THE UNIVERSITY
of EDINBURGH

Development of Machine Learning Schemes for Segmentation, Characterisation, and Evolution Prediction of White Matter Hyperintensities in Structural Brain MRI

Muhammad Febrian Rachmadi

Doctor of Philosophy
Institute of Perception, Action and Behaviour
School of Informatics
University of Edinburgh
2020

Abstract

White matter hyperintensities (WMH) are neuroradiological features seen in T2 Fluid-Attenuated Inversion Recovery (T2-FLAIR) brain magnetic resonance imaging (MRI) and have been commonly associated with stroke, ageing, dementia, and Alzheimer's disease (AD) progression. As a marker of neuro-degenerative disease, WMH may change over time and follow the clinical condition of the patient. In contrast to the early longitudinal studies of WMH, recent studies have suggested that the progression of WMH may be a dynamic, non-linear process where different clusters of WMH may shrink, stay unchanged, or grow. In this thesis, these changes are referred to as the "evolution of WMH".

The main objective of this thesis is to develop machine learning methods for prediction of WMH evolution in structural brain MRI from one time (baseline) assessment. Predicting the evolution of WMH is challenging because the rate and direction of WMH evolution varies greatly across previous studies. Furthermore, the evolution of WMH is a non-deterministic problem because some clinical factors that possibly influence it are still not known. In this thesis, different learning schemes of deep learning algorithm and data modalities are proposed to produce the best estimation of WMH evolution. Furthermore, a scheme to simulate the non-deterministic nature of WMH evolution, named auxiliary input, was also proposed. In addition to the development of prediction model for WMH evolution, machine learning methods for segmentation of early WMH, characterisation of WMH, and simulation of WMH progression and regression are also developed as parts of this thesis.

Lay Summary

Most previous researches in medical imaging performed cross-sectional analysis, where detection and diagnosis of pathologies are carried out from one assessment to find the current state of pathology in a patient. While cross-sectional analysis is sufficient for some pathologies, longitudinal analysis, where two or more assessments are carried out, is more suitable for degenerative pathologies as they increasingly affect tissues or organs and deteriorate over time. One example of degenerative disease is Alzheimer's disease (AD), which is a neuro-degenerative disease which affects the cognitive capability of a patient. This thesis propose a predictive model named Disease Evolution Predictor (DEP) for predicting the evolution (i.e., progression and regression) of white matter hyperintensities (WMH) which appear on brain magnetic resonance imaging (MRI) from a longitudinal dataset. WMH themselves have been associated with the progression of dementia and AD. This thesis examines whether the state-of-the-art deep learning algorithms can be used for such task. In addition to that, this thesis also demonstrates how segmentation of early and subtle WMH can be improved, proposes a novel unsupervised method for characterising WMH, and demonstrates how WMH progression and regression can be simulated using a novel irregularity map.

Acknowledgements

Asyhadu allaa ilaaha illallaah, wa asyhadu anna muhammadur-rasuulullaah.
Allahummaa shalli 'ala muhammad, wa 'ala alii muhammad.
Fil 'alamina innaka hamidun majid.

Allhamdulillaah hirabbil 'aalamiin. First and foremost, I would like to thank the God, *Allah subhaanallahu wata'aala*, who has guided me in the way of Islam since I was born until now. Without the God's help and will, I would go astray and eventually would not even survive for even one second in this vast and challenging world.

My sincere gratitude goes to my supervisors, Prof. Taku Komura and Dr. Maria Valdés-Hernández, who have helped me from the beginning of my master's research project to the end of this doctoral thesis. Thank you very much for your time, advice, and knowledge that both of you have passed to me. Furthermore, I also would like to thank all of people who have shared their expertise and advice throughout this journey.

My biggest gratitude in this world goes to my family in Indonesia, especially my mother, my father, and my older brother, who have prayed for me in the middle of the night every day so that I can complete my study as soon as possible. I also would like to thank my big and extended families who have shown supports and love to me since I was a kid. I am really glad to be a member of this family. I miss you all.

Furthermore, I would like to express my gratitude and love to Novianti Sri Wahyuni, who has been very patient and understanding to me while I was studying in the UK, far away from Indonesia. May the God gives us more patient, gratitude, and blessing in the next phase of our lives. I am indeed very lucky to be able to meet you.

