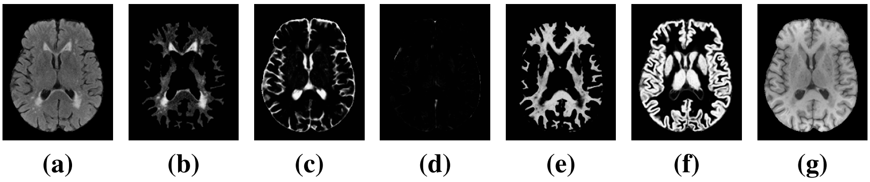


Figure 2.15. The tissue and WMH segmentation output from SegAE. ***(a)*** *and* ***(g)*** *show the original FLAIR and T1-w images respectively, and* ***(b)-(f)*** *show the segmentations of WMH, CSF, meninges that remain after skullstripping, WM, and GM, respectively.*

2.8 Prediction and post-processing

After training, the 5 Softmax output volumes ($\boldsymbol{S}$ in Figure 2.11) were used for prediction, while the reconstructed images ($\hat{\boldsymbol{Y}}$ in Figure 2.11) 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 2.15).

In this dissertation 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) [107] on the training data [108], and structures smaller than 3 voxels were removed from the segmentation results from the WMH challenge data due to noise in the cerebellum.

2.9 Evaluation

2.9.1 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)* [107]
  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:
  $$d_{\mathrm{H}}(X,Y) = \max\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\},$$
  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, the best linear fit was identified between the predicted and manually delineated volumes and the Pearson's correlation coefficient (r) was used for comparison.

2.9.2 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) [109] and the Lesion Prediction Algorithm (LPA) [110] as implemented in the LST toolbox[1] version 2.0.15, and one method developed previously for the AGES-Reykjavik data set based on an artificial neural network classifier (ANNC) [31]:

- *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 (κ) for an initial binary map, which is grown to include voxels that appear hyperintense in the FLAIR image for a final lesion probability map [109]. We used $\kappa = 0.1$ as determined by the result on our 5 validation images.

[1] www.statisticalmodelling.de/lst.html

Table 2.4. 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 (± 0.13)*	10.16 (± 10.40)	60.49 (± 29.75)*	0.44 (± 0.12)*	0.39 (± 0.10)
LGA	0.66 (± 0.15)*	15.22 (± 9.93)	**26.50** (± 23.58)	0.29 (± 0.12)	0.36* (± 0.11)
LPA	0.66 (± 0.19)	**9.20** (± 6.56)	62.28 (± 73.75)	0.53 (± 0.27)	0.40 (± 0.20)
SegAE	**0.77** (± 0.11)	10.97 (± 11.45)	33.31 (± 36.30)	**0.64** (± 0.19)	**0.47** (± 0.09)

- *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 [110].
- *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 voxel-wise intensities of FLAIR, T1-w, T2-w, and PD-w images and the classifier was trained on 11 manually annotated subjects [31].

2.9.3 Evaluation on the AGES-Reykjavik data set

Figure 2.16 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 2.4 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 2.17, top), as well as corresponding DSC of the four methods against the manual masks (see Figure 2.17, 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 2.18. ANNC and SegAE achieve $r = 0.98$, while LGA and LPA have $r = 0.78$ and $r = 0.73$, respectively.

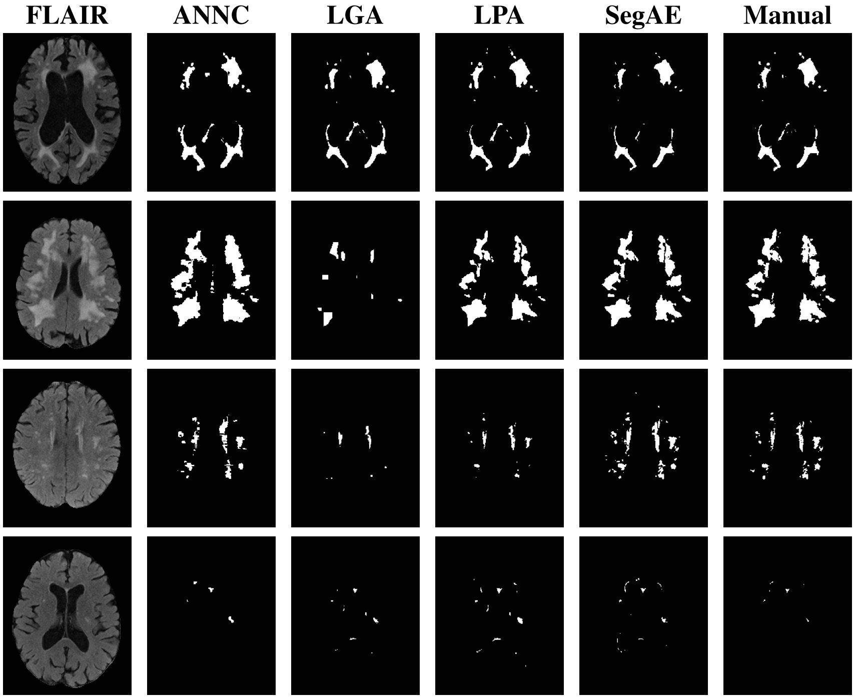


Figure 2.16. Visual comparison of the four methods with a manual rater for four different subjects, two with the largest and second largest lesion load (1st and 2nd row), one with a medium lesion load (3rd row), and one with the smallest lesion load (4th row).

2.9.4 Evaluation on the WMH challenge data set

Figure 2.19 shows a visual comparison between the WMH segmentation of SegAE and the manually delineated masks for 3 subjects in the WMH challenge training set. Table 2.5 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[2]. Furthermore, the website shows boxplots for all 5 metrics comparing the results obtained for each scanner.

2.10 Discussion

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

[2] https://wmh.isi.uu.nl/results/himinn/

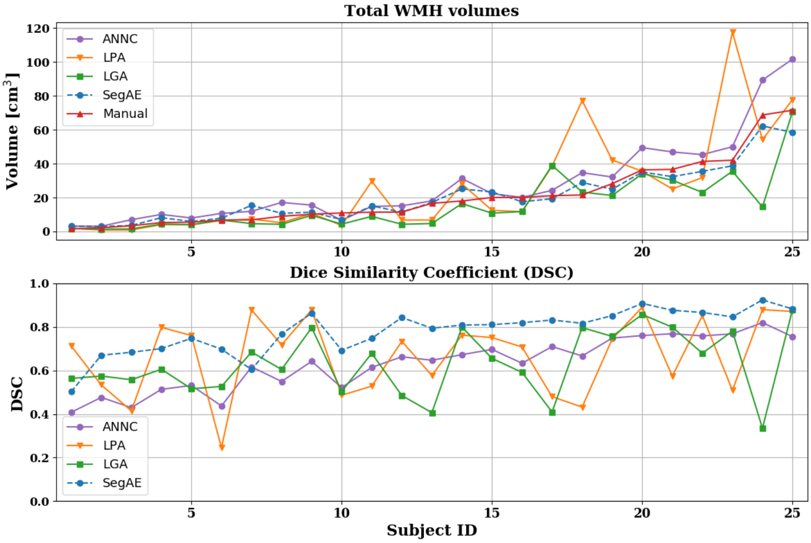


Figure 2.17. The top graph shows the overall WMH volume for the manual masks (red) and masks generated by ANNC (purple), LPA (orange), LGA (green), SegAE (blue, dotted), ordered by the volume of the manual masks. The bottom graph shows the DSC for the same methods compared with the manual masks.

Table 2.5. 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.

	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

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 2.20). 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

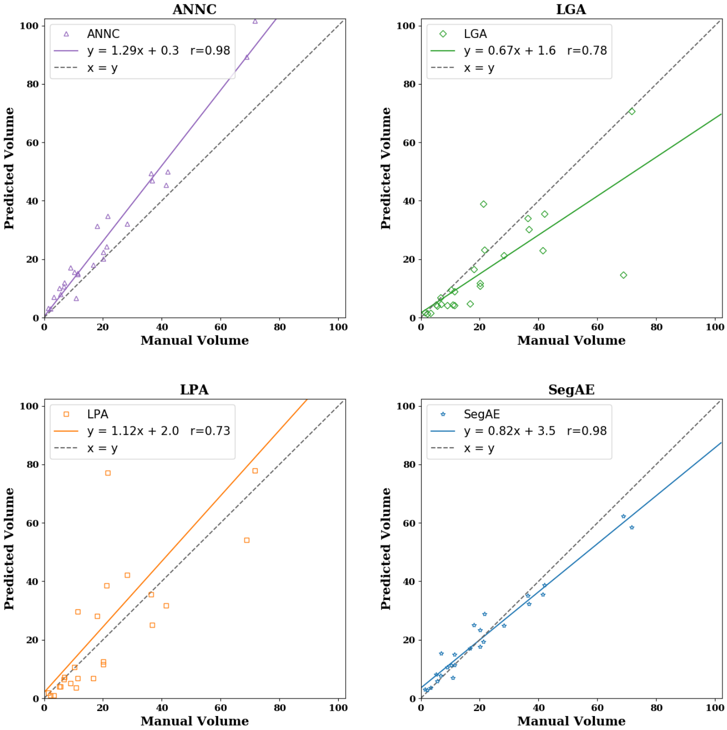


Figure 2.18. Predicted lesion volumes versus manual lesion volumes for the four methods. The solid lines show a linear fit of the points and the dashed black line has unit slope. Numbers are in cm^3. Slope, intercept, and Pearson's correlation coefficient between manual and predicted masks can be seen for the different methods.

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 DSC, L-TPR, and L-F1 scores, while LPA achieved the best average H95 score (cf. Table 2.4). A larger test set would be preferable to increase statistical power. 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

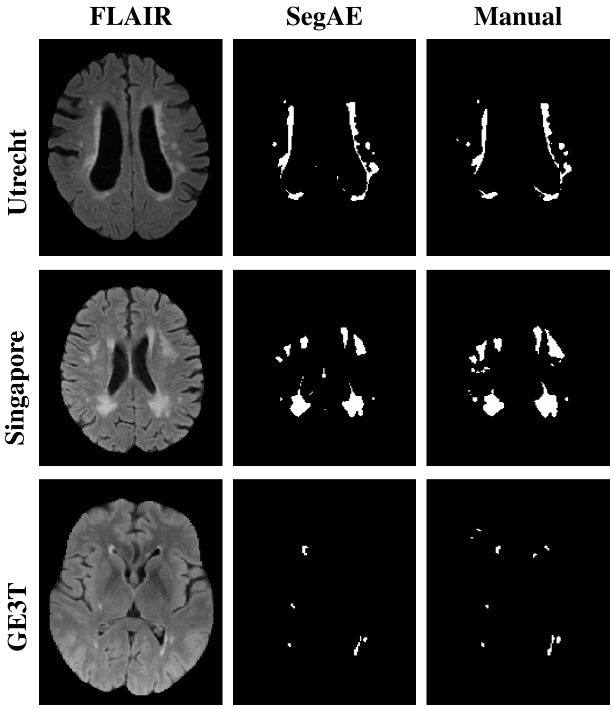


Figure 2.19. Visual comparison between the WMH segmentation of SegAE and the manually delineated masks for subjects in the WMH challenge training set. The top row shows the first subject (ID: 0) from the Utrecht scanner, the middle row shows the first subject from the Singapore scanner (ID: 50) and the bottom row shows the first subject in the GE3T scanner (ID: 100).

score despite having a volume correlation of only 0.78, seemingly because the AVD score penalizes undersegmentations less than oversegmentations, as mentioned in [64]. SegAE and ANNC achieved the highest volume correlation ($r = 0.98$), however, ANNC seems to systematically overestimate the lesion volumes, as indicated in Figure 2.18, 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 2.17. Bottom part of Figure 2.17 demonstrates the robustness of SegAE to a variety of WMH volumes and in Figure 2.16, 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 2.5. On the challenge website[3], methods are ranked according to the average

[3] Team Himinn. https://wmh.isi.uu.nl/results

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 2.4 and 2.5 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 2.43 in Appendix B.4 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 2.19, 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 set are caused by errors in skullstripping, since the test set and its results are blinded. The much higher H95 and AVD for some images from Utrecht and the AMS PETMR results may suggest that this might be the case.

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 MS and TBI. Methods for unsupervised segmentation of FLAIR hyperintensities are often used interchangeably [96], and we believe that the proposed method should be able to segment any lesions with similar intensities in the MRI sequences that we use.

2.11 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 2.5. The scores can be compared

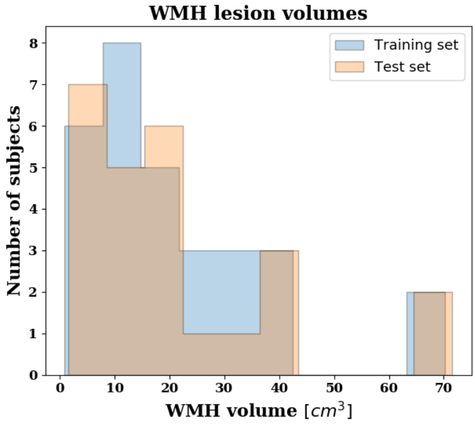


Figure 2.20. A histogram showing the WMH lesion volumes of the AGES-Reykjavik training (blue) and test (peach) sets. The volumes were predicted from SegAE since manual delineations do not exist for the training images.

with any method sent to the WMH segmentation challenge via the WMH challenge website[4].

[4] Team Himinn. https://wmh.isi.uu.nl/results

3 Large-scale parcellation of the ventricular system using convolutional neural networks

3.1 Introduction

Enlarged ventricles are a marker of several brain diseases; however, they are also associated with normal aging. Better understanding of the distribution of ventricular sizes in a large population would be of great clinical value to robustly define imaging markers that distinguish health and disease. The AGES-Reykjavik study includes magnetic resonance imaging scans of 4811 individuals from an elderly Icelandic population. Automated brain segmentation algorithms are necessary to analyze such a large data set but state-of-the-art algorithms often require long processing times or depend on large manually annotated data sets when based on deep learning approaches. In an effort to increase robustness, decrease processing time, and avoid tedious manual delineations, we selected 60 subjects with a large range of ventricle sizes and generated training labels using an automated whole brain segmentation algorithm designed for brains with ventriculomegaly. Lesion labels were added to the training labels, which were subsequently used to train a patch-based three-dimensional U-net CNN for very fast and robust labeling of the remaining subjects. Comparisons with ground truth manual labels demonstrate that the proposed method yields robust segmentation and labeling of the four main sub-compartments of the ventricular system.

3.2 Training data synthesis

For developmental purposes we selected 90 subjects (age 67-92) from the AGES-Reykjavik cohort. These subjects were selected based on the ventricle volume estimated with an atlas based method in [54]. The quality of this ventricle segmentation was not assessed systematically, however, it was sufficient to roughly group subjects into three groups of 30: Group 1 containing the smallest, Group 2 the medium, and Group 3 the largest ventricle sizes. This way our developmental sample covered the entire spectrum of ventricle sizes of the AGES-Reykjavik cohort (smallest to largest), with no overlap between the ventricle volumes in the three groups. RUDOLPH [52], a multi-atlas segmentation method specifically designed for extreme ventricle enlargement, was used to segment a total of 60 subjects, i.e. 20 subjects from each of the three groups described above into 138 regions [52,53]. For our training data set, this number of labels was reduced to 8 labels: background, left and right lateral ventricles, third and fourth ventricles, gray and white matter, and sulcal CSF. Included in these 8 labels were the four ventricles we wish to evaluate in addition to a tissue classification that helped the CNN model distinguish the ventricular system from other parts of the brain. RUDOLPH does not provide a lesion label, although many subjects in the AGES-Reykjavik data set have white matter lesions adjacent to the lateral ventricles. We used a preliminary version of SegAE to generate lesion labels that we added to the training data [111]. These modifications resulted in a final training set of 60 subjects, each comprising 9 labels.

Our training set for the CNN comprised the T1-w and T2-w MRIs, and the corresponding segmentations. As a last step before training the CNN, each T1-w and T2-w images were intensity normalized by dividing by the 99th percentile of the non-zero elements of the image. The images were cropped to the smallest cuboid containing the entire brain and two-channel 80×80×80 voxel patches were extracted with a 40 voxel stride over the entire volume.

3.3 Network architecture and training procedure

We used a 3D U-net architecture [77, 112] on four resolution scales to process large patches of T_1-w and T_2-w images. We downsampled using strided convolutions instead of max pooling to preserve spatial structure, and we included batch normalization layers to speed up convergence [112, 113]. All layers use 3×3×3 kernels, except the final output layer, which uses 1×1×1 kernels. The CNN architecture is shown in Figure 3.21.

The network was trained on 70% of the extracted training patches that included the most voxels belonging to classes of high class weight. Doing this reduced training time and helped the model learn underrepresented classes. A weighted categorical crossentropy loss function was applied using the class weights of the remaining patches. A GTX1080 Ti GPU was used to train the model for 150 epochs with a learning rate of $3 \cdot 10^5$, and another 50 epochs with a learning rate of $3 \cdot 10^6$, using an Adam optimizer [103] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Predictions were made with a patch overlap of 20 voxels in each dimension by averaging corresponding voxels and taking argmax of the class probability for each voxel.

3.4 Evaluation

A total of 15 subjects that were not in the training set (five from each Group described above) were manually delineated to use as ground truth labels when validating the proposed method. This was done by first identifying the ventricular system from the T_1-w image followed by a parcellation of the ventricular mask into four sub-compartments (left and right lateral ventricles, and 3rd and 4th ventricles). These were reviewed by an expert in neuroanatomy, who performed corrections where needed. The 15 subjects were processed using the proposed CNN model, and two state-of-the-art brain segmentation methods: FreeSurfer 6.0 [51] and RUDOLPH. To speed up the processing

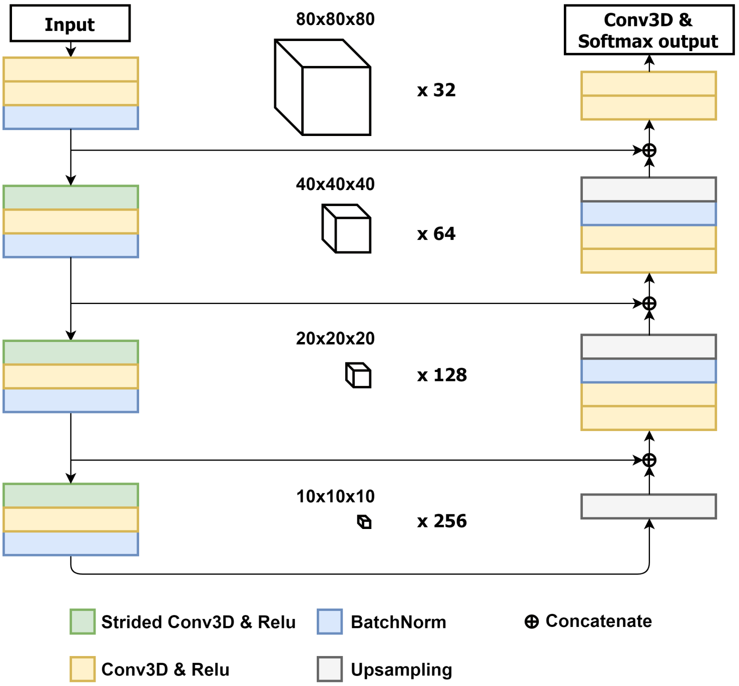


Figure 3.21. The CNN model architecture. The number of channels used for convolutions is denoted next to the boxes representing each resolution scale. All convolutional layers use kernels of size 3×3×3, except the final output layer which performs 1×1×1 convolutions and the number of channels equals the number of output labels.

with FreeSurfer we provided the algorithm with skull-stripped data. We ran FreeSurfer with the `-bigventricles` flag as well as with the default settings, to see which performed better on a range of ventricle sizes. The CNN method processed the subjects in 60 seconds on average while the processing times of RUDOLPH and FreeSurfer with the `-bigventricles` flag were 6 and 7.5 hours, respectively, on average.