I also would like to thank all of my colleagues in the Computer Graphics and Visualisation (CGVU) research group, especially Yunhee, Floyd, Daniel, Geng, Ian, Cian, Levi, He, Kunkun, and the others. Huge appreciation also goes to my fellow Indonesian PhD students at the University of Edinburgh, especially Aji, Danny, Krisna, Clara, Clarissa, and Viona. Thank you for the discussions, games, foods, and lunches shared together in the past four years. Hope I can see you guys in the next future.

The last but not the least, I would like to thank the Indonesian Government, especially the Indonesian Endowment Fund for Education (Indonesian: *Lembaga Pengelola Dana Pendidikan* or LPDP) of the Ministry of Finance, who has given a chance for me to continue my study in the UK. The future of Indonesia is not easy and challenging, and it is up to us to make Indonesia a better nation for all its citizens. Let's do our best for our nation.

Funds from Row Fogo Charitable Trust (Grant No. BRO-D.FID3668413), Wellcome Trust (Ref No. 088134/Z/0), European Union Horizon 2020 (PHC-03-15, project No 666881, 'SVDs@Target'), Fondation Leducq (CVD 16/05), and the UK Dementia Research Institute at the University of Edinburgh are gratefully acknowledged. The Titan Xp used for this research was donated by the NVIDIA Corporation.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute ofBiomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Tokyo, 17/MAR/2020

(*Muhammad Febrian Rachmadi*)

“All knowledge in this world belongs to the God,
so it must be returned back to the God
as good deeds in this world.”
– My late grandfather, Muhammad Zubet (1933-2018).

Contents

Nomenclature

$*$	Convolution operation
DSC	Dice similarity coefficient
D	Volumetric Disagreement of intra-/inter-observer reliability measurement
FN	False negative
FP	False positive
GT	ground truth
IM	Irregularity map
IM_1	Irregularity map from 1×1 patch
IM_2	Irregularity map from 2×2 patch
IM_4	Irregularity map from 4×4 patch
IM_8	Irregularity map from 8×8 patch
PPV	Positive predictive value (i.e., precision)
PS	predicted segmentation
TPR	True positive rate (i.e., recall)
TP	True positive
VD	Volume Difference
vol	Volume
α	Trainable parameter of Parametric Rectifier Linear Units (PreLU)

α_{IM}	Blending weight of IM_1
β	Bias
β_{IM}	Blending weight of IM_2
β_m	Translation weight of F_m at the m-th FiLM layer
cosh	Hyperbolic cosine
δ_{IM}	Blending weight of IM_8
η	Step of brain abnormalities simulation
exp	Exponential function
γ_{IM}	Blending weight of IM_4
γ_m	Linear transformation map of F_m at the m-th FiLM layer
κ	Kappa value of LST-LGA
λ_1	Weights of the first DEP-GAN's regularisation term
λ_2	Weights of the second DEP-GAN's regularisation term
λ_3	Weights of the third DEP-GAN's regularisation term
$\mathbb{E}$	Expected value (i.e., expectation)
$\mathbb{P}_g$	Distribution of "fake" (generated) images ($\mathbf{x}'$)
$\mathbb{P}_r$	Distribution of "real" images ($\mathbf{x}$)
$\mathbf{s}$	Source patch of LOTS-IM
$\mathbf{t}$	Target patch of LOTS-IM
$\mathbf{x_1'}$	"Fake" follow-up (year-1) image in DEP-GAN
$\mathbf{x}'$	"Fake" image generated by GAN
$\mathbf{x_0}$	Baseline (year-0) image in DEP-GAN
$\mathbf{x_1}$	Follow-up (year-1) image in DEP-GAN

$\mathbf{x}$	"Real" image for the input of GAN
$\mathbf{x}^\top \mathbf{w}$	Matrix multiplication between one-dimensional input vector $\mathbf{x}$ and one-dimensional kernel vector $\mathbf{w}$ in linear transformation operation
$\mathbf{y}'$	"Fake" DEM in DEP-GAN
$\mathbf{y}$	"Real" DEM in DEP-GAN
$\mathbf{z}$	Gaussian noise vector with distribution of $\mathcal{N}(0,1)$
$\mathcal{F}$	1-Lipschitz function for GAN optimisation
σ	Non-linear function
sinh	Hyperbolic sine
tanh	Hyperbolic tangent
$\mathbf{h}$	Vector of DBM's hidden layer
$\mathbf{v}$	Vector of DBM's visible layer
$\mathbf{W}$	Symmetric weight matrix of DBM
ReLU	Rectified Linear Unit non-linear function
sig	Sigmoid non-linear function
Θ	Enclosed DBM's parameters
ξ	Slack variable of SVM
a, b	Ranges of valid values in the two-dimensional convolutional kernel
$BCEL$	Binary cross-entropy loss
$blend$	Voxel value from the $IM_{blended}$
$C(\mathbf{x})$	Critic function for distinguishing "real" DEM ($\mathbf{y}$) and "fake" DEM ($\mathbf{y}'$) in DEP-GAN
d	Distance value between source patch ($\mathbf{s}$) and target patch ($\mathbf{t}$) in LOTS-IM computation

$D(\mathbf{x})$ Critic function for distinguishing "real" follow-up image ($\mathbf{x_1}$) and "fake" follow-up image ($\mathbf{x'_1}$) in DEP-GAN

E Energy function of DBM

F_m Feature map at m-th FiLM layer

f_w Discriminator/critic function of GAN

$FiLM$ Feature-wise Linear Modulation function

$Fl(t)$ Voxel value of T2-FLAIR MRI at time t in brain abnormalities simulation

g_θ Generator function of GAN

h Output to the neuron of neural networks

I Input image of convolution

i, j Indices of fature map (S)

$Irr(t)$ Irregularity value at time t in brain abnormalities simulation

K Convolutional kernel

$M(\mathbf{x})$ Generator function in DEP-GAN

ori Voxel value from the original T2-FLAIR MRI in LOTS-IM computation

pen Penalised voxel value in LOTS-IM computation

q_i Predicted probability of class i

S Feature map, the output of convolution

st Simulated time in brain abnormalities simulation

y_i Binary indicator (0 or 1) if class i is correctly predicted

$Z(\Theta)$ Partition function over all possible configurations of DBM with Θ parameters

MAE mean absolute error

MSE mean square error

abs Absolute function

Acronyms

AAE Adversarial Auto-Encoder. 64

AD Alzheimer's disease. iii, iv, xi, 1, 7, 26, 29, 31, 73, 86, 92, 95, 125

ADNI Alzheimer's Disease Neuroimaging Initiative. 30, 31, 39, 73, 75

ANCOVA analysis of covariance. 47, 60, 93, 104, 105, 118, 119

AnoGAN Anomaly GAN. 64

ARWMC Age-Related White Matter Change. 11, 14

AUC-PR area under Precision-Recall curve. 27, 45, 48–51, 53, 54

BCEL binary cross-entropy loss. 42

BG PVS basal ganglia perivascular spaces. 104, 108, 119

C-GAN Conditional GAN. 23

CAD computer-aided detection and diagnosis. 3, 5, 11, 15–19, 24

CADe computer-aided detection. 16

CADx computer-aided diagnosis. 16

CEN Convolutional Encoder Network. 64, 75, 77, 79, 83, 84, 87, 89, 126

CN cognitive normal. 31

CNN Convolutional Neural Network. ix, x, 3, 19–22, 24–32, 34–62, 125

CNN-GSI CNN with GSI. 29, 30, 40, 43, 48, 51, 54–56, 58, 62

CNN-GSI-xyz CNN with X, Y, and Z GSI. 52–54, 57, 58

CNN-GSI-xyz-rad CNN with X, Y, Z, and radial GSI. 52–54, 57, 58

CPU Central Processing Unit. 59, 65, 70, 76, 77, 79, 80

CSF cerebrospinal fluid. 17, 27, 34, 36, 66, 67, 69, 89, 101

D Disagreement. xiii, 46, 57, 58

DBM Deep Boltzmann Machine. x, xv, xvi, 29, 30, 37, 38, 44, 45, 47–49, 51, 53–55, 57–60, 64, 75, 77, 79, 83, 84, 87, 89, 126, 130

DEM Disease Evolution Map. xv, 24, 92–99, 103, 105–107, 115–118

DEP Disease Evolution Predictor. iv, xi, xii, 2–4, 24, 92, 93, 95–97, 99, 100, 102–124, 126–128, 131