We computed the total ventricle volume estimated by the three methods and compared with the volume of the manual masks (see Figure 3.22, top). Then the Dice coefficient [107] was used to measure the overlap of the ventricle labels provided by the three methods against the manual masks (see Figure 3.22, bottom). Figure 3.22 shows the results for every test subject ordered by the volume of the manual masks and indicates the robustness of the proposed CNN method on the entire spectrum of ventricle sizes. We note that FreeSurfer with the `-bigventricles` flag performed much better on large ventricles than FreeSurfer with default settings, while maintaining high accuracy on small ventricles, so hereafter, only the results generated with the `-bigventricles` flag will be used in our statistical analysis. Table 3.6 shows the Dice coefficients for the whole ventricular system and each sub-compartment. Figure 3.23 shows the segmentations of two subjects, one with the largest ventricles in the

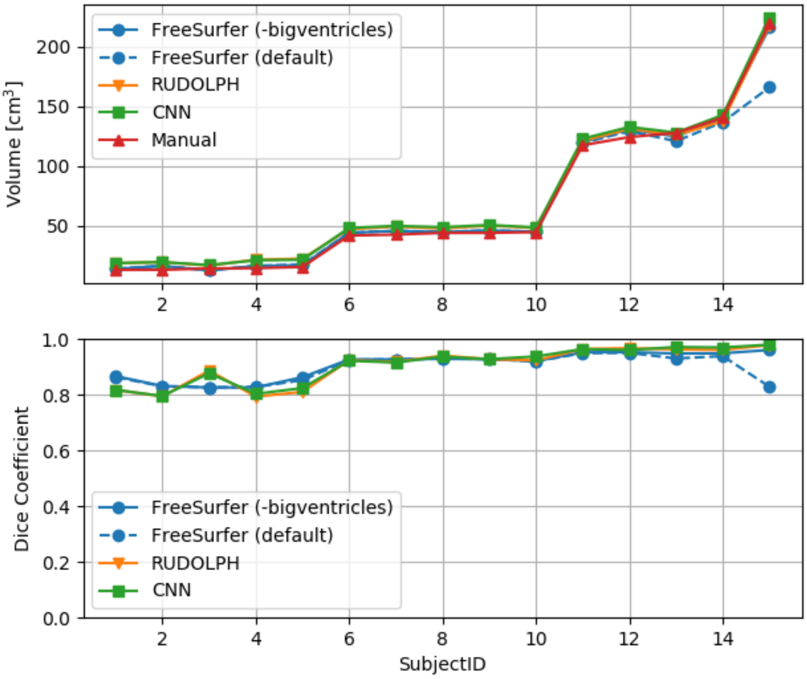


Figure 3.22. The top plot shows the overall ventricular volume for the manual masks (red) and masks generated by FreeSurfer (blue) showing both results generated using the `-bigventricles` *flag (solid line) and default settings (dotted line), RUDOLPH (orange), and CNN (green), ordered by volume of the manual masks. The bottom plot shows the Dice score for the same methods compared with the manual masks for the whole ventricular system.*

evaluation set (See Figure 3.23, top row) and one with small ventricles (See Figure 3.23, bottom row) showing the failure of FreeSurfer on the largest ventricles when using the default settings. A paired Wilcoxon signed-rank test (without correction for multiple comparisons) was used to compare CNN vs. FreeSurfer, CNN vs. RUDOLPH, and RUDOLPH vs. FreeSurfer. We found no significant differences between these methods in terms of Dice overlap ($p > 0.005$), except that the CNN method and RUDOLPH performed better than FreeSurfer on the 4th ventricle in this data set ($p < 0.005$) demonstrating the viability of the proposed method in producing fast and robust segmentations without any compromise in terms of accuracy and without the need for manual labels.

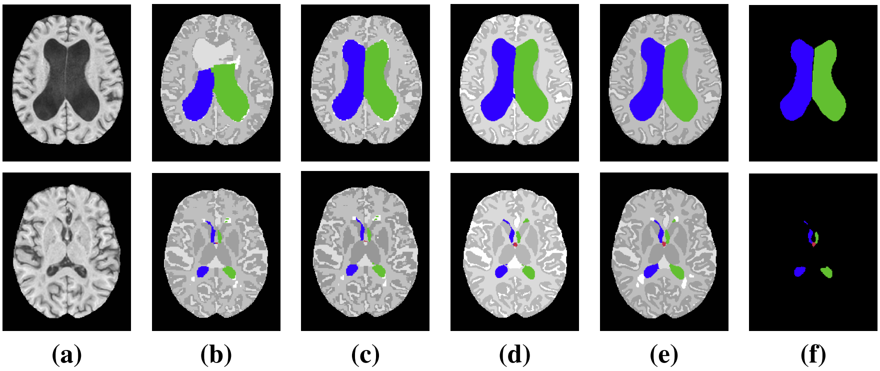


Figure 3.23. Visual comparison of the proposed CNN method with two state-of-the-art methods. ***(a)*** *T1-w image of a test subject and the segmentations generated by* ***(b)*** *FreeSurfer using default settings,* ***(c)*** *FreeSurfer with the* `-bigventricles` *flag,* ***(d)*** *RUDOLPH,* ***(e)*** *CNN,* ***(f)*** *a manual rater for the left (green) and right (blue) lateral ventricles, and the third ventricle (red). The fourth ventricle is not visible in these slices. The remaining labels generated by each method are shown in grayscale.*

Table 3.6. The mean Dice coefficient and standard deviation for the three methods on the entire ventricle system (Entire), the left lateral ventricle (LLV), the right lateral ventricle (RLV), the third ventricle (3rd) and the fourth ventricle (4th). The `-bigventricles` *flag was used to generate the FreeSurfer results. Asterisk (*) denotes values that are significantly different from the proposed CNN (p < 0.005).*

	FreeSurfer	RUDOLPH	CNN
Entire	0.906 (± 0.048)	0.904 (± 0.065)	0.906 (± 0.064)
LLV	0.918 (± 0.043)	0.904 (± 0.074)	0.908 (± 0.066)
RLV	0.915 (± 0.044)	0.908 (± 0.062)	0.909 (± 0.062)
3rd	0.853 (± 0.045)	0.884 (± 0.051)	0.887 (± 0.051)
4th	0.689 (± 0.081)*	0.807 (± 0.067)	0.790 (± 0.054)

3.5 Conclusions and discussion

We have presented a method for fast and robust segmentation and parcellation of the ventricular system in a large-scale study of the elderly, a population with high variability in ventricle size. A 3D patch-based U-net CNN model was trained on images that captured a large spectrum of ventricle sizes in our cohort from the AGES-Reykjavik study, using segmentations generated by RUDOLPH — an algorithm developed for robust segmentation and labeling of brains with ventriculomegaly [52]. Furthermore, lesion labels were added to the training images using an unsupervised lesion segmentation method [111] to prevent over-classification of the ventricles onto adjacent lesions.

The CNN segmentation of the entire ventricular system and parcellation into its four main sub-compartments (left and right lateral, third, and fourth ventricles) was evaluated on 15 manually labelled subjects. We found no significant differences between the proposed CNN method and two state-of-the-art methods in terms of accuracy, except the CNN method and RUDOLPH performed better than FreeSurfer on the 4th ventricle in this data set ($p < 0.005$). However, the CNN was significantly faster (360x speedup compared with RUDOLPH), without any compromise in terms of segmentation accuracy.

Figures 3.22 and 3.23 highlight the importance of using the `-bigventricles` flag when running FreeSurfer on a data set where enlarged ventricles can be expected due to high age or neurological diseases.

In summary, the proposed method provides segmentation and parcellation of the ventricular system that is robust to a wide spectrum of ventricle sizes, from healthy ventricles to severe ventriculomegaly. After training, the method generates results in a matter of seconds — a critical feature when studying large data sets — while the state-of-the-art brain segmentation algorithms take hours to process a single subject.

3.6 Conclusions

We have presented a method for fast and robust segmentation and parcellation of the ventricular system in a large-scale study of the elderly, a population with high variability in ventricle size. A 3D patch-based U-net CNN model was trained on images that captured a large spectrum of ventricle sizes in our cohort from the AGES-Reykjavik study, using segmentations generated by RUDOLPH — an algorithm developed for robust segmentation and labeling of brains with ventriculomegaly [52]. Furthermore, lesion labels were added to the training images using SegAE to prevent over-classification of the ventricles onto adjacent lesions.

The CNN segmentation of the entire ventricular system and parcellation into its four main sub-compartments (left and right lateral, third, and fourth ventricles) was evaluated on 15 manually labelled subjects. We found no significant differences between the proposed CNN method and two state-of-the-art methods in terms of accuracy, except the CNN method and RUDOLPH performed better than FreeSurfer on the 4th ventricle in this data set ($p < 0.005$). However, the CNN was significantly faster (360x speedup compared with RUDOLPH), without any compromise in terms of segmentation accuracy.

4 Skull-stripping U-net for brain MRIs

4.1 Introduction

Isolating the intracranial matter from brain MRIs is important for subsequent processing in many image processing pipelines. Multiple methods have been developed for brain extraction [114–116], however their accuracy is often not consistent between subjects due to atrophy, enlarged ventricles, traumatic brain injury, or random errors. Atlas based methods are common, however, many atlases are needed to capture the wide anatomical variability [101]. One atlas based brain extraction method is the Multi-cONtrast brain STRipping (MONSTR) [101] method, which was developed to be robust to traumatic brain injury by utilizing multi-contrast information. Although highly accurate in most cases, and extensively validated previously against state-of-the art methods [101], we found that MONSTR fails in some subjects of the AGES-Reykjavik data set. State-of-the art methods for most brain segmentation tasks are based on CNNs [117–119]. They are usually more consistent because they are robust to random errors in the training set, which usually contains multiple manually delineated images. CNN based methods can also be much faster than methods based on multi-atlas registration with multiple degrees of freedom. However, manually delineated brainmasks are often not available and are time-consuming to generate. Here we show that by training a three-dimensional U-net CNN using the brainmasks generated by MONSTR as training data we can generate more accurate brainmasks than MONSTR, if images with the most visible errors are removed from the training set. This training method can be used when a skull-stripping method, such as MONSTR, generates near perfect brainmasks for a large subset of the data at hand, which can be used for training. Alternatively, brainmasks can be manually corrected, however, this is a much more time-consuming approach.

4.2 Preparation of training data and CNN architecture

The development and evaluation of the skull-stripping U-net was performed using brain MRIs from the AGES-Reykjavik data set (cf. Appendix A.1). The brainmasks used for supervised training of the skullstripping CNN were generated by the MONSTR method [101]. Brainmask atlases for MONSTR were created by manually delineating the brain in 6 subjects from our AGES-Reykjavik development set of 120 subjects. Manual inspection of 60 of the generated MONSTR brainmasks led to the exclusion of 13 masks due to skullstripping failures (see Figure 4.24); hence the remaining 47 masks

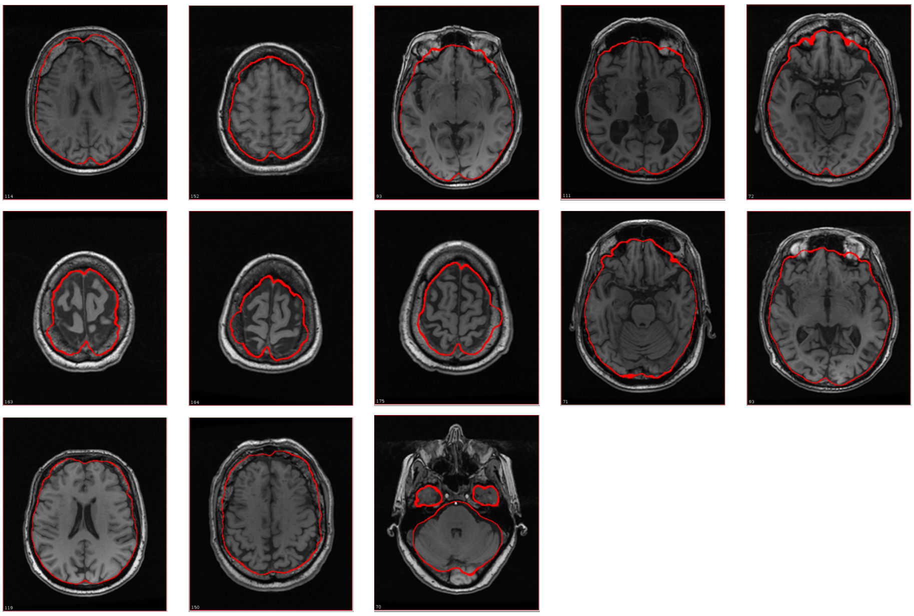


Figure 4.24. Erroneous skull-stripping results from MONSTR that were removed from our training set. The figure shows T1-w images and the corresponding skull-stripping boundaries generated by MONSTR (red).

were used for training. Our training set comprised the T1-w, T2-w, and FLAIR images and the corresponding brainmasks. The network architecture can be seen in Figure 4.25.

4.3 Training

The 47 training images were intensity normalized by dividing by the 99th percentile of the non-zero elements of the image and 80×80×80 voxel patches were extracted with a 40 voxel stride. A weighted categorical cross-entropy loss function was used, wherein the weights were determined using class weights [120]. The network was trained for 200 epochs with a learning rate of $1 \cdot 10^5$ using the Adam optimizer [103] with Nesterov momentum [104], with $\beta_1 = 0.9$, $\beta_2 = 0.999$, schedule decay of 0.004, and a batch size of 5.

4.4 Evaluation

The evaluation of our skull-stripping method was twofold: First, we compared the results of our method to results generated by MONSTR on the development set of 120 subjects. Second, we compared the ICVs of 2401 subjects on MRI scans that were

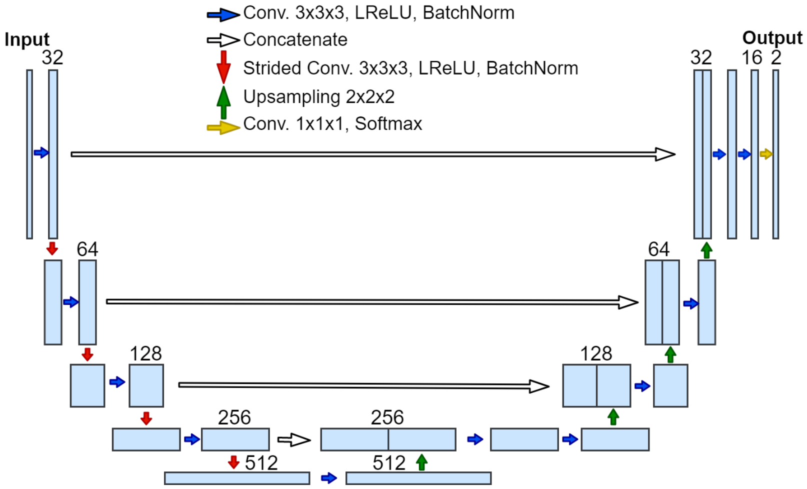


Figure 4.25. The proposed CNN architecture for the skullstripping U-net. The input comprises large 3D patches from FLAIR, T1-w, and T2-w images. Kernels of size 3×3×3 are used in all convolutional layers except size 1×1×1 is used in the final two layers.

acquired at two different time points (scans acquired 5 years apart on average). We visually inspected 9 slices of each of these 2401 subjects to detect failures and their causes.

Figure 4.26 shows a histogram of the Dice dissimilarity (one minus the Dice similarity coefficient [107]) between the 120 MONSTR brainmasks and the U-net brainmasks. The 8 subjects with the highest Dice dissimilarity were inspected. Three of these 8 subjects were a part of the MONSTR results that were removed from the training set. Figure 4.27 shows one slice from each of these 8 subjects that have the largest error. Of those 8 brainmasks, the brainmask with the smallest error had small enough error so that we concluded that a further comparison of our U-net and MONSTR results was not necessary.

One limitation to this evaluation strategy is that by comparing the overlap of the masks generated by the U-net to the masks from MONSTR we would not expect to find large values for Dice dissimilarity if both MONSTR and the U-net systematically fail in the same way. Therefore, we visually inspected 9 slices from each of the 2401 subjects with longitudinal MRI scans and found that the U-net was very robust, except in cases when: 1) One or more MRI sequences had registration errors (24 registration errors in total); and/or 2) there were visible skullstripping errors (9 cases of which 3 were caused by registration errors). These registration errors are marked on Figure 4.28, which compares the brainmask volumes at two timepoints. The figure shows that the predicted ICV is generally very consistent, except the few cases that had registration and/or skullstripping failures.

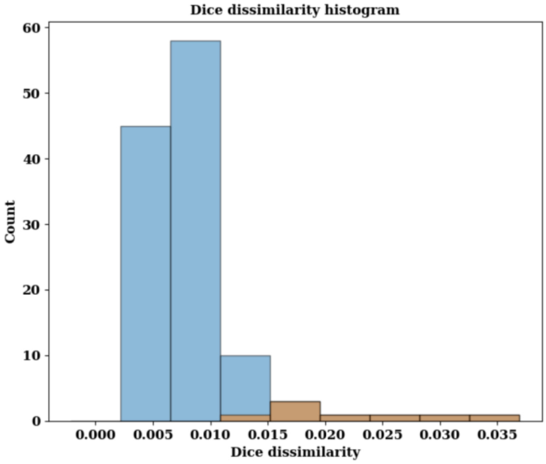
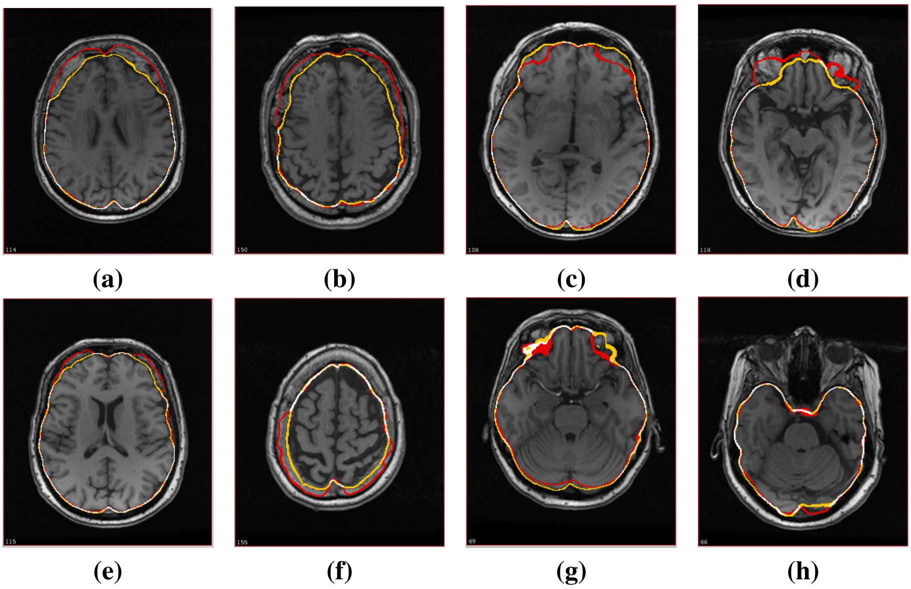


Figure 4.26. A histogram showing the number of the skullstripping U-net brain segmentations within each group of Dice dissimilarity when compared with the MONSTR segmentations. The brown columns show the values of the 8 brain segmentations with the highest dissimilarity and the rest is shown in blue.