DEP-GAN DEP based on Generative Adversarial Network. xii, xiv–xvi, 24, 93, 95–100, 102, 103, 109–111, 113, 114, 116–123, 126–128

DEP-GAN-1C DEP-GAN with 1 critic. 93, 103, 105–108

DEP-GAN-2C DEP-GAN with 2 critics. 93, 103, 105–109, 121, 122

DEP-UResNet DEP based on U-Residual Network. 24, 93, 95, 97, 99, 100, 102, 103, 109–111, 113–115, 118, 119, 121–123, 126–128

DSC Dice similarity coefficient. xiii, 26–28, 36, 45, 46, 48–55, 57–59, 61, 75–77, 79–87, 98, 103, 104, 106, 111–114, 117, 120, 121

DTI diffusion tensor imaging. 7, 10, 58

DWMH deep WMH. 7, 11–13, 16, 31, 46, 87

EM Expectation–Maximization. 17

FCN Fully Connected Network. 22, 27, 96

FiLM Feature-wise Linear Modulation. xiv, xvi, 97, 99, 100

FN False Negative. xiii, 45, 46

FP False Positive. xiii, 45, 46, 76

GAN Generative Adversarial Network. xii, xiv–xvi, 3, 19, 23, 64, 92, 93, 95, 102, 103, 105–107, 109

GLCM grey-level co-occurrence matrix. 18, 26

GLRLM grey-level run-length matrix. 18, 26

GM grey matter. 17, 27, 34, 36

GPU General Processing Unit. 59, 65, 66, 70, 79, 80, 85, 89, 90

GSI Global Spatial Information. ix, x, 3, 25, 26, 28–30, 32, 34, 36–40, 42, 44, 46–54, 56, 58–62, 75, 77, 79, 86, 87, 125

GT ground truth. xiii, 32, 33, 36, 45, 46, 51, 55–57, 59, 60, 73, 75, 84, 85, 112, 113

IAM Irregularity Age Map. xi, 64, 65, 67, 69, 70, 76, 77, 79, 80, 89

ICV intracranial volume. 66, 69, 101, 105, 108, 113, 119

IM irregularity map. xi, xiii–xv, 3, 24, 64–73, 76–82, 88–90, 93–96, 99, 102, 103, 105–111, 113, 114, 116–123, 125–128

IQR interquartile range. 101

k-NN k-nearest neighbour. 26, 27

LBL binary WMH label. 94, 95, 99

LBL-DEM three-class DEM label. 94, 95, 99, 105, 109, 121

LDA Linear Discriminant Analysis. 26

LoA limit of agreement. 104, 105, 109, 111, 112, 121

LOTS Limited One-time Sampling. 79

LOTS-IM Limited One-time Sampling Irregularity Map. xi, xiv–xvi, 2, 3, 24, 63, 65–67, 70–72, 74–87, 89, 90, 93, 95, 102, 118, 125, 126, 129, 131

LST-LGA Lesion Growth Algorithm from Lesion Segmentation Tool. x, xiv, 17, 30, 35, 36, 47–49, 51–55, 57, 58, 64, 75–77, 79, 81, 82, 84, 85, 87, 89, 126

MAE mean absolute error. xvi, 98

MCI mild cognitive impairment. xi, 7, 27, 31, 73, 86, 92

MICCAI Medical Image Computing and Computer Assisted Intervention. 27, 91

MR magnetic resonance. 13, 17, 35, 39, 40, 43, 58, 89

MRI magnetic resonance imaging. iii, iv, x, xi, xvi, 1–3, 5, 6, 10–13, 16, 18, 24–26, 28–32, 34, 36, 38–40, 44, 46, 47, 49–54, 58, 60, 61, 63–66, 68–75, 82, 84–86, 90–93, 95, 100–103, 122, 125, 127–130

MS multiple sclerosis. 7, 26

MSE mean square error. xvi, 98, 99

MTI magnetization transfer image. 58

MTS Multiple-time Sampling. 69, 70, 80

NAWM normal appearing white matter. 9, 34, 35, 67, 89

OPF Optimum Path Forest. 26

OTS One-time Sampling. 65, 69, 70

OTS-IAM One-time Sampling Irregularity Age Map. xi, 64, 65, 67, 70, 76, 77, 79, 80, 89

PCA principal component analysis. 26, 44

PD positron density. 58

PET positron emission tomography. 31

PM probability map. 24, 64, 71, 72, 77, 78, 94–96, 99, 102, 103, 105–111, 113, 114, 116–123, 126–128

PPV Positive Predictive Value. xiii, 45, 46, 75–77, 80

PreLU Parametric Rectifier Linear Units. xiii, 41

PS predicted segmentation. xiii, 46

PV periventricular. 12, 101

PVWMH periventricular WMH. 7, 11–13, 16, 31, 46, 87

RBF radial basis function. 18, 44

RBM Restricted Boltzmann Machine. 37, 38, 45

ReLU Rectified Linear Unit. 21, 22

ResBlock Residual Block. 97, 99, 100

RF Random Forest. 3, 17, 18, 26, 27, 30, 35, 36, 44, 47–49, 51, 53–55, 57–61, 64, 75, 77, 79, 83, 84, 87, 89, 126

RMSprop Root Mean Square propagation. 43

ROI region of interest. 26, 27, 38, 49

rprop resilient propagation. 43

RPS Rotterdam Progression Scale. 14

SD standard deviation. 31–33, 48, 57, 77, 81, 83–85, 104, 105, 109, 111, 121

SL stroke lesions. 2, 22, 28, 93, 100–103, 105, 110, 116–121, 123

SPC Specificity. 76

SPC Schmidt Progression Scale. 14, 76, 77, 80

SVD small vessel disease. 1, 5–7, 10, 101

SVM Support Vector Machine. xv, 3, 17, 18, 26, 27, 30, 35, 36, 44, 47–49, 51, 53–55, 57–61, 64, 75, 77, 79, 83, 84, 87, 89, 126

T1-W T1-weighted. 22, 26–28, 32–34, 36, 48–51, 61, 101, 128

T2-FLAIR T2 Fluid-Attenuated Inversion Recovery. iii, xvi, 1–3, 6, 22, 24, 26–28, 32–36, 38, 39, 46, 48, 60, 61, 64, 66, 67, 69–74, 83, 88, 89, 92, 93, 95, 99, 101, 115, 116, 123, 125–128

T2-W T2-weighted. 1, 6, 26, 32, 58, 61, 95, 101, 123, 128

TN True Negative. 76

TNR True Negative Rate. 76

TP True Positive. xiii, 45, 46

TPR True Positive Rate. xiii, 45, 46, 75–77, 80

UNet U-Network. 22, 27, 28, 75, 77–79, 84–87, 89, 90, 126

UResNet U-Residual Network. 22, 28, 75, 77–79, 84–87, 89, 90, 92, 93, 95–97, 99, 102, 118, 126

URL Uniform Resource Locator. 33, 75, 130, 131

VA-GAN Visual Attribution GAN. 93, 95, 97, 98, 102, 103, 105–108, 120, 122

VD Volume Difference. xiii, 46, 53, 57, 58

vol volume. xiii, 46, 53, 98

WADRC Wisconsin Alzheimer's Disease Research Centre. 26

WGAN-GP Wasserstein GAN with gradient penalty. 23, 93, 98, 103, 105–108, 122

WM white matter. 17, 36

WMH white matter hyperintensities. iii, iv, ix–xi, 1–3, 5–11, 13–19, 21–36, 38–40, 42–66, 68–114, 116–128

Y1 Year 1. 47, 86

Y2 Year 2. 47, 86

Y3 Year 3. 47, 86

Chapter 1

Introduction

This thesis focuses on the development of machine learning methods for segmentation, characterisation, and prediction of the evolution of white matter hyperintensities (WMH) using T2 Fluid-Attenuated Inversion Recovery (T2-FLAIR) brain magnetic resonance imaging (MRI). This first chapter provides a general motivation for this thesis, its scope, and its contributions. Finally, the structure of this thesis is also described at the end of this chapter.