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

(e) **(f)** **(g)** **(h)**

*Figure 4.27. T1-w images and the corresponding skullstripping boundaries generated by MONSTR (red) and the U-net (yellow) and their overlap (white). Figures **(a)**-**(h)** show the 8 subjects from the development set that have the lowest Dice similarity in ascending order.*

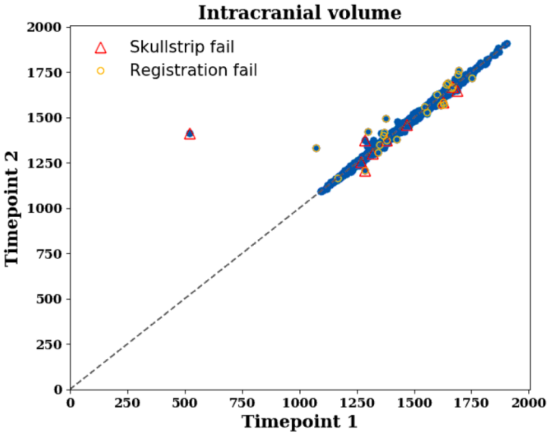


Figure 4.28. The ICV predictions from the skullstripping U-net for MRIs at the first and second visit (timepoint 2 vs. timepoint 1) shown in blue. The line representing equal volume is shown with a dashed black line.

4.5 Conclusions

We have shown that training a CNN using the brainmasks generated by MONSTR can lead to better results than MONSTR itself, if images with errors are removed from the training set. This training method can be used when a skullstripping method such as MONSTR generates near-perfect brainmasks for a large subset of subjects which can be used for training. We show visually that the 8 brainmasks with the largest Dice dissimilarity between the MONSTR brainmasks and the Skullstripping U-net brainmasks are due to errors in MONSTR and not the skullstripping U-net. Of those 8 brainmasks, the one with the smallest error is small enough that we conclude that comparing the results for more subjects is not needed. We manually inspected 9 slices from 2401 subjects, which came back for another visit 5 years later on average, and found that the U-net was robust except in cases with registration errors. These registration errors were marked on Figure 4.28 comparing the brainmask volumes at two timepoints, from the skullstripping U-net. The figure shows that the predicted ICV is consistent except in the few cases which had already been marked with registration and skullstripping failures after visual inspection.

5 A joint ventricle and WMH segmentation from MRI for evaluation of age-related changes in the brain

5.1 Introduction

Age-related changes in brain structure include atrophy of the brain parenchyma and white matter changes of presumed vascular origin. Enlargement of the ventricles may occur due to atrophy or impaired cerebrospinal fluid (CSF) circulation [13]. The co-occurrence of these changes in neurodegenerative diseases and in aging brains often requires investigators to take both into account when studying the brain, however, automated segmentation of enlarged ventricles and WMHs can be a challenging task. Here we present a hybrid multi-atlas segmentation and convolutional autoencoder approach for joint ventricle parcellation and WMH segmentation from MRIs. Our fully automated approach uses the convolutional autoencoder SegAE to generate a standardized image of grey matter, white matter, CSF, and WMHs, which, in conjunction with labels generated by a multi-atlas segmentation approach, is then fed into a CNN to parcellate the ventricular system. Hence, our approach does not depend on manually delineated training data for new data sets. The segmentation pipeline was validated on both healthy elderly subjects and subjects with NPH using ground truth manual labels and compared with state-of-the-art segmentation methods. We then applied the method to a cohort of 2401 elderly brains to investigate associations of ventricle volume and WMH load with various demographics and clinical biomarkers. Our results indicate that the ventricle volume and WMH load are both highly variable in a cohort of elderly subjects and there is an independent association between the two, which highlights the importance of taking both the possibility of enlarged ventricles and WMHs into account when studying the aging brain.

5.2 Materials

For developmental purposes we selected 90 subjects (age 67-92) from the AGES cohort. As in Chapter 3, the subjects were selected based on previously reported total ventricle volumes [54] to roughly group subjects into three groups of 30: Group 1 containing the smallest, Group 2 the medium, and Group 3 the largest ventricle sizes. This way our development sample covered the entire spectrum of ventricle sizes of the AGES cohort (smallest to largest). Out of the development set of 90 subjects, 60 subjects were used for training, 5 for validation of model parameters, and the remaining 25 were used for testing.

A second data set, consisting of NPH patients, from the Johns Hopkins Hospital, Baltimore, USA (see details in Appendix A.2) was used to test the robustness of the proposed method to a different scanner type and subject population. The selection of the SegAE training subjects from the NPH data set was performed by rating the severity of WMHs in the 80 NPH subjects on a scale from 0 to 3 and randomly selecting 10 images with a severity of 3. To determine the hyperparameters for SegAE, 3 subjects were randomly selected for validation. The other 77 subjects were used for testing the quality of the ventricle segmentation. Out of the subjects with a WMH score of more than 0, 10 subjects were randomly selected for testing the quality of the WMH segmentation.

5.2.1 Preprocessing

The images in the AGES-Reykjavik data set and the NPH data set were pre-processed by resampling to $0.8 \times 0.8 \times 0.8$ mm^3 voxel size, rigidly registering the baseline T1-w images to the MNI-152 atlas space [46] and, in turn, registering the baseline T2-w, FLAIR and follow-up images to the corresponding baseline T1-w images in the MNI-152 atlas space. All images were skullstripped using the skullstripping U-net described in Chapter 4. Since inhomogeneity correction is a part of the training process for SegAE (as described in [60] and in Chapter 2, Section 2.5), no inhomogeneity correction was needed during pre-processing.

5.3 Methods

Two different CNN architectures were used for two sequential tasks in the pipeline; the segmentation autoencoder (SegAE) for unsupervised tissue and WMH segmentation and a U-net specifically designed for parcellating the ventricular system, hereafter referred to as the Ventricle CNN or V_CNN, using standardized images made from the SegAE segmentations as input. Figure 5.29 shows the complete segmentation pipeline.

5.3.1 Generation of training data and CNN architecture

SegAE was used to segment each subject's brain into the GM, WM, CSF, and WMHs from the skullstripped T1-w, T2-w, and FLAIR images, as described in Chapter 2. The resulting CSF segmentations sometimes contained unwanted signal decay due to pulsation artifacts, which appeared bright as WMHs within the ventricles in FLAIR images (see the 3rd ventricle in Figure 5.30(a)). This sometimes resulted in the pulsation artifact being classified as WMHs. We corrected for these artifacts using a pulsation artifact segmentation obtained with an element-wise multiplication of the CSF and WMH segmentations from SegAE. The pulsation artifact segmentation was then added to the CSF segmentation and subtracted from the WMH segmentation. Results from this correction procedure are shown in Figure 5.30.

The RUDOLPH [52] algorithm was run on the development set of 90 subjects

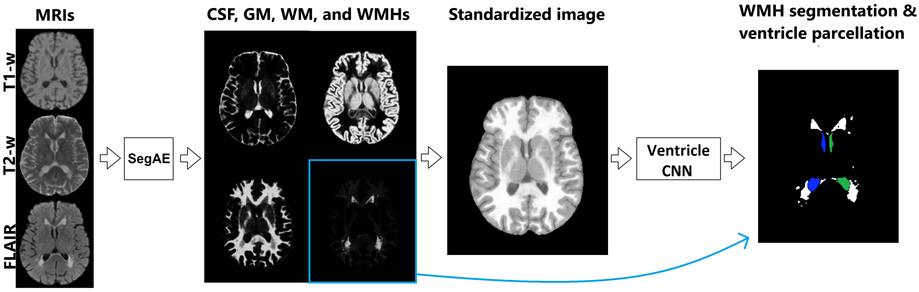


Figure 5.29. The proposed pipeline for joint ventricle and WMH segmentation. SegAE is used to decompose T1-w, T2-w and FLAIR images into four images, where the proportion of CSF, GM, WM, and WMHs is represented in each voxel. These are in turn used to create a standardized image from which the Ventricle CNN parcellates the ventricular system into the left and right lateral ventricles, and the 3rd and 4th ventricles.

from the AGES-Reykjavik data set and the ventricle labels for the left and right lateral ventricles and the 3rd and 4th ventricles were isolated. The RUDOLPH ventricle labels corresponding to subjects in the training set were manually inspected and in 25 images, lateral ventricle labels erroneously appearing in the sulcal CSF were manually removed. Subsequently, the CSF segmentation from SegAE was multiplied with the RUDOLPH ventricle labels to generate parcellated ventricle training labels that were consistent with the tissue segmentation from SegAE. This improved the quality of the training labels due to RUDOLPH's consistent over-segmentation of the ventricles. The new ventricle labels were further processed with morphological closing to fill holes in the segmentation of the ventricles due to the brighter choroid plexus within the ventricles (see Figure 5.31). Morphological closing was performed with a 3×3×3 cube for two iterations in the lateral and third ventricles.

The input into the ventricle segmentation network was a weighted combination of the soft segmentation outputs from SegAE

$$I_{standard} = 1 \cdot S_{CSF} + 2 \cdot S_{GM} + 3 \cdot (S_{WMH} + S_{WM})$$

where $I_{standard}$ is the standardized image and S_{CSF}, S_{GM}, S_{WMH} and S_{WM} are the soft segmentations of the CSF, GM, WMHs, and WM, respectively. Scalar multiplication $(\cdot)$ with the weights 1, 2, and 3 is used to distinguish the tissues when they are combined into one image. Doing this allows us to use a single homogeneous image with a sharp tissue contrast as input (see Figure 5.32 (c)), fit larger patches into GPU memory than if all the SegAE segmentations or MRI sequences were used as input, and to standardize tissue contrast when different sequences or MRIs from different scanners are used. The CNN architecture for the combined SegAE and V_CNN pipeline can be seen in Figure 5.33.

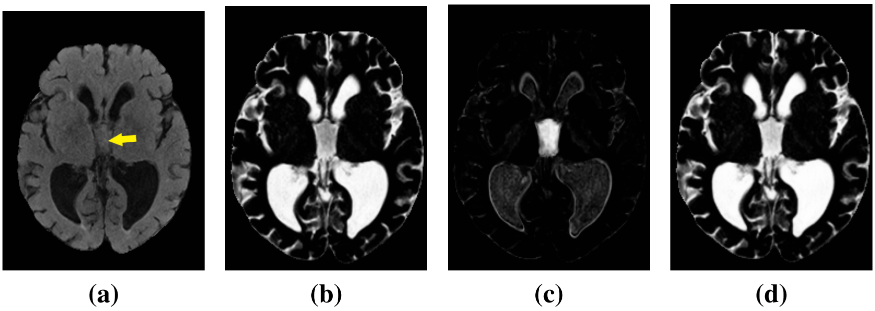


Figure 5.30. Identification and removal of pulsation artifact. Image ***(a)*** *shows a FLAIR image with a pulsation artifact in the third ventricle (yellow arrow). Image* ***(b)*** *shows the corresponding CSF output from SegAE before thresholding. Image* ***(c)*** *shows a pulsation artifact segmentation obtained with element-wise multiplication of the CSF and WMH outputs (non-binarized). Image* ***(d)*** *shows the CSF segmentation that has been corrected for pulsation artifacts by adding the pulsation artifact segmentation shown in* ***(c)****.*

5.3.2 Training and prediction

AGES-Reykjavik data set
The SegAE network was trained using T1-w, T2-w and FLAIR images from 30 subjects, as described in Chapter 2, Section 2.7. For the evaluation of input sequence dependence in Section 5.4.2, two other trained SegAE networks were prepared using the same method: One SegAE network was trained using only T1-w images as input, and another using only T1-w and T2-w images as input to the network. However, in all three SegAE networks, the same three inhomogeneity corrected T1-w, T2-w and FLAIR sequences were used for calculation of the loss function during training. While this training scheme requires all the same MRI sequences as before for the training set, the omitted input sequences are not needed for prediction for subjects that are not in the training set.

Standardized images were created from the SegAE segmentations of the AGES-Reykjavik training set and 128×128×128 voxel patches were extracted with a 40 voxel stride. The V_CNN was trained on standardized images and corresponding ventricle labels from 60 subjects. The SegAE network trained using T1-w and T2-w images as input did not show the strong pulsation artifacts that were apparent in the SegAE CSF segmentation when FLAIR images were included as input. Therefore, the SegAE network using only T1-w and T2-w images as input was used for post-processing of the RUDOLPH ventricle labels that were used for training the V_CNN. The V_CNN was trained using a Dice loss for 200 epochs (due to memory constraints, 15 subjects were selected 4 times to train for 50 epochs) with a learning rate of $1 \cdot 10^6$, using the Adam optimizer [103] with Nesterov momentum [104] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, schedule decay of 0.004, and a batch size of one. The learning rates were chosen manually with six tries and comparisons with the validation set of 5 subjects (labels generated in the same way as the training data in Section 5.3.1). Other hyperparameters were not changed from the default values of Tensorflow [106]. After training, ventricle label

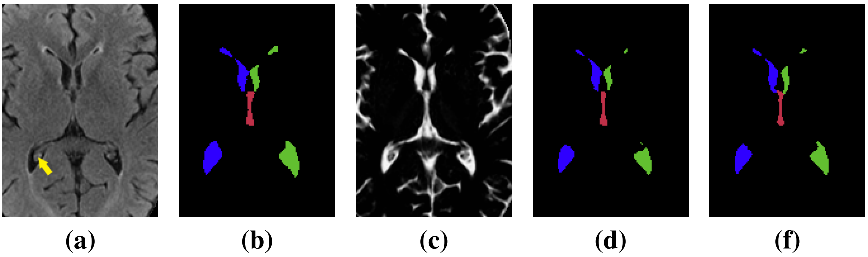


Figure 5.31. Preparation of training labels. Image ***(a)*** *shows a slice of a FLAIR image showing choroid plexus in the right lateral ventricle (yellow arrow). Image* ***(b)*** *shows the corresponding ventricle segmentation from RUDOLPH, and* ***(c)*** *shows the corresponding CSF segmentation from SegAE. Image* ***(d)*** *shows a ventricle segmentation obtained by element-wise multiplication of each label from RUDOLPH with the CSF segmentation in* ***(c)*** *and a morphological closing of the lateral and third ventricles.* ***(f)*** *shows a corresponding manual delineation.*

prediction was performed with a stride of 64, and patches were assembled using the average of overlapping voxels.

NPH data set
The SegAE network trained on the AGES-Reykjavik images was further trained using T1-w, T2-w and FLAIR images of the 10 training subjects in the NPH data set using a learning rate of 0.0001. The V_CNN that was trained on the AGES-Reykjavik data set was used directly on the NPH data set, with no retraining, and prediction was performed in the same way as for the AGES-Reykjavik data. An automatic post-processing of the ventricle parcellation was performed by changing sporadic lateral- and 3rd ventricle labels in the same connected component as the fourth ventricle to fourth ventricle label.

5.4 Evaluation

We present three experiments for the proposed joint ventricle and WMH segmentation method. First, we conduct a comparison with widely used, publicly available segmentation methods, using manual delineations as a reference. Second, we experiment with different combinations of input sequences into our segmentation pipeline. Finally, we compare the segmented volumes to various biomarkers in the AGES-Reykjavik data set and explore the strength of association between ventricle size and WMH load.

5.4.1 Visual and quantitative comparison

The four ventricle compartments and WMHs in a total of 25 subjects (8-9 from each Group of different ventricle sizes described in Section 5.2) from the AGES-Reykjavik cohort were manually delineated for evaluation of the proposed method. In addition, the method was evaluated on 77 subjects with manual ventricle labels and 10 subjects

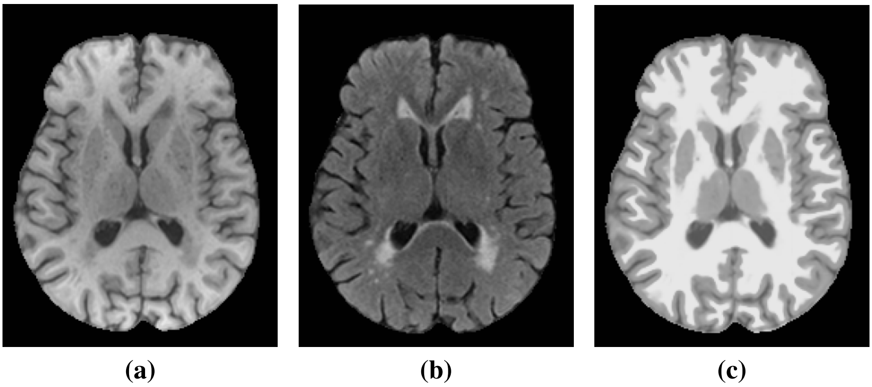


*Figure 5.32. Images **(a)** and **(b)** show T1-w and a FLAIR images of a subject, respectively, and image **(c)** shows a standardized image made of SegAE segmentations, which is free of inhomogeneity artifacts and WMHs.*

with manual WMH labels from the NPH data set.

For both of these data sets, the entire ventricular system was labeled first as a single binary mask from the T1-w image. Then each ventricle mask was parcellated into the left and right lateral ventricles, and the 3rd and 4th ventricles. The WMHs were manually segmented from the FLAIR images. All test subjects were processed using the proposed method, as well as two whole brain segmentation methods: The widely used FreeSurfer 6.0 [51] and RUDOLPH, which was specifically developed to be robust to severely enlarged ventricles. Furthermore, the method was compared to two publicly available and widely used WMH segmentation methods: LGA [109] and LPA [110]. These four methods could be applied to our data sets without the need for a new set of manually delineated training segmentations. We ran FreeSurfer both with default parameters and with the `-bigventricles` flag to account for enlarged ventricles.

For each subject in the test set the following metrics were computed to evaluate the performance of the proposed method and alternative methods compared to the manually delineated structures:

- *Dice Similarity Coefficient (DSC)*
  A measure of overlap between the ground truth and predicted segmentations, DSC is defined as $2\frac{|\mathbf{A}\cap\mathbf{B}|}{|\mathbf{A}|+|\mathbf{B}|}$, where **A** and **B** are binary masks. A DSC of 1 indicates a perfect overlap and 0 indicates no overlap between **A** and **B**.

- *Log Volume Ratio (LVR)*
  A log transformed ratio of the predicted volume V_P to the true volume V_T. LVR is defined as $log(\frac{V_P}{V_T})$. Lower LVR indicates a more accurate prediction.

- *Lesion-wise F1-score (L-F1)*
  Let N_P be the number of correctly detected lesions after comparing the predicted lesion mask P to the ground truth lesion mask 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.