1.1 Motivation

WMH are neuroradiological features seen in T2-weighted (T2-W) and T2-FLAIR brain MRI. WMH are considered a feature of small vessel disease (SVD) (Wardlaw et al., 2013), partly because in many occasions they have been reported as having vascular origin. Nevertheless, they have been also seen in autoimmune diseases that have effects on the brain (Theodoridou and Settas, 2006), in neurodegenerative diseases (Ge, 2006), and in psychiatric illnesses (Kempton et al., 2008; Videbech, 1997); none of which necessarily encompasses the presence of SVD indicators. Clinically, WMH have been commonly associated with stroke, ageing, dementia, and AD progression (Wardlaw et al., 2013; Prins and Scheltens, 2015). For example, in AD patients, higher load of WMH volume has been associated with higher amyloid beta deposits, presence of markers of SVD, and reduced amyloid beta clearance; all these contributing to an overall worsening of the cognitive functions in these patients (Birdsill et al., 2014).

In early studies, WMH and their severity were presumed to be linearly progressing over time with age (Veldink et al., 1998; Schmidt et al., 2003) due to lack of data with more than one follow-up assessment (Van Leijsen et al., 2017). With increasing

longitudinal data over the years, recent studies have suggested that progression of WMH may be a non-linear process over time (Wardlaw et al., 2017; Van Leijsen et al., 2017). For example, the WMH volume of a patient may grow in the first follow-up assessment, but shrink in the second (or vice versa). Furthermore, in an individual patient, different WMH clusters may simultaneously shrink (i.e., regress), stay unchanged (i.e., stable), or grow (i.e., progress) over a period of time (Ramirez et al., 2016; Chappell et al., 2017). In this thesis, "evolution of WMH" is used to refer theses changes.

The main aim of this thesis is to develop a model for predicting the evolution of WMH from T2-FLAIR brain MRI using a data-driven deep learning method. To achieve this, various machine learning and deep learning methods to automatically quantify WMH through segmentation are first explored. Then, a novel computer graphics-based method for WMH characterisation named Limited One-time Sampling Irregularity Map (LOTS-IM) is proposed to better quantify subtle WMH from T2-FLAIR brain MRI. Lastly, a novel deep learning model to automatically predict the evolution of WMH from brain MRI named Disease Evolution Predictor (DEP) model is proposed. Using DEP, it is hoped that clinicians can estimate the size and location of WMH in time to study their progression/regression in relation to clinical health and disease indicators ultimately to design more effective therapeutic interventions.

1.2 Scope

The scope of this thesis is limited to specific neuroradiological features of WMH, and encompasses the development and analysis of their segmentation, characterisation, and evolution prediction methods in T2-FLAIR brain MRI in brains of individuals with mild-to-moderate vascular pathology, where they are considered a biomarker for progression to dementia and AD. Other brain features also observed in these images (e.g., stroke lesions (SL), perivascular spaces) and pathologies, such as multiple sclerosis, are out of the scope of this thesis.

1.3 Thesis Contributions

The main contributions of this thesis are listed below.

1. Analysing the performances of conventional machine learning and deep learning methods for WMH segmentation in routine clinical brain MRI with none or mild

vascular pathology (Rachmadi et al., 2017b,a).

2. Proposing the use of spatial information in deep learning methods to improve their performance on segmenting early and subtle WMH (Rachmadi et al., 2018b).

3. Proposing a novel unsupervised quantitative method for WMH characterisation and analysis named LOTS-IM (Rachmadi et al., 2017c, 2018c, 2020).

4. Demonstrating the use of irregularity map (IM), a novel modality of brain MRI produced from T2-FLAIR by using LOTS-IM, for simulating the evolution of abnormalities inside the brain (Rachmadi et al., 2018a).

5. Proposing novel deep learning methods to model and predict the evolution of WMH (Rachmadi et al., 2019a,b).

1.4 Structure

The rest of the thesis is organised as follows:

1. **Chapter 2** introduce WMH, their significance for clinical studies, and different means to assess WMH quantitatively. Different computer-aided detection and diagnosis (CAD) methods that have been proposed by previous studies for quantitative assessment of WMH are also briefly discussed in this chapter.

2. **Chapter 3** explains how GSI is important for WMH segmentation, especially when using CNNs. In this chapter, the performance of deep learning methods are compared to the performance of conventional machine learning methods, such as Support Vector Machine (SVM) and Random Forest (RF), for WMH segmentation.