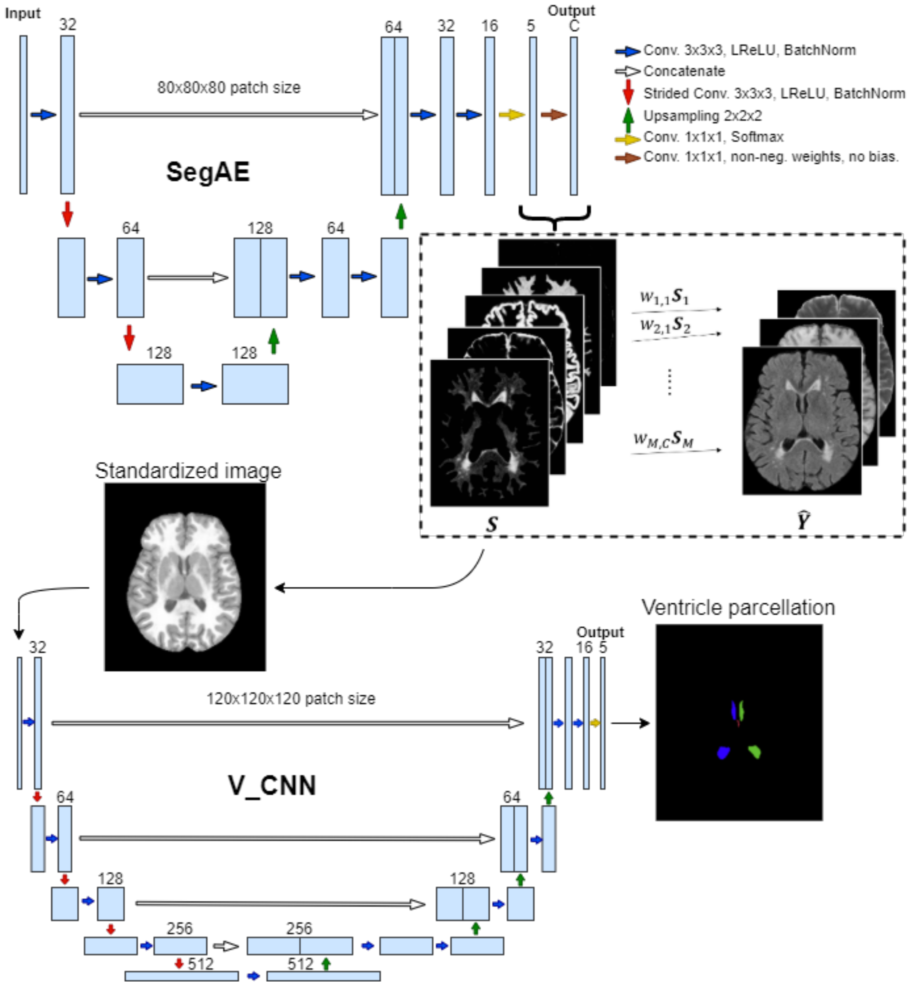


Figure 5.33. The proposed CNN architecture. The input comprises large 3D patches of MRI images that are segmented in an unsupervised manner using SegAE into WMHs, WM, GM, CSF, and meninges. The segmentations are used to create standardized images, which in turn are used as the input into the V_CNN. Kernels of size 3×3×3 are used in all convolutional layers except size 1×1×1 is used in the final two layers of both SegAE and the V_CNN. The V_CNN output is a segmentation of the four ventricle compartments, which in conjunction with the SegAE output provides a consistent ventricle and WMH segmentation.

- *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:

$$d_{\mathrm{H}}(X,Y) = \max\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\},$$

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.

Figure 5.34 shows the ventricle volumes (of the entire ventricular system combined) of the manual masks and the estimated ventricle volumes using the three methods, as well as the DSC between the corresponding ventricle segmentations and the manual masks, ordered by the volume of the manual masks, for the AGES-Reykjavik and the NPH test sets, respectively. This way, the performance of the methods relative to ventricle volume is demonstrated, since the DSC is known to be sensitive to the size of the segmented volume [121]. The proposed method shows the most stable performance on all three groups of ventricle sizes in the AGES-Reykjavik data set, and achieves the highest DSC on 20 out of the 25 subjects. RUDOLPH has the lowest DSC score on the smallest ventricles. FreeSurfer with default settings fails when presented with the largest ventricles. In subsequent analysis we omit results from default FreeSurfer, since FreeSurfer with the `-bigventricles` flag has a similar performance as the default version on subjects with smaller ventricles. The proposed method shows a stable performance on all 77 subjects with NPH compared to FreeSurfer, which shows poor performance on two subjects with severely enlarged ventricles, and RUDOLPH, which has a consistently lower DSC score.

Table 5.7 shows the DSC, LVR, and H95 metrics for the entire ventricular system and each sub-compartment. Table 5.8 shows the DSC, LVR and L-F1 metrics for the WMHs. Scores are averaged over all subjects and the best scores are shown in bold. Statistical significance was determined using a Wilcoxon signed-rank test and values that are significantly different ($p < 0.05/15$) from the proposed method are denoted with an asterisk (*). The proposed method achieves the highest average DSC and H95 scores on the entire ventricular system (significantly better than FreeSurfer and RUDOLPH on both the AGES-Reykjavik and NPH data sets). On the NPH data set, the proposed method also achieves the highest LVR score (significantly better than FreeSurfer and RUDOLPH), however, FreeSurfer achieves a slightly better LVR score on average than the proposed method on the AGES-Reykjavik data set, although it is not significantly different. The corresponding scores for the left- and right lateral ventricles and the 3rd and 4th ventricles separately are also shown in Table 5.7. The proposed method achieves the highest average WMH segmentation score on all three metrics compared to FreeSurfer, LGA, and LPA on the AGES-Reykjavik data set. LPA achieves the highest DSC and L-F1 scores on the NPH data sets, although, was not significantly better than the proposed method. Our method achieved the best LVR score, which was not significantly different from the comparison methods.

A visual comparison of the ventricle and WMH segmentations from the proposed method and the alternative segmentation methods can be seen in Figure 5.35. LPA and LGA provide WMH labels but not ventricle labels. RUDOLPH and FreeSurfer provide a whole brain segmentation with ventricle parcellation, however RUDOLPH does not provide WMH labels, as is common in multi-atlas segmentation approaches, and the WMH labels from FreeSurfer are not accurate, as expected, given that FreeSurfer's segmentation is entirely based on the T1-w sequence, where WMHs have similar intensity values to GM structures. The proposed method is able to provide accurate

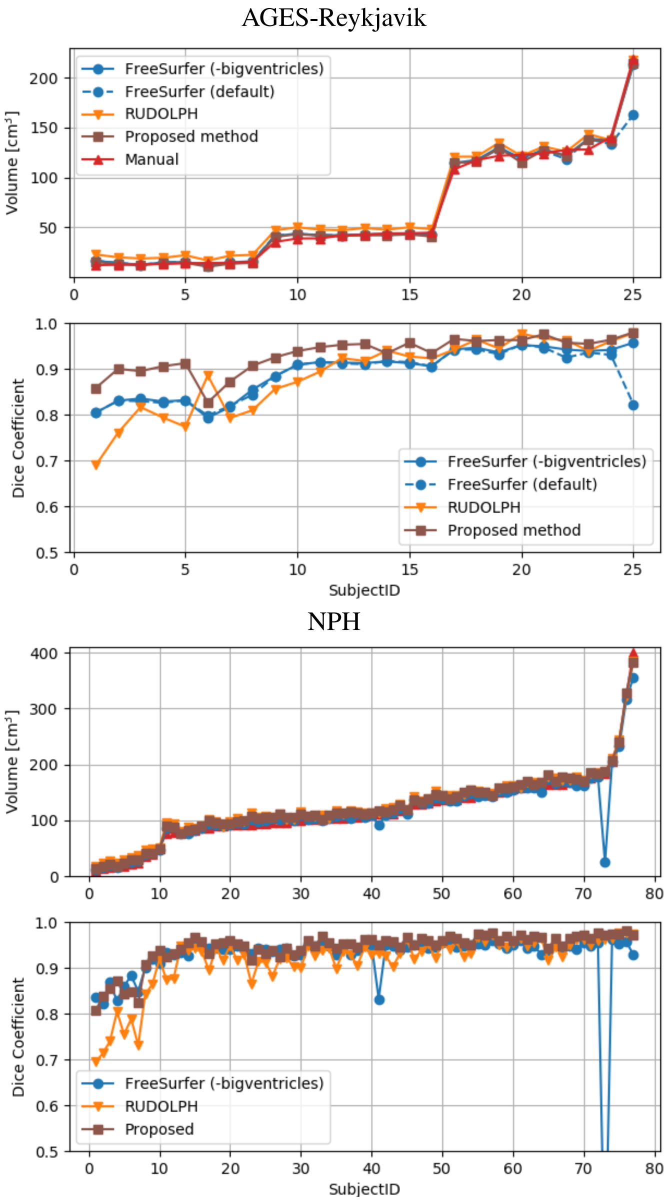


Figure 5.34. Volume and DSC comparison for the entire ventricle system in subjects from the AGES-Reykjavik data set (above) and NPH data set (below). The top graph show the overall ventricle volume for the manual masks (red) and masks generated by FreeSurfer (blue), RUDOLPH (orange), and the proposed method (brown), ordered by the volume of the manual masks. The bottom graphs show the DSC for the same methods compared with the manual masks.

Table 5.7. Evaluation of the ventricle segmentation. The mean and standard deviation of the DSC, LVR, and H95 for FreeSurfer, RUDOLPH and the proposed CNN pipeline on the entire ventricular system (Entire), the left lateral ventricle (LLV), the right lateral ventricle (RLV), the third ventricle (3rd) and the fourth ventricle (4th). A paired Wilcoxon signed-rank test was used to obtain the p-values for determining statistical significance. Asterisk (*) denotes values that are significantly different from the proposed CNN ($p < 0.05/15$), and bold figures denote the best score for each metric.

		AGES-Reykjavik data set (N = 25)		
		FreeSurfer	RUDOLPH	Proposed
Entire	DSC	0.894 (± 0.048)*	0.888 (± 0.079)*	**0.932 (± 0.038)**
	LVR	**0.071 (± 0.072)**	0.200 (± 0.171)*	0.072 (± 0.068)
	H95	6.939 (± 7.229)*	6.624 (± 8.384)*	**2.816 (± 5.408)**
LLV	DSC	0.906 (± 0.044)*	0.889 (± 0.081)*	**0.938 (± 0.039)**
	LVR	0.072 (± 0.077)	0.203 (± 0.175)*	**0.070 (± 0.077)**
	H95	7.155 (± 8.370)*	7.451 (± 0.175)*	**2.942 (± 5.459)**
RLV	DSC	0.900 (± 0.052)*	0.890 (± 0.081)*	**0.935 (± 0.038)**
	LVR	**0.073 (± 0.070)**	0.195 (± 0.179)*	0.078 (± 0.074)
	H95	7.646 (± 0.916)*	7.205 (± 9.222)*	**3.848 (± 7.750)**
3rd	DSC	0.853 (± 0.044)	0.867 (± 0.056)	**0.869 (± 0.039)**
	LVR	**0.136 (± 0.098)**	0.188 (± 0.132)*	0.152 (± 0.119)
	H95	**2.260 (± 0.781)**	2.540 (± 0.993)	2.573 (± 0.910)
4th	DSC	0.687 (± 0.077)*	0.777 (± 0.092)*	**0.824 (± 0.054)**
	LVR	0.525 (± 0.191)*	0.417 (± 0.224)*	**0.199 (± 0.144)**
	H95	12.857 (± 3.231)*	2.901 (± 1.434)	**2.615 (± 1.473)**
		NPH data set (N = 77)		
Entire	DSC	0.923 (± 0.088)*	0.916 (± 0.060)*	**0.944 (± 0.036)**
	LVR	0.076 (± 0.227)*	0.110 (± 0.130)*	**0.074 (± 0.064)**
	H95	3.884 (± 5.788)*	17.564 (± 7.175)*	**2.562 (± 2.303)**
LLV	DSC	0.928 (± 0.084)*	0.921 (± 0.062)*	**0.945 (± 0.036)**
	LVR	**0.073 (± 0.214)***	0.104 (± 0.134)	0.083 (± 0.068)
	H95	3.540 (± 5.493)*	17.820 (± 7.419)*	**2.180 (± 1.709)**
RLV	DSC	0.925 (± 0.097)*	0.915 (± 0.059)*	**0.946 (± 0.034)**
	LVR	0.079 (± 0.276)*	0.108 (± 0.126)*	**0.073 (± 0.038)**
	H95	**3.883 (± 6.404)***	20.394 (± 7.936)*	3.999 (± 7.977)
3rd	DSC	0.830 (± 0.076)	**0.851 (± 0.095)**	0.837 (± 0.104)
	LVR	0.155 (± 0.169)*	0.234 (± 0.215)	**0.219 (± 0.238)**
	H95	3.512 (± 1.852)	**2.642 (± 1.321)***	4.192 (± 5.308)
4th	DSC	0.739 (± 0.078)*	**0.805 (± 0.070)**	0.775 (± 0.127)
	LVR	0.417 (± 0.179)*	**0.309 (± 0.189)**	0.315 (± 0.364)
	H95	9.771 (± 3.815)	**2.885 (± 1.493)***	8.583 (± 7.499)

and consistent (i.e., non-overlapping labels) WMH and ventricle segmentation, while FreeSurfer's WMH labels tend to bleed into the labels of the lateral ventricles.

Table 5.8. Evaluation of the WMH segmentation. The mean and standard deviation for DSC, LVR, and L-F1 for the WMH segmentations from FreeSurfer, LGA, LPA, and SegAE. A paired Wilcoxon signed-rank test was used to obtain the p-values for determining statistical significance. Asterisk (*) denotes values that are significantly different from the proposed CNN ($p < 0.05/15$), and bold figures denote the best score for each metric.

AGES-Reykjavik data set (N = 25)

	FreeSurfer	LGA	LPA	Proposed
DSC	0.284 (± 0.107)*	0.634 (± 0.146)*	0.669 (± 0.175)	**0.774** (± **0.100**)
LVR	0.697 (± 0.255)*	0.322 (± 0.352)	0.558 (± 0.607)*	**0.297** (± **0.307**)
L-F1	0.127 (± 0.068)*	0.309 (± 0.117)*	0.354 (± 0.185)	**0.437** (± **0.085**)

NPH data set (N = 10)

	FreeSurfer	LGA	LPA	Proposed
DSC	0.482 (± 0.0.120)*	0.665 (± 0.110)	**0.778** (± **0.053**)	0.721 (± 0.070)
LVR	0.754 (± 0.288)	0.634 (± 0.233)	0.360 (± 0.156)	**0.334** (± **0.173**)
L-F1	0.088 (± 0.037)	0.086 (± 0.020)	**0.163** (± **0.057**)	0.146 (± 0.070)

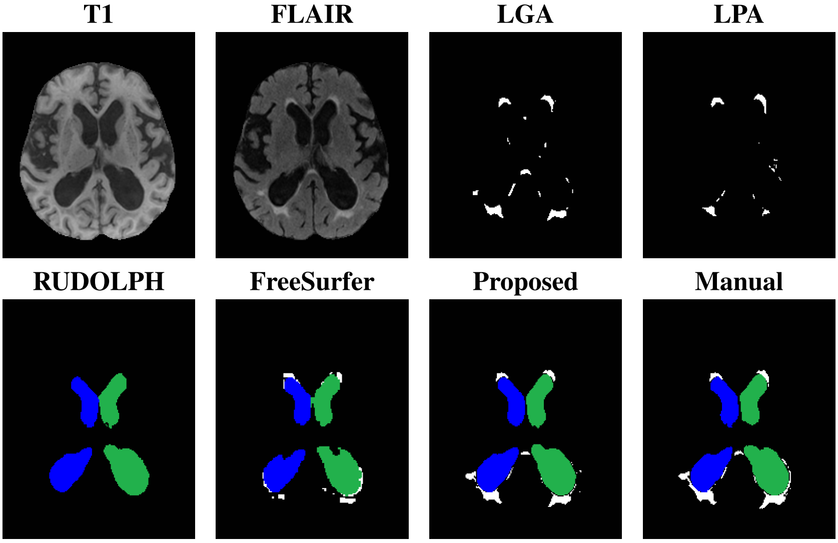


Figure 5.35. Visual comparison of the output of the proposed method and the five methods used for comparison showing the left (green) and right (blue) lateral ventricles (the 3rd and 4th ventricles are not visible in these slices), and WMHs (white). LPA and LGA provide WMH labels but not ventricle labels. RUDOLPH and FreeSurfer provide a whole brain segmentation with ventricle labels, however RUDOLPH does not provide WMH labels and the WMH labels from FreeSurfer are not accurate. The proposed method provides accurate ventricle and WMH labels.

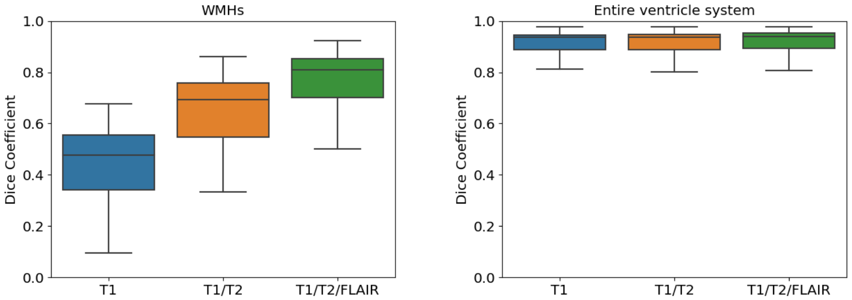


Figure 5.36. The boxplots show the DSC of the proposed method using the different number of sequences as input for the WMHs (left) and ventricular system (right): 1) Only T1-w (blue), 2) only T1-w and T2-w (orange), and 3) T1-w, T2-w, and FLAIR images (green).

5.4.2 Input sequence dependence

In our second experiment we wanted to explore whether the proposed method was able to segment the WMHs and the ventricles using fewer input sequences than the network was trained on. This would be beneficial, for example, for WMH segmentation when FLAIR images are missing. Three SegAE networks trained on the AGES-Reykjavik data set were used for this experiment, one using only the T1-w image as input, another using T1-w and T2-w images as input, and one using T1-w and T2-w and FLAIR images as input. Only one V_CNN network was used to segment the ventricles in standardized images created with segmentations from the three SegAE networks separately. The AGES-Reykjavik test set was used for evaluation of the input sequence dependence. Figure 5.36 shows boxplots of the DSC coefficients for WMHs (Figure 5.36, left) and the entire ventricular system (Figure 5.36, right) for the proposed method when using the following sequences as input: 1) only T1-w images, 2) only T1-w and T2-w images, and 3) using T1-w, T2-w, and FLAIR images as input. As expected, the segmentation accuracy for WMHs is not as accurate when some of the sequences are missing, however, we note that our method is able to produce similar WMH segmentation results as the LGA method but without using the FLAIR image as input (mean DSC 0.647 ± 0.140 for the proposed method vs. 0.634 ± 0.146 for LPA). Likewise, the proposed method produced a better average DSC for the WMH segmentation than FreeSurfer when using only the T1-w image as input (mean DSC 0.442 ± 0.162 for the proposed method vs. 0.284 ± 0.107 for FreeSurfer). For the ventricle segmentation, there is no significant difference in accuracy when using one, two, or all three sequences.