3. **Chapter 4** explains a novel unsupervised method for WMH characterisation, analysis and segmentation named LOTS-IM. In this chapter, the performance of LOTS-IM for WMH segmentation is compared to that of conventional machine learning and deep learning methods. This chapter also demonstrates the use of IM for simulating the evolution of abnormalities inside the brain.

4. **Chapter 5** explains the newly proposed deep learning methods named DEP model to predict the evolution of WMH. In this chapter, an ablation study of GAN based DEP models, different learning approaches, and an ablation study of

auxiliary input modalities for DEP model are discussed and compared to each other.

5. **Chapter 6** summarises all important results in this thesis. Furthermore, conclusions and future work are described.

Chapter 2

White Matter Hyperintensities - Characteristics and Assessment

In this chapter, the background and basic knowledge that underpin this thesis are described. Firstly, white matter hyperintensities (WMH) are introduced, including their clinical significance and their evolution over time. Secondly, an overview of different means to quantitatively assess WMH from brain MRI is described. Lastly, an overview of computer-aided detection and diagnosis (CAD) methods for WMH assessment from brain MRI is discussed.

2.1 White Matter Hyperintensities

In this section, the nature of WMH, their appearance in brain MRI, their significance in clinical studies, and their progression over a period of time are described. This section provides brief explanations for each of the topics mentioned above, mostly from a clinical point of view, to describe the clinical background of this thesis.

2.1.1 WMH and their significance in medical study

WMH, together with lacunar ischaemic strokes, lacunes, cerebral microbleeds, and perivascular spaces, are neuroradiological features or markers of cerebral small vessel disease (SVD) (Wardlaw et al., 2013), partly because in many occasions they have been reported as having vascular origin. Nevertheless, they have been also seen in autoimmune diseases that have effects on the brain (Theodoridou and Settas, 2006), in neurodegenerative diseases (Ge, 2006), and in psychiatric illnesses (Kempton et al.,

2008; Videbech, 1997); none of which necessarily encompasses the presence of small vessel disease indicators. WMH are usually diagnosed using imaging techniques, such as MRI, as it is difficult to visualise them in vivo (Shi and Wardlaw, 2016). WMH appear as brighter (i.e., hyperintense) region in T2-weighted (T2-W) and T2 Fluid-Attenuated Inversion Recovery (T2-FLAIR) of MRI. The appearance of WMH in T2-FLAIR and the general schematic of their ill-posed boundary (i.e., the boundary between WMH and non-WMH is not clear cut) can be seen in Figure 2.1.

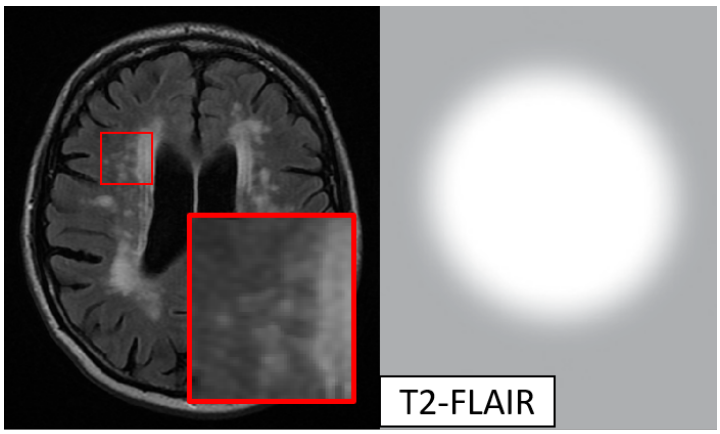


Figure 2.1: Appearance of WMH in T2-FLAIR and the general schematic of their ill-posed boundary. This figure is modified from (Wardlaw et al., 2013).

The underlying pathology of WMH mostly reflects demyelination and axonal loss as a consequence of chronic ischaemia caused by cerebral SVD (Prins and Scheltens, 2015). Clinically, WMH have been commonly associated with stroke and dementia progression (Wardlaw et al., 2013; Prins and Scheltens, 2015; Wardlaw et al., 2015). Between 1990 and 2010, about 15 million people had a stroke and 35.6 million were estimated to be living with dementia worldwide (Lozano et al., 2012). The prevalence of WMH increases with increasing vascular risk factors such as hypertension, diabetes, and smoking (Wardlaw et al., 2015). However, WMH are often found on MRI in virtually every individual over 60 years old with highly variable degree of WMH volume load (de Leeuw et al., 2001; van Leijsen et al., 2017).