5.4.3 Association between ventricle size and lesion load

In our final experiment, we explored associations between the segmentation volumes and various biomarkers in the AGES-Reykjavik study on data from 2371 subjects. We ran the pipeline on 2401 subjects for which all the required data existed and omitted 30 subjects due to failures in processing (24 due to registration errors and 6 due to

Table 5.9. Demographics and biomarkers [mean and standard deviation (SD)] at baseline for the 2371 subjects used in the multiple regression model. Age, sex, body mass index (BMI), systolic (sys) and diastolic (dia) blood pressure, hypertension medication (htnmed), diabetes mellitus type 2 (DM2), and history of smoking (Smoking).

	Baseline
Age [mean, SD]	74.662, 4.750
Sex [% male]	41%
BMI [mean, SD]	27.190, 4.086
sys [mean, SD]	141.294, 19.880
dia [mean, SD]	74.102, 9.347
htnmed [count]	1430
Type 2 diabetes [count]	214
Smoking status [count]	[0: 1026, 1: 1091, 2: 254]

skullstripping errors).

First, we explored the relationship of age and the total volume of the ventricles divided by ICV as well as the WMH load divided by ICV for men and women (see Figure 5.37). We show the 3 year average and standard deviation of the volumes for each age group. While there was an overall increase in both ventricle volume and WMH load with age, the individual variability was high within each age group.

Second, we compared selected segmentation volumes from our pipeline (ventricle- and sulcal CSF volumes, WMH load, and ICV) to several demographics and biomarkers in the AGES-Reykjavik study (see Table 5.9) to explore risk factors for either ventricle enlargement or high WMH load and the individual association between the two. Table 5.10 shows the results from two multiple linear regression models on data from 2371 subjects at first visit (baseline) to predict the volume of the entire ventricular system (Table 5.10, top) and the WMH load (Table 5.10, bottom), respectively. The ventricle model has the WMH load as input, and the WMH load model has the entire ventricle volume as input. Furthermore, both models include the sulcal CSF volume, ICV, age, sex, body mass index (BMI), systolic and diastolic blood pressure, use of hypertension medication, diabetes mellitus type 2, and history of smoking. The values were normalized by subtracting the mean and dividing by the standard deviation. The use of hypertension medication and the presence of diabetes mellitus type 2 is represented with the dichotomous variables 0 and 1, for absence and presence, respectively. Smoking status is represented with the categorical variables 0, 1, and 2 for non-, former-, and current smoker, respectively. Using the multiple regression model we found that WMHs, ICV, age, and diabetes mellitus type 2 are significantly associated with the ventricle volume ($p < 0.05/15$). Furthermore, we found that ventricle volume, age, sex, diastolic blood pressure and smoking are significantly associated with WMH load ($p < 0.05/15$). Thus, there may be other underlying reasons for the association between WMHs and ventricle volume than increasing age and the other biomarkers and demographic factors mentioned above.

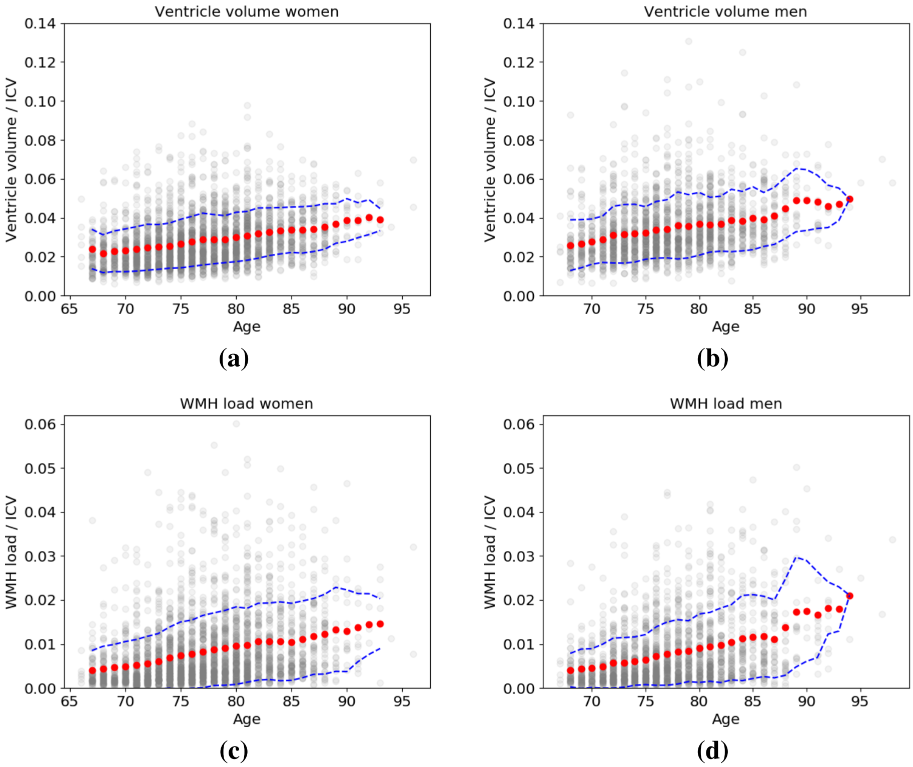


*Figure 5.37. The 3 year average (red dots) and standard deviation (dashed blue line) of ventricle volumes and WMH load for each age group. The association between age and the total volume of all the ventricles divided by ICV (top) for **(a)** women and **(b)** men, as well the the association between age and WMH load divided by ICV (bottom) for **(c)** women and **(d)** men. Individual subjects are shown in grey.*

5.5 Discussion

Our hybrid multi-atlas segmentation and convolutional autoencoder approach jointly provides a segmentation of WMHs and a parcellation of the ventricular system into its four main compartments. First, a segmentation of the WMHs, CSF, WM, and GM is aquired with the unsupervised CNN method SegAE. Then the training labels for the ventricle parcellation network, i.e., the V_CNN, are acquired without manual delineation by merging labels from the multi-atlas segmentation method RUDOLPH with the CSF segmentation provided by SegAE. The input to the V_CNN is a standardized image created with a linear combination of the SegAE segmentations. Our results imply that standardized images allow us to segment brain structures, such as the ventricles, using different types of sequences as input to the pipeline. In contrast, if a CNN was trained using the MRI sequences directly, different CNNs would have to be trained for each combination of input sequences [122] or by incorporating image synthesis [123]. Doing

Table 5.10. Multiple linear regression models to predict the entire ventricle volume (top) and WMH load (bottom). Input parameters are the WMH load and ventricle volume (Ventr.) for the ventricle volume and WMH load models, respectively, sulcal CSF (sCSF) volume, ICV, age, sex, body mass index (BMI), systolic (Sys) and diastolic (Did) blood pressure, hypertension medication (htnmed), Diabetes mellitus type 2 (DM2), and history of smoking (Smoking). The regression coefficients (β_n), the standard error (S), the t statistic and p-values, as well as the 95% confidence interval are reported in the table.

Predicting ventricle volume

	β_n	S	t	p	[0.025	0.975]
Constant	40.7473	0.355	114.688	0.000	40.051	41.444
WMH	4.1334	0.375	11.028	0.000	3.398	4.868
sCSF	0.5497	0.477	1.152	0.250	-0.386	1.486
ICV	6.9992	0.536	13.052	0.000	5.948	8.051
Age	3.3780	0.413	8.177	0.000	2.568	4.188
Sex	-1.3190	0.506	-2.609	0.009	-2.310	-0.328
BMI	0.6169	0.367	1.681	0.093	-0.103	1.337
Sys	0.3602	0.423	0.851	0.395	-0.470	1.190
Dia	-0.0727	0.429	-0.170	0.865	-0.913	0.768
Htnmed	-0.0216	0.369	-0.059	0.953	-0.746	0.703
DM2	1.1835	0.365	3.246	0.001	0.469	1.898
Smoking	0.3159	0.368	0.859	0.391	-0.406	1.037

Predicting WMH load

	β_n	S	t	p	[0.025	0.975]
Constant	9.3225	0.180	51.742	0.000	8.969	9.676
Ventr.	2.2806	0.207	11.028	0.000	1.875	2.686
sCSF	0.1301	0.242	0.537	0.591	-0.345	0.605
ICV	0.5573	0.281	1.981	0.048	0.006	1.109
Age	1.8497	0.209	8.850	0.000	1.440	2.260
Sex	0.9884	0.256	3.862	0.000	0.487	1.490
BMI	0.2958	0.186	1.589	0.112	-0.069	0.661
Sys	0.4248	0.215	1.980	0.048	0.004	0.845
Dia	0.7570	0.217	3.491	0.000	0.332	1.182
Htnmed	0.5315	0.187	2.843	0.005	0.165	0.898
DM2	0.0216	0.185	0.117	0.907	-0.342	0.385
Smoking	0.9274	0.186	4.996	0.000	0.563	1.291

that would limit the method to data sets with similar MRI parameters and scanner characteristics. This method can serve as an alternative to training on ground truth from multiple data sets at once or as an alternative to domain adaptation techniques for translating images between different domains [87].

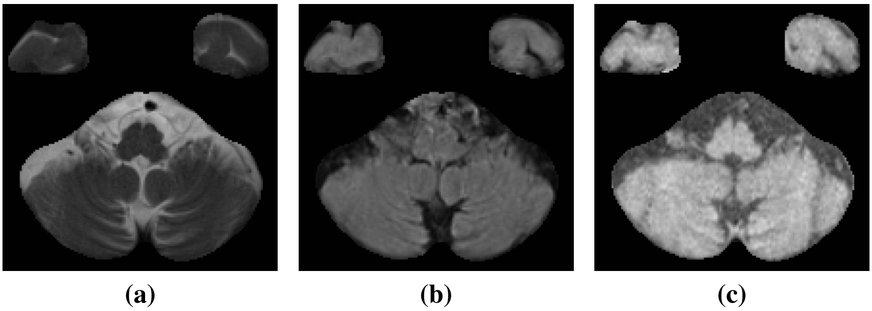


Figure 5.38. Images ***(a)****,* ***(b)****, and* ***(c)*** *show an axial slice of the cerebellum in T2-w, FLAIR, and T1-w images, respectively, of a subject in the NPH data set. The T2-w images have a higher in-plane resolution, which shows the thin lines of CSF in the cerebellum. Meanwhile, the upsampling of the lower resolution FLAIR and T1-w images gives them a blurry appearance, leading to brighter voxels instead of fine dark lines corresponding to the CSF in the T2-w image.*

Simply by training an unsupervised tissue and WMH segmentation method on the NPH data set, the V_CNN trained only on AGES-Reykjavik data could be used to further parcellate the ventricular system in standardized images of NPH subjects. This data set was especially challenging because of severely enlarged ventricles and, in many cases, strong pulsation artifacts in the FLAIR images. Our method is a step towards making CNNs, trained in a supervised manner using manually delineated labels or labels from multi-atlas segmentation methods, able to directly segment new brain MRI data (using different scanners or protocols) without the need to generate new training labels.

Pulsation artifacts were removed from the CSF and WMH segmentations from SegAE using a pulsation artifact mask obtained with element-wise multiplication of the soft CSF and WMH segmentation masks. A limitation of this approach is that if the pulsation artifact is too strong, such that it is exclusively present in the WMH segmentation, the multiplication will be zero. A pulsation artifact output could potentially be incorporated into SegAE to generalize the method further and avoid pulsation artifacts affecting the CSF and WMH segmentations.

One limitation that we came across when inspecting the SegAE WMH segmentations of the NPH data set was sporadic WMH labels erroneously appearing in the cerebellar region in some subjects of the NPH data set. We believe that this is due to resampling during pre-processing. Resampling may create a problem for unsupervised multi-contrast methods such as SegAE when there is a large difference in resolution between the available MRI sequences. For instance, when thin lines of CSF in high-resolution T2-w images of the NPH data set correspond to brighter voxels in FLAIR and T1-w images due to blurring (see Figure 5.38). One solution could be to use the lower FLAIR resolution as a reference for registration as proposed in [124]. However, the ground truth manual delineations that existed for our data set were only available in MNI-space. Future solutions may involve more advanced super-resolution techniques.

We conducted an ablation experiment to test if our method could be used to generate accurate segmentations without using all three MRI sequences as input (i.e., the T1-w, T2-w, and FLAIR images). Our method was shown to give a robust ventricle segmentation when changing the input MRI sequences used to generate the standardized images with SegAE (see Figure 5.36). Furthermore, we showed that the method could be used to segment WMHs without using the FLAIR images as input, i.e., by using only T1-w or only T1-w and T2-w images, although with some resignation in DSC. However, the DSC for the WMH segmentations when only T1-w and T2-w images were used as input were still comparable to the LPA method. Similarly, using only T1-w images as input, the average DSC was higher when using the proposed method rather than FreeSurfer. Therefore, this strategy may be a viable option in data sets where FLAIR images are not available for all subjects.

Finally, we conducted an experiment where we compared the ventricle and WMH segmentation volumes to various demographics and clinical biomarkers in the AGES-Reykjavik data set. The aim was to determine the variability in the elderly population and to explore the strength of association between ventricle sizes and WMHs and risk factors for both.

First we demonstrated in Figure 5.37 how the average ventricle sizes increase with age for both sexes and how a population data set could be used to determine enlarged ventricles using the standard deviation for each age group. Similarly, the average WMH load increases with age, and notably, the standard deviation also generally increases with age. Fewer data points between ages 90-97 cause the standard deviation to decrease. The results demonstrate the high variability of ventricle volumes and WMHs in the elderly population.

Ventricle volume and WMH load depend on multiple factors and to explore the individual association between the ventricle size and WMH load, we used multiple linear regression models that take several confounders into account. Previous studies have found hypertension to be a major risk factor for severe WMHs and that hyperintensive drugs reduce the risk of severe WMHs [125]. Our results showed a positive association with systolic and diastolic blood pressure, although only statistically significant for diastolic blood pressure, and a non-significant positive association with the use of hyperintensive drugs. The blood pressure variables were measured at the time of study and lack information about duration of high blood pressure over a longer time period for each subject. The use of hypertensive medication may indicate a longer history of high blood pressure and be positively associated with high WMH load even if they have a lowering effect on blood pressure. Diabetes mellitus type 2 has been associated with ventricle enlargement [126], as supported by our results, and a moderately elevated risk for lacunar infarction in older men [127]. Our results found a significant positive association with ventricle volume, but not with WMH load. Smoking has been associated with a higher WMH load [128], as seen in our results, however, we do not have an accurate measurement of how much or for how long each subject has smoked. Other lifestyle factors that are associated with smoking, such as alcohol consumption and/or less physical activity [129], may influence our results.

Previous studies have found associations between WMH load and region specific atrophy [39], which are both biomarkers of small vessel disease [27]. The WMH and CSF volumes increase with age while the GM and WM volumes decrease [31], and disproportionate ventricular dilation is associated with WMH load [40]. The association of WMH load and ventricular volume has also been shown to be independent of demographics, vascular burden and APOE genotype [41]. In our results, the ventricle volume but not the sulcal CSF volume is associated with WMH load. There could be different causes for this in different individuals. WMHs possibly indicate reduced white matter integrity around the ventricles in some individuals, and perhaps the expansion of the ventricles might cause periventricular WMHs in the case of NPH patients [130]. Our results indicate that it is important to investigate the ventricles in the elderly, diabetes patients, and in people with small vessel disease; and WMHs in elderly people with high blood pressure, enlarged ventricles, and smokers.

The significant positive association between ventricle volume and WMH load indicate that there are other underlying reasons for this association (such as cerebral small vessel disease) than the variables used in the multiple linear regression models in Table 5.10. Simultaneous segmentation of both the ventricles and WMHs in large scale studies of the elderly population may shed further light on this connection and differentiate between causes of ventricle enlargement and increased WMH load [41], and how they contribute to dementia [131].

The proposed method currently provides segmentation of WMHs and a detailed parcellation of the ventricular system, which has been a challenging task in the segmentation of brain MRIs of the elderly and people with neurodegenerative diseases [62,132]. The method has the potential of being extended to include segmentations of other brain structures, both cortical and subcortical, that are usually segmented with methods that do not take WMHs or other tissue abnormalities into account (e.g., multi-atlas segmentation methods or supervised CNNs with labelled training data). The proposed method also enables us to segment directly from standardized images that can be created using different MRI protocols and scanners if appropriate measures are taken to correct for image artifacts.

5.6 Conclusions

We have introduced a hybrid multi-atlas segmentation and convolutional autoencoder approach for a joint segmentation of WMHs and the four ventricular compartments in the human brain. The method was compared with the whole-brain segmentation methods FreeSurfer and RUDOLPH and the WMH segmentation methods LGA, and LPA. The proposed method achieved the best average DSC on the entire ventricular system in the AGES-Reykjavik cohort and in the NPH patient data set (comparison of individual ventricle structures and alternative metrics can be seen in Table 5.7). The proposed method achieved the highest average DSC, LVR and L-F1 for the WMH segmentation on the AGES-Reykjavik data set (see statistical significance in Table 5.8). LPA achieved the highest DSC and L-F1 for WMHs in the NPH data set (not significantly better than the proposed method), and the proposed method the best LVR (not significantly better than the comparison methods). We showed that WMH load and the ventricle volumes in the AGES-Reykjavik cohort are independently associated using a multiple linear regression model taking several potential confounders into account.

6 Discussion and conclusion

This dissertation presents novel deep neural network methods for brain segmentation. The methods are specifically designed to not require new sets of manually delineated training data when presented with new data sets. Our first approach, the Segmentation Auto-Encoder, SegAE, is the first CNN that learns to simultaneously segment tissues and WMHs in an unsupervised manner. The method enables researchers to use unsupervised learning in a way that the output components represent the same proportions of tissue and WMHs across all the subjects in the data set - i.e., subjects with low and high lesion load - instead of using conventional clustering methods that may output different proportions of tissue in each component for different subjects, such as when no component corresponds to WMHs in subjects with no WMHs; or instead of using supervised methods that require manual delineations and may not be robust to differences from the training set when it comes to MRI protocol or large variations in subject anatomy. Furthermore, the SegAE segmentations can be used to create standardized images, which in turn make supervised CNNs robust to differences in MRI parameters if they are trained to use the standardized images as input, e.g. for parcellation of brain structures. Our second approach is a CNN that segments and parcellates brain structures using available multi-atlas segmentation methods to generate data for training the CNN, which subsequently performs the segmentation task at hand much faster than the multi-atlas segmentation methods used on the training data. Furthermore, the automatically generated training data can be improved by either manually removing erroneous segmentations, as we did before training the Skull-stripping U-net, or by correcting the segmentations with the SegAE tissue segmentation output in conjunction with other post-processing approaches, as we did before training the Ventricle CNN. These modifications yielded better brain masks and ventricle labels, respectively, for training, which in turn resulted in CNNs that performed better than the methods used to generate the original training data. By validating the methods on outliers and images of patients with severe ventricle enlargement, we showed that the CNNs learned features that could extrapolate well to subjects outside the training data distribution of brain anatomy. Hence, we developed a fully automatic and robust pipeline for brain segmentation without the need for new manual delineations. The methods were validated on challenging brain MRI data sets from multiple centers (Iceland, USA, the Netherlands, and Singapore) with high variability in both lesion load and ventricle volume, including healthy elderly subjects with high variability in lesion load and NPH patients with severe ventricle pathology. The new pipeline was used to segment brain MRIs of 2401 subjects in the AGES-Reykjavik cohort, which provided an opportunity for a further qualitative and quantitative evaluation and for exploring the associations of ventricle volume, WMH load, and various demographics and clinical biomarkers.

6.1 Limitations and future work

In Chapter 1, I stated 5 major challenges in automatic brain MRI segmentation that we address in this work. Here I discuss how these challenges where met and uncertainties that arise from methodological constraints. I include ideas on what may need to be done in the future to completely solve these problems.

***Challenge 1**: The lack of training data for CNN segmentation methods.*

The amount of training data needed for the development of a CNN depends on the complexity of problem being solved, and how similar the training set is to the test set. Solving complex problems requires a CNN with a large number of model parameters, such as number layers and number of filters in each layer. Usually images of at least a few dozen subjects are needed for CNNs to capture the inter-subject variability, even when training subjects are selected specifically to reflect this variation. As an example, the WMH challenge data set includes a training set of 60 manually delineated subjects, and VParNet [63] was trained using 40 manual delineations of the ventricular system in images from two data sets. The procurement of so many gold standard manual delineations to segment WMHs and the ventricular system of subjects in new data sets is both time-consuming and expensive.

In Chapters 4, 3, and 5 training data was acquired automatically using multi-atlas segmentation methods. Consequently, the training data includes erroneous segmentations made by the imperfect methods used to generate the training data. However, the generated training data can be modified automatically, or with a quick manual removal of failures, resulting in a trained CNN that is more accurate than the method used to generate the training data.

The major limitation of generating training data this way is that the CNNs are not trained on gold standard manual delineations. There is a tradeoff between competing on accuracy metrics with the state-of-the-art deep learning methods trained using manual labels, and generating labels efficiently using the power of existing segmentation methods. This could be mitigated by pre-training on a large training set generated by state-of-the art multi-atlas segmentation methods and fine-tuning the CNN using manual delineations in case any are available. Furthermore, data augmentation techniques may be used to synthesize more data from a limited set of manually delineations.

In Chapter 2 we developed SegAE, a completely unsupervised method for WMH and tissue segmentation in brain MRI. SegAE does not need any training segmentations; neither manually delineated nor automatically generated labels. Instead, the network is trained to reconstruct the input images themselves under constraints that result in the output of one layer of the network to represent the proportion of WMHs and tissues in each voxel. One limitation of this method is that different data sets may have different MRI artifacts that need to be corrected for. E.g. inhomogeneity artifacts are different between scanners, and pulsation artifacts depend on the MRI parameters and ventricle volume [59]. Meanwhile, supervised methods would inherently take these artifacts into account. State-of-the-art methods for WMH segmentation are supervised CNNs trained on manual delineations, however, there is a high intra- and inter-rater variability in the manual delineations and manually segmenting the tissue classes in the same way would be immensely difficult and impractical. SegAE solves both of these problems, and is a promising method to segment objects of interest in the brain that are visible with different MRI sequences. This could be highly beneficial if in the future new and improved MRI sequences will emerge and previously generated manual delineations become absolute.

Another limitation of SegAE is that the brain MRIs in the training set need to have sufficient WMH contrast and amount of WMHs, which has yet to be adequately defined. Supervised segmentation methods also need sufficient amount of labelled voxels of a certain class for training, however, in the case of labelled training data it is possible to quantify the size of labelled structures and use class sampling or class weighting of the cost function to mitigate class imbalance.

The number of parameters used in the proposed models were 1,388,128 and 23,514,005 for SegAE and the Ventricle CNN, respectively. SegAE was trained using images from 30 subjects and the Ventricle CNN with images from 60 subjects. Training with more images may be beneficial, especially if biases in the training data that result from the training data generation process are minimized. Acquiring a larger training set for these methods depends on the availability of images, the memory limitations of the computer hardware, and the time needed to run RUDOLPH on all subjects in the training set.

***Challenge 2**: The segmentation failures of conventional multi-atlas registration based methods due to high variance in brain structures and abnormalities.*

MAS methods can fail due to variability between the atlases and target images. High variability is present in cases such as when images of healthy subjects and NPH patients are registered together, in cases of unusual extra-cerebral formations (before the application of a skull-stripping method), and in subjects with lesions.

In Chapter 4, we trained a Skullstripping U-net using brainmasks generated by MONSTR after removing brainmasks with visible failures from the training set. A visual inspection of the brain segmentations with the highest Dice dissimilarity between the output of the Skullstripping U-net and MONSTR showed that the Skull-stripping U-net was more robust than MONSTR on our development set. This indicates that a swift and effective way to improve brain segmentation, instead of relying on time-consuming and tedious manual delineations, is to generate training data for CNNs using conventional multi-atlas segmentation methods and manually removing segmentations from the training set with visible failures. Furthermore, these results suggest that the Skullstripping U-net learns features from the training data that generalize well to difficult subjects that are not in the training set.

One limitation of this work was the evaluation of this method; making a qualitative comparison of the segmentations with the highest Dice dissimilarity between two methods, instead of comparing the proposed Skullstripping U-net with ground truth segmentations. A low Dice dissimilarity would be measured if the skullstripping U-net and MONSTR made exactly the same errors. Furthermore, if the segmentations that were removed from the training set due to failures were caused by certain characteristics of the subjects, the trained CNN would perhaps not be robust to those characteristics because they would be out of distribution compared to the training data. However, our qualitative and quantitative analysis of MRIs of 2401 elderly subjects, with highly variable brain shapes and presence of abnormalities due to atrophy and neurodegenerative diseases, suggests that the Skullstripping U-net was accurate and robust.

In Chapter 3 and 5 we made CNNs for ventricle segmentation that were robust to severe ventricle enlargement by using the entire range of ventricle sizes in the AGES-Reykjavik data set for training. The training labels were automatically generated using the multi-atlas segmentation method RUDOLPH, which was specifically designed to segment NPH patients with extreme ventricle enlargement. The Ventricle CNN in Chapter 5 was further improved by removing oversegmentations caused by RUDOLPH from the training data by using the CSF segmentation from SegAE.

One limitation of this method is that the selection of training data, by selecting subjects that cover the entire range of ventricle sizes, utilized information that was known a priori. I.e., we used results from a method that was previously used to segment the AGES-Reykjavik data set [54], and although those results were not validated, it was enough to roughly group the training subjects into the smallest, medium, and largest ventricle sizes, without any overlap in ventricle volumes between groups. This procedure does not make the Ventricle CNN depend on external methods moving forward, because prior segmentations are not necessary for using the method on new data sets. One of the main aims in the development of the Ventricle CNN in Chapter 5 was to train the Ventricle CNN using standardized images. They can then be generated using different MRI scanners if appropriate measures are taken to correct for image artifacts. Doing so would allow us to use a Ventricle CNN trained on subjects with a range of ventricle sizes from one data set, even data sets including subjects with severe ventricle enlargement such as NPH patients, and implement the Ventricle CNN on new data sets with different scanner parameters and an unknown distribution of ventricle sizes.

***Challenge 3**: The long processing time of conventional methods.* The required processing time for RUDOLPH and FreeSurfer (with the `-bigventricles` flag) was 6 and 7.5 hours respectively. For the analysis of the 2401 subjects in the AGES-Reykjavik cohort, that came in for two visits, a 6 hour processing time for each visit would have resulted in over 3 years of processing time, unless further computational resources were employed.

The CNN method developed in Chapter 3 achieved a 360x speedup. Similarly, SegAE and the Ventricle CNN have a much shorter running time than RUDOLPH and FreeSurfer. Using RUDOLPH to generate 60 brain segmentations for training was the most time consuming part from training to implementation. However, the number of labels provided by FreeSurfer and RUDOLPH is much larger, and while the speed of prediction using a trained CNN is very fast, the training process can be slow. A CNN has been used to replicate FreeSurfer's anatomical segmentation of 95 classes in under 1 minute [133], which suggests that adding more labels would not slow down the speed of producing anatomical segmentation using CNNs considerably. The training time of CNNs may decrease in the future with improvements in hardware, software, and training methodology, which is outside the scope of this dissertation.

***Challenge 4**: Inconsistent segmentation results of images aquired with different MRI parameters or MRI scanners*

Automatic segmentation methods usually only work for a specific MRI sequence or a specific combination of sequences using relatively consistent MRI scanner parameters. This complicates analysis of data sets where different types of sequences are available, and data sets where sequences (e.g. T1-w) may be acquired with different parameters. Furthermore, future MRI sequences are likely to have improved image characteristics, such as more tissue contrast and higher resolution. Ideally, segmentation methods should provide consistent segmentation results across all these image variations, without the need to develop a new method for each case.

In Chapter 5, the WMHs and ventricles were segmented using our pipeline with different combinations of input sequences and data from different scanners. The tissue and WMH segmentation output from SegAE was combined into standardized images. A parcellation into anatomical structures can be then derived from standardized images from different sequences using a single CNN. This could be advantageous if there were missing sequences in the data set or if more sequences have been acquired for a subset of subjects. One potential limitation of using SegAE to generate the standardized images is that SegAE has not yet been validated for GM and WM segmentation using gold standard manual delineations, nor was the method specifically optimized for these tissues. Ground truth segmentations for WM and GM were lacking, and producing accurate manual delineations of these structures is difficult and time-consuming. A preliminary evaluation of the GM, WM and CSF segmentations from SegAE can be seen in Appendix B.6 using a single synthetic T1-w image from the BrainWEB database [134].

An intensity normalization is often performed as a preprocessing step before automatic MRI analysis due to standardize tissue intensity ranges [135]. As described in Chapter 2 the SegAE loss function uses Cosine proximity which is scale-invariant. However, the loss function is not bias-invariant. The Pearson correlation coefficient is both scale and bias-invariant due to a subtraction of the mean. However, the mean value of a brain image patch is not a good estimate of the total bias. A potential way to remove the bias is by using a filter with zero bias. The loss function used in Chapter 2 for training SegAE includes two terms; a Cosine proximity between the true images patches $\boldsymbol{Y}$ and the predicted image patches $\hat{\boldsymbol{Y}}$, and another term where high-pass filtering is performed on both $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ before applying the Cosine proximity function. Therefore, bias invariance could be achieved by simply removing the first term and selecting a high-pass filter $\boldsymbol{K}$ that optimizes the quality of tissue and WMH segmentation from SegAE. Another potential way of acquiring intensity range invariance with SegAE would be to apply intensity normalization to the training data.

Our method for creating standardized images does not take different resolutions of sequences and data sets into account. Images and datasets must be resampled to the same resolution before being processed by SegAE. This can create inconsistencies, such as when low resolution FLAIR images are upsampled to the resolution of high resolution T2-w images. Thin CSF tracts in the cerebellum can appear bright in the FLAIR images instead of being attenuated. In Chapter 5 this resulted in WMH oversegmentation in the cerebellum when the pipeline was run on the NPH data set. Potential solutions would be to use the lower FLAIR resolution as reference or incorporate super-resolution techniques into the pipeline, before training SegAE.

***Challenge 5**: Effective pre- and post-processing to adjust for MRI artifacts (such as inhomogeneity and pulsation artifacts)*
Preprocessing MRIs using an inhomogeneity correction method must be considered carefully in subjects with WMHs. One must make certain that WMHs are not being removed or degraded as the methods may consider WMHs as inhomogeneities [26] (see Figure 2.12 and Appendix B.2). In Chapter 2 we proposed alternating between tissue and WMH segmentation with SegAE and perforing N4 bias correction with pure-tissue probability masks excluding WMH regions. While this method was highly effective for inhomogeneity correction and preserving WMH contrast, the process convoluted the training process and requires manual analysis and intervention after each iteration. A simplification and automation of this process would be highly beneficial, potentially by using a CNN for inhomogeneity correction.

Pulsation artifacts and the choroid plexus can both cause holes in the segmentation mask of the ventricles. In Chapter 5 we proposed two solutions: 1) Creating a pulsation-artifact mask by multiplying the CSF segmentation and WMH segmentation from SegAE, and adding the pulsation artifacts back to the CSF segmentation; and 2) morphologically closing holes in the segmentation masks of the lateral and third ventricle. The former solution may not work if pulsation artifacts are strong enough such that they are almost exclusively apparent in the WMH segmentation, then the multiplication with the CSF segmentation would diminish the pulsation artifact segmentation. Furthermore, this method might alter the WMH and CSF segmentation in areas of lesions where both WMHs and CSF contribute to image intensities. The second solution, morphologically closing holes in the segmentation masks, did not have apparent downsides. A further validation of the method is warranted. It was solely intended to close small holes due to the choroid plexus and remnants of holes due to pulsation artifacts.

6.2 Overall conclusions

In this dissertation we have addressed the challenges of tissue and WMH segmentation, skullstripping, and segmentation of the ventricular system. New methods were developed and validated on large cohorts of the elderly and NPH patients with a high variability in WMH load and ventricle volumes. The five major technical contributions presented in this dissertation are: 1) the first unsupervised CNN for segmentation of brain tissues and WMHs; 2) methods that build on training data from MAS methods to improve accuracy and robustness to abnormalities; 3) two orders of magnitude faster processing speed compared the MAS methods; 4) a method for segmentation of images acquired with different MRI parameters and MRI scanners; and 5) methods for artifact removal, such as inhomogeneity correction in presence of WMHs and removal of pulsation artifacts. This work enables researchers to take a further look into the association of ventricle enlargement and WMHs, as well as to look at the effect of WMH location, ventricle shape, and how the WMHs and ventricles change over time in relation to their biomechanical causes and their effect on cognition. Finally, new large scale data sets with different MRI parameters may be incorporated into such a study without supervised retraining of the CNNs.

A Materials

A.1 AGES-Reykjavik MRI data

The AGES-Reykjavik study cohort comprises 5764 participants (female and male, age 66-93 at first visit), 4811 of which underwent brain MRI [54]. A total of 2644 out of the 4811 subjects had a second visit on average 5 years later. The MRIs were acquired using a dedicated 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 $0.94 \times 0.94 \times 1.5$ mm^3 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 mm^2; 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 $0.86 \times 0.86 \times 3.0$ mm^3 voxel size and 54 slices.

A.2 NPH MRI data

The NPH data set was acquired from the Johns Hopkins Hospital, Baltimore, USA. Brain MRIs of 80 NPH patients (age range 26-90 years with average age 66.8 ± 15) were acquired with a 3-Tesla scanner. MPRAGE sequence (TR: 2110 ms, TE: 3.24 ms, FA: 8°, TI: 1100 ms) with a 0.9 mm isotropic voxel size, axial T2-w sequence (TR: 6500 ms, TE: 134 ms, TA: 2:38) with a 3 mm slice thickness, and an axial FLAIR sequence (TR: 9000, TE: 94 ms, TI: 2500 ms, TA: 2:44) with a 3 mm slice thickness.

A.3 The WMH challenge data

The WMH challenge [64], initiated at MICCAI 2017, aims to provide a benchmark for automatic segmentation of WMHs of presumed vascular origin and remains open and ongoing[5]. 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 1.11 shows an overview of how the data set is separated into training and test sets. Table 1.12 shows scanning parameters for the 5 scanners.

[5] https://wmh.isi.uu.nl/

Table 1.11. 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 1.12. 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	-	$1.00 \times 1.00 \times 1.00$	192
	2D FLAIR	11,000	125	2,800	$0.96 \times 0.95 \times 3.00$	48
Singapore	3D T1-w	2,300	1,9	900	$1.00 \times 1.00 \times 1.00$	N/A
	2D FLAIR	9,000	82	2,500	$1.00 \times 1.00 \times 3.00$	N/A
AMS GE3T	3D T1-w	7.8	3.0	-	$0.94 \times 0.94 \times 1.00$	176
	3D FLAIR	8,000	126	2,340	$0.98 \times 0.98 \times 1.20$	132
AMS GE1.5T	3D T1-w	12.3	5.2	-	$0.98 \times 0.98 \times 1.50$	172
	3D FLAIR	6,500	117	1,987	$1.21 \times 1.21 \times 1.30$	128
AMS PETMR	3D T1-w	9.9	4.6	-	$0.87 \times 0.87 \times 1.00$	180
	3D FLAIR	4,800	279	1,650	$1.04 \times 1.04 \times 0.56$	321

B Further analysis of the methodology from Chapter 2

B.1 Changes made to SegAE after the original conference paper

A preliminary version of SegAE was presented earlier in conference format [4]. The improvements made in the journal paper [60] were not discussed in Chapter 2. Although the CNN architecture remains the same, the input MRIs and training methodology was substantially modified to improve the performance of the method presented here. The major modifications and the motivation behind each one of them are listed below:

1. *A scale-invariant loss function and a new regularizer to stabilize training*: The loss function used in the preliminary version presented in our conference paper was $L = (Y^3 - \hat{Y}^3)$, where Y is a 3D patch from a FLAIR image and $\hat{Y}^3$ is the corresponding estimated FLAIR image patch. The power of 3 was used to put more weight on the FLAIR hyperintensities, however, there are several drawbacks with this approach: a) It would not work as intended when we add new MRI sequences because WM and CSF have high intensity in T1-w and T2-w images, respectively; b) this image transformation tends to amplify noise and skew the relative tissue intensities; c) difference between true vs. predicted patches depends on the intensity scale, which makes training noisy because of imperfect image normalization. This effect was to some extent mitigated by using higher beta parameters in the Adam optimizer to reduce learning noise. In the new version of SegAE presented in Chapter 2 we use the Cosine proximity function to construct a cost function, which is scale-invariant and does not cause the aforementioned problems (see details in the main text).
   The preliminary version presented in the conference paper had no regularizer as part of the cost function; only scaling of the CNN activations before applying the Softmax function. The reasoning for this was that the Softmax function approaches the argmax function when the input is in the range $[0, \infty)$. However, the CNN eventually learns weights that give non-binary Softmax outputs to lower the cost, so early stopping was needed. In the method presented in Chapter 2, we have an explicit regularization term in the cost function so the solution converges to the expected segmentations.

2. *More MR sequences contributing to the calculation of the loss function*: In the preliminary version we used only FLAIR images for the calculation of the loss function. In the new version presented in Chapter 2 we use T1-w, T2-w, and

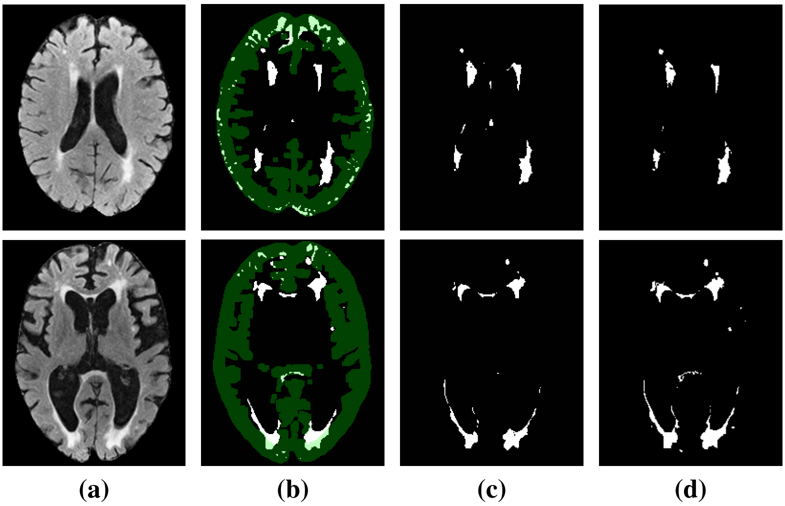


*Figure 2.39. The figure shows **(a)** a FLAIR image, **(b)** the output of the preliminary version of SegAE [4]; green overlay shows the areas that were removed in a special post-processing step using a morphologically eroded brainmask (without sulcal CSF) from FreeSurfer, **(c)** the result from the proposed way of training SegAE, and **(d)** the manually delineated WMHs.*

FLAIR, which makes the method more robust to inhomogeneity artifacts and noise. A comparison of the method using fewer MR sequences is shown below in Section B.4.

3. *An inhomogeneity correction performed during the training phase*: In the previous paper we didn't use any inhomogeneity correction because we found that the N4 bias-correction [48] tended to degrade the WMH lesion segmentations. This resulted in over-segmentation of WMHs in areas around the brain cortex that were removed afterwards with a morphologically eroded brainmask. This is explained in the conference paper and is an obvious disadvantage of the previous method in terms of complexity, processing time, and sensitivity to WMHs (see Figure 2.39 (b)). In the SegAE version presented in Chapter 2 we perform inhomogeneity correction during the training phase, as described below.

B.2 Improved inhomogeneity correction

Chapter 2 includes a section about the N4 bias-correction method and how it can be used to process FLAIR images with WMHs. For the AGES-Reykjavik data set we observed that when using the default settings of N4 (single mesh over the entire domain) the inhomogeneity artifacts were not adequately removed and decreasing the B-spline distance parameter tended to degrade the lesions (see Figure 2.40). This effect is demonstrated in Figure 2.40 (c) and (d), where we decreased the B-spline

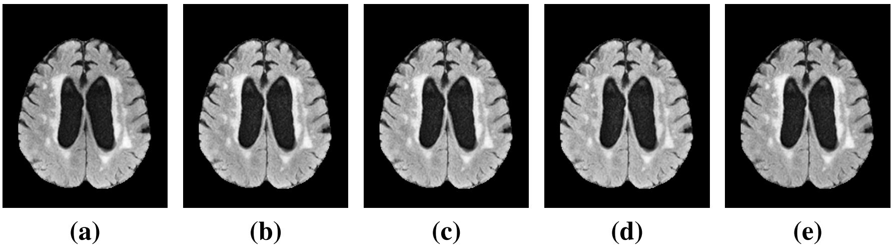


Figure 2.40. Comparison of the N4 bias-correction method when using different B-spline distance showing ***(a)*** *the original FLAIR image (after skullstripping),* ***(b)*** *showing the default settings of N4 (single mesh),* ***(c)*** *the FLAIR image when using 150 mm distance,* ***(d)*** *the FLAIR image when using 100 mm distance, and* ***(e)*** *the FLAIR image when using a pure-tissue probability mask.*

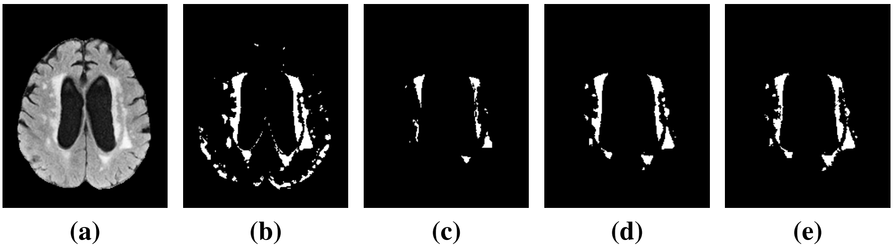


Figure 2.41. The figure shows the WMH output of SegAE trained on ***(b)*** *T1-w, T2-w, and FLAIR images after standard N4 bias-correction; (c) T1-w, T2-w, and FLAIR after N4 correction with pure-tissue probability masks; and (d) intensity transformed T1-w and T2-w images (using corresponding PD-w images) and N4 bias-corrected FLAIR images using pure-tissue probability masks. Figure (a) shows the FLAIR image and (e) shows the manually delineated WMH mask for comparison.*

distance to 150 mm and 100 mm, respectively, leading to substantial improvement of the inhomogeneity artifacts in the cortical regions, however, at the same time causing the WM lesions to lose contrast. To address this we used the pure-tissue probability mask option proposed in [102]. This enabled us to use N4 for estimating the bias field without affecting the WMHs, by excluding the WMH regions when doing the bias field estimation.

Furthermore, the T1-w and T2-w images of the AGES-Reykjavik data set were processed differently in the proposed method: They were intensity transformed using the corresponding PD-w images to correct the bias-field, in particular, inhomogeneity artifacts in the lateral ventricles in the T2-w images (see Figure 2.13 in the main text), and for contrast enhancement of the T1-w images. Comparison of three SegAE models, trained to reconstruct images that have been bias corrected with 1) standard N4, 2) N4 with pure-tissue probability mask, and 3) N4 with pure-tissue probability mask for processing the FLAIR image but intensity transformation with a PD-w image for the T1-w and T2-w images, can be seen in Figure 2.41.

B.3 DSC and volume comparison

A comparison of the lesion volumes and Dice Similarity Coefficient (DSC) for the old SegAE scheme, SegAE using standard N4 bias-correction, and the proposed SegAE for 15 subjects (same subjects as in [4]) can be seen in Figure 2.42. Using standard N4 bias correction gives lower DSC in all cases, and using the old method of training and post-processing SegAE gives lower DSC in 10 out of 15 cases.

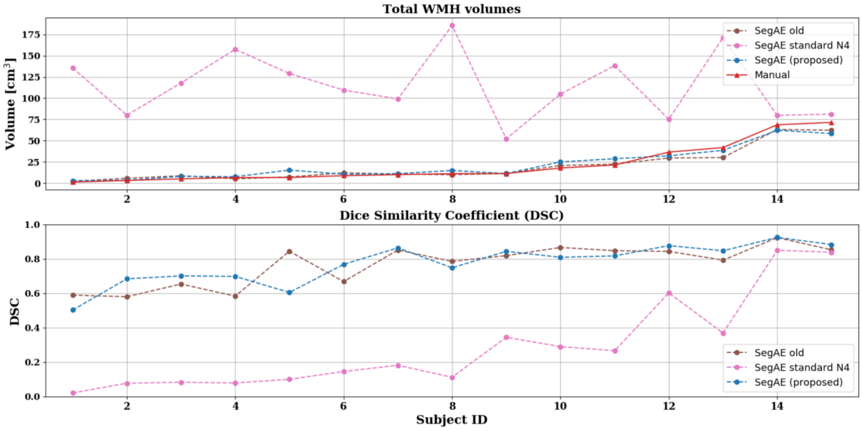


Figure 2.42. The top graph shows the overall WMH volume for the manual masks (red) and masks generated by old SegAE (brown, dotted), SegAE with standard N4 (pink, dotted), and proposed SegAE (blue, dotted), ordered by the volume of the manual masks. The bottom graph shows the DSC for the same methods compared with the manual masks.

B.4 Number of input sequences

In Chapter 2 we present results on two separate data sets, i.e. the AGES-Reykjavik data set and the WMH challenge data set. The AGES-Reykjavik data set consists of T1-w, T2-w, FLAIR, and PD-w images, while the WMH challenge data set comprises only T1-w and FLAIR images.

To evaluate the effects of using a different number of input sequences we conducted several experiments by training SegAE using 1) only FLAIR images, 2) T1-w and FLAIR images and, 3) T1-w, T2-w, and FLAIR images (and PD-w images for image enhancement). A visual comparison of the WMH segmentations can be seen in Figure 2.43. Using more input sequences, if available, seems to improve the robustness to cases with severe inhomogeneity artifacts that can not be completely removed with the N4 bias-correction, and improve robustness to noise.

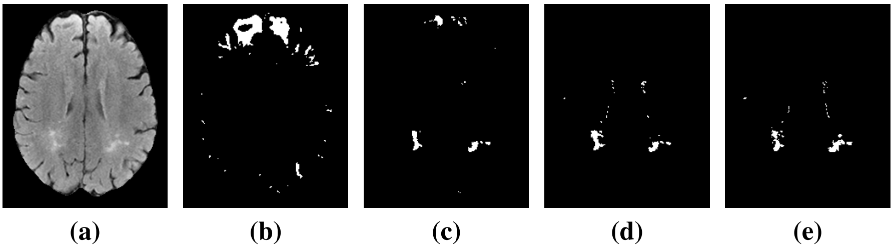


Figure 2.43. WMH segmentation output from SegAE using different input sequences. ***(a)*** *the FLAIR image;* ***(b)*** *using only FLAIR images as input;* ***(c)*** *using T1-w, and FLAIR images as input;* ***(d)*** *the proposed SegAE using T1-w, T2-w, and FLAIR images as input; and* ***(e)*** *the manually delineated mask.*

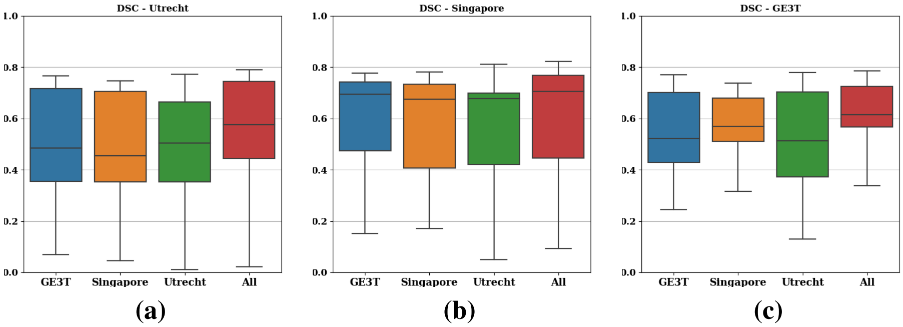


Figure 2.44. DSC comparision of SegAE models trained using only images from GE3T (blue), Singapore (orange), Utrecht (green), as well as images from all scanners combined (red). ***(a)*** *shows the results evaluated on images from the Utrecht training set,* ***(b)*** *shows results evaluated on the Singapore training set, and* ***(c)*** *shows results evaluated on the GE3T training set.*

B.5 WMH challenge models

SegAE was submitted to the WMH challenge (MICCAI 2017 [64]). Training data from three scanners were provided; GE3T (20 subjects), Singapore (20 subjects), and Utrecht (20 subjects). We trained SegAE on training data from all three scanners simultaneously and submitted this model to the challenge.

During training, SegAE is sensitive to the contrast of the training images, so here we explore whether the performance on the training set would improve if SegAE was trained on data from each scanner separately. Figure 2.44 shows boxplots of the DSC scores achieved for each training set separately when SegAE models are trained on each training set separately, as well as all the data simultaneously, as was done in the challenge. These results suggest that training SegAE using all available training data improves the performance on all the WMH challenge data sets, even though the data come from different scanners.

Table 2.13. Results of SegAE models trained using 6 different configurations on a simulated T1-w image from the BrainWEB database. DSC values are shown for the CSF, GM, and WM using both 0% inhomogeneity (RF0) and 40% inhomogeneity (RF40). The weights a and b determine which terms are used in the loss function.

Loss weights	CSF_{RF0}	GM_{RF0}	WM_{RF0}	CSF_{RF40}	GM_{RF40}	WM_{RF40}
$a=1,b=0$	0.884	0.892	0.889	0.521	0.000	0.831
$a=0,b=1$	0.685	0.626	0.219	0.904	0.738	0.584
$a=1,b=1$	0.930	0.927	0.923	0.909	0.916	0.936

B.6 Evaluation on a synthetic image

During the development of SegAE, a synthetic image from the BrainWEB database [134] was used to evaluate the effect of using the Laplace operator in the loss function and the quality of the tissue segmentation. A single synthetic T1-w image of a normal subject with no lesions was used as input. The voxel size was 1x1x1 mm and 3% noise was added (calculated relative to the brightest tissue). Three SegAE outputs for materials were used (M=3). The training was performed with regularization coefficient α=0.01, and 1000 epochs. SegAE was trained seperately on images with both 0% and 40% intensity non-uniformity. In equation 8, the weights a and b are introduced to the SegAE loss function. Separate training runs were performed to evaluate the effect of the two terms of the loss function by assigning the values 0 and 1 to the weights a and b as shown in Table 2.13. Furthermore, Table 2.13 shows the DSC values of six SegAE models trained using three different configurations of a and b and two different inhomogeneity levels.

$$L(\boldsymbol{Y},\hat{\boldsymbol{Y}}) = -\frac{1}{C}\sum_{c=1}^{C}(a\cdot f(\mathrm{vec}(\boldsymbol{Y}_c),\mathrm{vec}(\hat{\boldsymbol{Y}}_c)) + b\cdot f(\mathrm{vec}(K*\boldsymbol{Y}_c),\mathrm{vec}(K*\hat{\boldsymbol{Y}}_c))), \qquad (8)$$

This test indicates that including the term with the Laplace operator in the loss function increases the robustness of SegAE to field inhomogeneity compared to using only Cosine proximity. Furthermore, this test shows that SegAE can segment the GM, WM, and CSF using only one simulated T1-w image as input.

References

[1] J. Virhammar, K. Laurell, K. G. Cesarini, and E.-M. Larsson, “Increase in callosal angle and decrease in ventricular volume after shunt surgery in patients with idiopathic normal pressure hydrocephalus,” *Journal of neurosurgery*, vol. 130, no. 1, pp. 130–135, 2018.

[2] W. A. Evans, “An encephalographic ratio for estimating ventricular enlargement and cerebral atrophy,” *Archives of Neurology & Psychiatry*, vol. 47, no. 6, pp. 931–937, 1942.

[3] F. Bre, J. M. Gimenez, and V. D. Fachinotti, “Prediction of wind pressure coefficients on building surfaces using artificial neural networks,” *Energy and Buildings*, vol. 158, pp. 1429–1441, 2018.

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

[5] F. Fazekas, J. B. Chawluk, A. Alavi, H. I. Hurtig, and R. A. Zimmerman, “Mr signal abnormalities at 1.5 t in alzheimer’s dementia and normal aging,” *American journal of roentgenology*, vol. 149, no. 2, pp. 351–356, 1987.

[6] Y. Hou, X. Dan, M. Babbar, Y. Wei, S. G. Hasselbalch, D. L. Croteau, and V. A. Bohr, “Ageing as a risk factor for neurodegenerative disease,” *Nature Reviews Neurology*, vol. 15, no. 10, pp. 565–581, 2019.

[7] J. A. Schneider, Z. Arvanitakis, W. Bang, and D. A. Bennett, “Mixed brain pathologies account for most dementia cases in community-dwelling older persons,” *Neurology*, vol. 69, no. 24, pp. 2197–2204, 2007.

[8] R. Harvey, M. Skelton-Robinson, and M. Rossor, “The prevalence and causes of dementia in people under the age of 65 years,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 9, pp. 1206–1209, 2003.

[9] 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*, vol. 79, no. 6, pp. 929–939, 2016.

[10] J.-P. Coutu, A. Goldblatt, H. D. Rosas, and D. H. Salat, “White matter changes are associated with ventricular expansion in aging, mild cognitive impairment, and alzheimer’s disease,” *Journal of Alzheimer’s Disease*, vol. 49, no. 2, pp. 329–342, 2016.

[11] D. Purves, G. J. Agustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, R. D. Mooney, M. L.Platt, and L. E. White, *Neuroscience*. 23 Plumtree Road, Sunderland MA01375 USA: Oxford University Press, 2018.

[12] J. Prince and J. Links, *Medical Imaging Signals and Systems*. Pearson Prentice Hall bioengineering, Pearson Prentice Hall, 2006.

[13] Y. Kurihara, T. M. Simonson, H. D. Nguyen, D. J. Fisher, C.-S. L. Y. Sato, and W. T. Yuh, "Mr imaging of ventriculomegaly—a qualitative and quantitative comparison of communicating hydrocephalus, central atrophy, and normal studies," *Journal of Magnetic Resonance Imaging*, vol. 5, no. 4, pp. 451–456, 1995.

[14] J. Kreutzer, J. DeLuca, and B. Caplan, eds., *Encyclopedia of Clinical Neuropsychology*. 233 Spring Street, New York, NY 10013, USA: SpringerScienceþBusiness Media, LLC, 2011.

[15] A. E. George, A. Holodny, J. Golomb, and M. J. de Leon, "The differential diagnosis of Alzheimer's disease. Cerebral atrophy versus normal pressure hydrocephalus," *Neuroimaging Clin. N. Am.*, vol. 5, pp. 19–31, Feb 1995.

[16] D. Arnone, J. Cavanagh, D. Gerber, S. Lawrie, K. Ebmeier, and A. McIntosh, "Magnetic resonance imaging studies in bipolar disorder and schizophrenia: meta-analysis," *The British Journal of Psychiatry*, vol. 195, no. 3, pp. 194–201, 2009.

[17] A. Brean and P. Eide, "Prevalence of probable idiopathic normal pressure hydrocephalus in a norwegian population," *Acta neurologica Scandinavica*, vol. 118, no. 1, pp. 48–53, 2008.

[18] D. Shprecher, J. Schwalb, and R. Kurlan, "Normal pressure hydrocephalus: diagnosis and treatment," *Current neurology and neuroscience reports*, vol. 8, no. 5, pp. 371–376, 2008.

[19] S. Hakim and R. Adams, "The special clinical problem of symptomatic hydrocephalus with normal cerebrospinal fluid pressure: observations on cerebrospinal fluid hydrodynamics," *Journal of the neurological sciences*, vol. 2, no. 4, pp. 307–327, 1965.

[20] A. M. Fjell and K. B. Walhovd, "Structural brain changes in aging: courses, causes and cognitive consequences," *Rev Neurosci*, vol. 21, no. 3, pp. 187–221, 2010.

[21] M. G. Erkkinen, M.-O. Kim, and M. D. Geschwind, "Clinical neurology and epidemiology of the major neurodegenerative diseases," *Cold Spring Harbor perspectives in biology*, vol. 10, no. 4, p. a033118, 2018.

[22] J. Virhammar, K. Laurell, K. G. Cesarini, and E.-M. Larsson, "The callosal angle measured on mri as a predictor of outcome in idiopathic normal-pressure hydrocephalus," *Journal of neurosurgery*, vol. 120, no. 1, pp. 178–184, 2014.

[23] A. K. Toma, E. Holl, N. D. Kitchen, and L. D. Watkins, "Evans' index revisited: the need for an alternative in normal pressure hydrocephalus," *Neurosurgery*, vol. 68, no. 4, pp. 939–944, 2011.

[24] N. M. Zahr, D. Mayer, T. Rohlfing, J. Orduna, R. Luong, E. V. Sullivan, and A. Pfefferbaum, "A mechanism of rapidly reversible cerebral ventricular enlarge-

ment independent of tissue atrophy," *Neuropsychopharmacology*, vol. 38, no. 6, pp. 1121–1129, 2013.

[25] A. M. Tawfik, L. Elsorogy, R. Abdelghaffar, A. A. Naby, and I. Elmenshawi, "Phase-contrast mri csf flow measurements for the diagnosis of normal-pressure hydrocephalus: observer agreement of velocity versus volume parameters," *American Journal of Roentgenology*, vol. 208, no. 4, pp. 838–843, 2017.

[26] J. M. Wardlaw, M. C. Valdés Hernández, and S. Muñoz-Maniega, "What are white matter hyperintensities made of? Relevance to vascular cognitive impairment," *Journal of the American Heart Association*, vol. 4, no. 6, p. e001140, 2015.

[27] 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*, vol. 12, no. 8, pp. 822–838, 2013.

[28] S. Debette and H. Markus, "The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis," *Bmj*, vol. 341, p. c3666, 2010.

[29] 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, and J. Coresh, "Association of Brain Magnetic Resonance Imaging Signs With Cognitive Outcomes in Persons With Nonimpaired Cognition and Mild Cognitive Impairment," *JAMA Netw Open*, vol. 2, p. e193359, May 2019.

[30] W. Longstreth, T. A. Manolio, A. Arnold, G. L. Burke, N. Bryan, C. A. Jungreis, P. L. Enright, D. O'Leary, and L. Fried, "Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people: the Cardiovascular Health Study," *Stroke*, vol. 27, no. 8, pp. 1274–1282, 1996.

[31] 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*, vol. 59, no. 4, pp. 3862–3870, 2012.

[32] S. B. Wharton, J. E. Simpson, C. Brayne, and P. G. Ince, "Age-Associated White Matter Lesions: The MRC Cognitive Function and Ageing Study," *Brain pathology*, vol. 25, no. 1, pp. 35–43, 2015.

[33] S. M. Maniega, M. C. V. Hernández, J. D. Clayden, N. A. Royle, C. Murray, Z. Morris, B. S. Aribisala, A. J. Gow, J. M. Starr, M. E. Bastin, *et al.*, "White matter hyperintensities and normal-appearing white matter integrity in the aging brain," *Neurobiology of aging*, vol. 36, no. 2, pp. 909–918, 2015.

[34] D. Jericó, E. O. Luis, L. Cussó, M. A. Fernández-Seara, X. Morales, K. M. Córdoba, M. Benito, A. Sampedro, M. Larriva, M. J. Ramírez, *et al.*, "Brain ventricular enlargement in human and murine acute intermittent porphyria," *Human Molecular Genetics*, vol. 29, no. 19, pp. 3211–3223, 2020.

[35] R. Geraldes, O. Ciccarelli, F. Barkhof, N. De Stefano, C. Enzinger, M. Filippi, M. Hofer, F. Paul, P. Preziosa, A. Rovira, *et al.*, "The current role of mri in

differentiating multiple sclerosis from its imaging mimics," *Nature Reviews Neurology*, vol. 14, no. 4, p. 199, 2018.

[36] T. van Eimeren, A. Antonini, D. Berg, N. Bohnen, R. Ceravolo, A. Drzezga, G. U. Höglinger, M. Higuchi, S. Lehericy, S. Lewis, *et al.*, "Neuroimaging biomarkers for clinical trials in atypical parkinsonian disorders: Proposal for a neuroimaging biomarker utility system," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 11, no. 1, pp. 301–309, 2019.

[37] B. S. Passiak, D. Liu, H. A. Kresge, F. E. Cambronero, K. R. Pechman, K. E. Osborn, K. A. Gifford, T. J. Hohman, M. S. Schrag, L. T. Davis, *et al.*, "Perivascular spaces contribute to cognition beyond other small vessel disease markers," *Neurology*, vol. 92, no. 12, pp. e1309–e1321, 2019.

[38] E. J. Vinke, M. De Groot, V. Venkatraghavan, S. Klein, W. J. Niessen, M. A. Ikram, and M. W. Vernooij, "Trajectories of imaging markers in brain aging: the Rotterdam Study," *Neurobiology of aging*, vol. 71, pp. 32–40, 2018.

[39] A. P. Appelman, L. G. Exalto, Y. Van Der Graaf, G. J. Biessels, W. P. Mali, and M. I. Geerlings, "White matter lesions and brain atrophy: more than shared risk factors? a systematic review," *Cerebrovascular Diseases*, vol. 28, no. 3, pp. 227–242, 2009.

[40] W. Palm, J. van der Grond, J. Milles, S. Sigurdsson, G. Eiriksdottir, V. Gudnason, L. Launer, and M. van Buchem, "Disproportionate ventricular dilatation in the elderly could be a manifestation of small vessel disease," *Ventricular dilation in agingi and dementia (PhD thesis)*, p. 75.

[41] R. Wang, L. Fratiglioni, A. Laveskog, G. Kalpouzos, C.-H. Ehrenkrona, Y. Zhang, L. Bronge, L.-O. Wahlund, L. Bäckman, and C. Qiu, "Do cardiovascular risk factors explain the link between white matter hyperintensities and brain volumes in old age? a population-based study," *European journal of neurology*, vol. 21, no. 8, pp. 1076–1082, 2014.

[42] A. L. Lumsden, A. Mulugeta, A. Zhou, and E. Hyppönen, "Apolipoprotein e (apoe) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the uk biobank," *EBioMedicine*, vol. 59, p. 102954, 2020.

[43] P. Scheltens, T. Erkinjunti, D. Leys, L.-O. Wahlund, D. Inzitari, T. del Ser, F. Pasquier, F. Barkhof, R. Mäntylä, J. Bowler, *et al.*, "White matter changes on ct and mri: an overview of visual rating scales," *European neurology*, vol. 39, no. 2, pp. 80–89, 1998.

[44] 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*, vol. 148, pp. 77–102, 2017.

[45] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, J. Lancaster, *et al.*, "A probabilistic atlas of the human brain: theory and rationale for its development," *Neuroimage*, vol. 2, no. 2, pp. 89–101, 1995.

[46] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. Almli, and D. Collins, "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood," *NeuroImage*, no. 47, p. S102, 2009.

[47] U. Vovk, F. Pernus, and B. Likar, "A review of methods for correction of intensity inhomogeneity in MRI," *IEEE transactions on medical imaging*, vol. 26, no. 3, pp. 405–421, 2007.

[48] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: improved N3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[49] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. B. Cuadra, "A review of atlas-based segmentation for magnetic resonance brain images," *Computer methods and programs in biomedicine*, vol. 104, no. 3, pp. e158–e177, 2011.

[50] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical image analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[51] B. Fischl, "Freesurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[52] L. M. Ellingsen, S. Roy, A. Carass, A. M. Blitz, D. L. Pham, and J. L. Prince, "Segmentation and labeling of the ventricular system in normal pressure hydrocephalus using patch-based tissue classification and multi-atlas labeling," in *Medical Imaging 2016: Image Processing*, vol. 9784, p. 97840G, International Society for Optics and Photonics, 2016.

[53] C. Ledig, R. A. Heckemann, A. Hammers, J. C. Lopez, V. F. Newcombe, A. Makropoulos, J. Lötjönen, D. K. Menon, and D. Rueckert, "Robust whole-brain segmentation: application to traumatic brain injury," *Medical image analysis*, vol. 21, no. 1, pp. 40–58, 2015.

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

[55] H. E. Atlason, M. Shao, V. Robertsson, S. Sigurdsson, V. Gudnason, J. L. Prince, and L. M. Ellingsen, "Large-scale parcellation of the ventricular system using convolutional neural networks," in *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10953, p. 109530N, International Society for Optics and Photonics, 2019.

[56] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 611–623, 2012.

[57] Y. Huo, A. J. Plassard, A. Carass, S. M. Resnick, D. L. Pham, J. L. Prince, and B. A. Landman, "Consistent cortical reconstruction and multi-atlas brain segmentation," *NeuroImage*, vol. 138, pp. 197–210, 2016.

[58] S. González-Villà, A. Oliver, Y. Huo, X. Lladó, and B. A. Landman, "Brain structure segmentation in the presence of multiple sclerosis lesions," *NeuroImage: Clinical*, vol. 22, p. 101709, 2019.

[59] R. Bakshi, S. D. Caruthers, V. Janardhan, and M. Wasay, “Intraventricular csf pulsation artifact on fast fluid-attenuated inversion-recovery MR images: analysis of 100 consecutive normal studies,” *American Journal of Neuroradiology*, vol. 21, no. 3, pp. 503–508, 2000.

[60] H. E. Atlason, A. Love, S. Sigurdsson, V. Gudnason, and L. M. Ellingsen, “Segae: Unsupervised white matter lesion segmentation from brain mris using a cnn autoencoder,” *NeuroImage: Clinical*, vol. 24, p. 102085, 2019.

[61] X. Wang, M. C. V. Hernández, F. Doubal, F. M. Chappell, and J. M. Wardlaw, “How much do focal infarcts distort white matter lesions and global cerebral atrophy measures?,” *Cerebrovascular Diseases*, vol. 34, no. 5-6, pp. 336–342, 2012.

[62] A. Carass, M. Shao, X. Li, B. E. Dewey, A. M. Blitz, S. Roy, D. L. Pham, J. L. Prince, and L. M. Ellingsen, “Whole brain parcellation with pathology: Validation on ventriculomegaly patients,” in *International Workshop on Patch-Based Techniques in Medical Imaging*, pp. 20–28, Springer, 2017.

[63] M. Shao, S. Han, A. Carass, X. Li, A. M. Blitz, J. Shin, J. L. Prince, and L. M. Ellingsen, “Brain ventricle parcellation using a deep neural network: Application to patients with ventriculomegaly,” *NeuroImage: Clinical*, vol. 23, p. 101871, 2019.

[64] 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.

[65] 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*, vol. 17, pp. 918–934, 2018.

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

[67] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira, “Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches,” *Information Sciences*, vol. 186, no. 1, pp. 164–185, 2012.

[68] J.-C. Souplet, C. Lebrun, N. Ayache, and G. Malandain, “An automatic segmentation of t2-flair multiple sclerosis lesions,” in *MICCAI-Multiple sclerosis lesion segmentation challenge workshop*, 2008.

[69] T. B. Harris, L. J. Launer, G. Eiriksdottir, O. Kjartansson, P. V. Jonsson, G. Sigurdsson, G. Thorgeirsson, T. Aspelund, M. E. Garcia, M. F. Cotch, H. J. Hoffman, and V. Gudnason, “Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics,” *Am. J. Epidemiol.*, vol. 165, pp. 1076–1087, May 2007.

[70] T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy, *et al.*, "The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[71] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults," *Journal of cognitive neuroscience*, vol. 22, no. 12, pp. 2677–2684, 2010.

[72] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.

[73] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. Vlassenko, *et al.*, "Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease," *MedRxiv*, 2019.

[74] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack Jr, W. Jagust, J. C. Morris, *et al.*, "The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement," *Alzheimer's & Dementia*, vol. 13, no. 5, pp. 561–571, 2017.

[75] B. C. Csáji *et al.*, "Approximation with artificial neural networks," *Faculty of Sciences, Etvs Lornd University, Hungary*, vol. 24, no. 48, p. 7, 2001.

[76] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A guide to convolutional neural networks for computer vision," *Synthesis Lectures on Computer Vision*, vol. 8, no. 1, pp. 1–207, 2018.

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

[78] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[79] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.

[80] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[81] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[82] M. Shao, S. Han, A. Carass, X. Li, A. M. Blitz, J. L. Prince, and L. M. Ellingsen, "Shortcomings of ventricle segmentation using deep convolutional networks," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pp. 79–86, Springer, 2018.

[83] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[84] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "GAN augmentation: augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018.

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

[86] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[87] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[88] 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 Medical Image Computing and Computer-Assisted Intervention 2017*, pp. 516–524, Springer, 2017.

[89] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato, "Cross-modality image synthesis from unpaired data using cyclegan," in *International workshop on simulation and synthesis in medical imaging*, pp. 31–41, Springer, 2018.

[90] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 240–248, Springer, 2017.

[91] M. M. Lopez and J. Ventura, "Dilated convolutions for brain tumor segmentation in mri scans," in *International MICCAI Brainlesion Workshop*, pp. 253–262, Springer, 2017.

[92] X. Tomas-Fernandez and S. K. Warfield, "A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation," *IEEE transactions on medical imaging*, vol. 34, no. 6, pp. 1349–1361, 2015.

[93] R. Khayati, M. Vafadust, F. Towhidkhah, and 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.*, vol. 38, pp. 379–390, Mar 2008.

[94] 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, and 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*, vol. 32, no. 1, pp. 24–34, 2010.

[95] K. Krupa and M. Bekiesińska-Figatowska, "Artifacts in magnetic resonance imaging," *Polish journal of radiology*, vol. 80, p. 93, 2015.

[96] C. Bowles, C. Qin, R. Guerrero, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "Brain lesion segmentation through image synthesis and outlier detection," *NeuroImage: Clinical*, vol. 16, pp. 643–658, 2017.

[97] C. Baur, B. Wiestler, S. Albarqouni, and 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, 2019.

[98] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study," *Medical Image Analysis*, p. 101952, 2021.

[99] B. Palsson, J. Sigurdsson, J. R. Sveinsson, and M. O. Ulfarsson, "Hyperspectral unmixing using a neural network autoencoder," *IEEE Access*, vol. 6, pp. 25646–25656, 2018.

[100] B. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Convolutional autoencoder for spectral–spatial hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 535–549, 2020.

[101] S. Roy, J. A. Butman, and D. L. Pham, "Robust skull stripping using multiple MR image contrasts insensitive to pathology," *Neuroimage*, vol. 146, pp. 132–147, Feb 2017.

[102] 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*, vol. 99, pp. 166–179, 2014.

[103] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[104] T. Dozat, "Incorporating nesterov momentum into adam." `https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ`, 2016.

[105] X. Glorot and 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, 2010.

[106] 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, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[107] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

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

[109] 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*, vol. 59, no. 4, pp. 3774–3783, 2012.

[110] P. Schmidt, *Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging*. PhD thesis, lmu, 2017.

[111] H. E. Atlason, A. Love, S. Sigurdsson, V. Gudnason, and L. M. Ellingsen, "Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder," *SPIE Medical Imaging 2019: Image Processing*, vol. accepted for publication, 2019.

[112] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer, 2016.

[113] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, 2015.

[114] M. Jenkinson, M. Pechaud, S. Smith, *et al.*, "Bet2: Mr-based estimation of brain, skull and scalp surfaces," in *Eleventh annual meeting of the organization for human brain mapping*, vol. 17, p. 167, Toronto., 2005.

[115] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, D. L. Collins, A. D. N. Initiative, *et al.*, "Beast: brain extraction based on nonlocal segmentation technique," *NeuroImage*, vol. 59, no. 3, pp. 2362–2373, 2012.

[116] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," *IEEE transactions on medical imaging*, vol. 30, no. 9, pp. 1617–1634, 2011.

[117] C. Wachinger, M. Reuter, and T. Klein, "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy," *NeuroImage*, vol. 170, pp. 434–445, 2018.

[118] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep MRI brain extraction: A 3d convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, 2016.

[119] S. Roy, A. Knutsen, A. Korotcov, A. Bosomtwi, B. Dardzinski, J. A. Butman, and D. L. Pham, “A deep learning framework for brain extraction in humans and animals with traumatic brain injury,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 687–691, IEEE, 2018.

[120] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[121] A. Carass, S. Roy, A. Gherman, J. C. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel, *et al.*, “Evaluating white matter lesion segmentations with refined sørensen-dice analysis,” *Scientific Reports*, vol. 10, no. 1, pp. 1–19, 2020.

[122] Y. Wang, Y. Zhang, Y. Liu, Z. Lin, J. Tian, C. Zhong, Z. Shi, J. Fan, and Z. He, “ACN: Adversarial co-training network for brain tumor segmentation with missing modalities,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 410–420, Springer, 2021.

[123] L. Shen, W. Zhu, X. Wang, L. Xing, J. M. Pauly, B. Turkbey, S. A. Harmon, T. H. Sanford, S. Mehralivand, P. L. Choyke, *et al.*, “Multi-domain image completion for random missing input data,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 4, pp. 1113–1122, 2020.

[124] M. Dadar, T. A. Pascoal, S. Manitsirikul, K. Misquitta, V. S. Fonov, M. C. Tartaglia, J. Breitner, P. Rosa-Neto, O. T. Carmichael, C. Decarli, *et al.*, “Validation of a regression technique for segmentation of white matter hyperintensities in Alzheimer’s disease,” *IEEE transactions on medical imaging*, vol. 36, no. 8, pp. 1758–1768, 2017.

[125] C. Dufouil, A. de Kersaint-Gilly, V. Besancon, C. Levy, E. Auffray, L. Brunnereau, A. Alperovitch, and C. Tzourio, “Longitudinal study of blood pressure and white matter hyperintensities: the eva mri cohort,” *Neurology*, vol. 56, no. 7, pp. 921–926, 2001.

[126] J. H. Lee, S. Yoon, P. F. Renshaw, T.-S. Kim, J. J. Jung, Y. Choi, B. N. Kim, A. M. Jacobson, and I. K. Lyoo, “Morphometric changes in lateral ventricles of patients with recent-onset type 2 diabetes mellitus,” *PLoS One*, vol. 8, no. 4, 2013.

[127] E. S. Korf, L. R. White, P. Scheltens, and L. J. Launer, “Brain aging in very old men with type 2 diabetes: the honolulu-asia aging study,” *Diabetes care*, vol. 29, no. 10, pp. 2268–2274, 2006.

[128] S. H. Kim, C.-H. Yun, S.-Y. Lee, K.-h. Choi, M. B. Kim, and H.-K. Park, “Age-dependent association between cigarette smoking on white matter hyperintensities,” *Neurological Sciences*, vol. 33, no. 1, pp. 45–51, 2012.

[129] T. Lohse, S. Rohrmann, M. Bopp, and D. Faeh, “Heavy smoking is more strongly associated with general unhealthy lifestyle than obesity and underweight,” *PloS one*, vol. 11, no. 2, p. e0148563, 2016.

[130] I. Siasios, E. Z. Kapsalaki, K. N. Fountas, A. Fotiadou, A. Dorsch, K. Vakharia, J. Pollina, and V. Dimopoulos, "The role of diffusion tensor imaging and fractional anisotropy in the evaluation of patients with idiopathic normal pressure hydrocephalus: a literature review," *Neurosurgical focus*, vol. 41, no. 3, p. E12, 2016.

[131] M. Habes, G. Erus, J. B. Toledo, T. Zhang, N. Bryan, L. J. Launer, Y. Rosseel, D. Janowitz, J. Doshi, S. Van der Auwera, *et al.*, "White matter hyperintensities and imaging patterns of brain ageing in the general population," *Brain*, vol. 139, no. 4, pp. 1164–1179, 2016.

[132] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review," *Neuroinformatics*, vol. 13, no. 3, pp. 261–276, 2015.

[133] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, p. 117012, 2020.

[134] C. A. Cocosco, V. Kollokian, R. K.-S. Kwan, G. B. Pike, and A. C. Evans, "Brainweb: Online interface to a 3D MRI simulated brain database," in *NeuroImage*, Citeseer, 1997.

[135] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu, *et al.*, "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014.