WMH are well associated with poor clinical outcome such as increasing risk of admission to a nursing home, stroke, and mortality (Debette and Markus, 2010; van der Holst et al., 2016; Schmidt et al., 2016; van Leijsen et al., 2017). Furthermore, the association between WMH and cognitive decline or dementia has also been well established (Schmidt et al., 2005; Van Dijk et al., 2008; Debette and Markus, 2010; Prins and Scheltens, 2015). More importantly, there have been studies showing that

greater WMH volume loads at baseline increased the likelihood of progression from normal aging to mild cognitive impairment (MCI) (Smith et al., 2008) and progression from amnestic MCI to Alzheimer's disease (AD) (van Straaten et al., 2008). All these previous studies highlight the clinical importance of WMH.

In the human brain, the rate of growth of WMH has been strongly correlated with regional grey matter atrophy, which contributes to the secondary reductions in global brain volume (Lambert et al., 2016). However, the clinical effect of WMH depends on their location in the brain. It has been reported that periventricular WMH (PVWMH) are more closely associated with cognitive decline and brain atrophy than the deep WMH (DWMH) (Huang et al., 2018). PVWMH were also reported to increase the likelihood of progression from amnestic MCI to dementia and AD (van Straaten et al., 2008). Regional WMH analyses revealed significant differences in WMH across regions that also differed significantly by diagnosis (Yoshita et al., 2006). Furthermore, WMH also may affect the white matter tracts of the brain. In a recent study using diffusion tensor imaging (DTI), it was reported that WMH are associated with two patterns of transformed diffusion characteristics in the surrounding white matter tract network while the diffusion characteristics along white matter tracts improve further away from WMH along its penumbra (Reginold et al., 2018). Different types of WMH also have different nature and effect to the brain. For example, punctate WMH are well known to be not ischaemic, not progressive, and thus benign. On the other hand, early confluent and confluent lesions are ischaemic, progressive, and thus malignant (Schmidt et al., 2003). Clinical studies are consistent with this categorisation which also showed that WMH progression cannot be considered benign (Longstreth Jr et al., 2005; Schmidt et al., 2005).

WMH may also be used as a surrogate marker for other clinical purposes. For example, it has been suggested that WMH may serve as a marker for the progression of SVD (Sachdev et al., 2007). Many cross-sectional and longitudinal studies have also provided strong evidence that WMH are clinically important markers of increased risk of stroke, dementia, depression, impaired gait, mobility, and death (Wardlaw et al., 2015). Another proof-of-concept trial study also proposed the use of confluent WMH, which show fast progression and has high correlation with cognitive decline, as a surrogate marker to show treatment effects on lesion progression (Schmidt et al., 2016). In fact, neuroimaging has been proposed as a way to achieve surrogate markers to assess treatment effects in SVD since an earlier study (Pantoni, 2010). A similar concept and model has also been suggested for other white matter diseases such as multiple sclerosis

(MS) (Schmidt et al., 2004).

2.1.2 Evolution of WMH over time

In early studies, the WMH and their severity were presumed to be linearly progressing over time with age (Veldink et al., 1998; Schmidt et al., 2003), this was found to be due to the lack of data with more than one follow-up assessment (Van Leijsen et al., 2017). With increasing longitudinal data over the years, recent studies have suggested that the evolution of WMH may be a non-linear process over time (Wardlaw et al., 2017; Van Leijsen et al., 2017) and have a dynamic behaviour in each patient (Ramirez et al., 2016). For example, WMH volume may grow in the first follow-up assessment and shrink in the second follow-up assessment or vice versa (see Figure 2.2 for example) (Van Leijsen et al., 2017). This is different to most of the longitudinal studies dated from more than a decade ago in which only one follow-up assessment was used (Veldink et al., 1998; Schmidt et al., 2003). Furthermore, more recent studies have also reported that different clusters of WMH in a patient may simultaneously either grow, shrink, or remain stable in the same follow-up assessment (Ramirez et al., 2016; Jiaerken et al., 2019; van Leijsen et al., 2018).

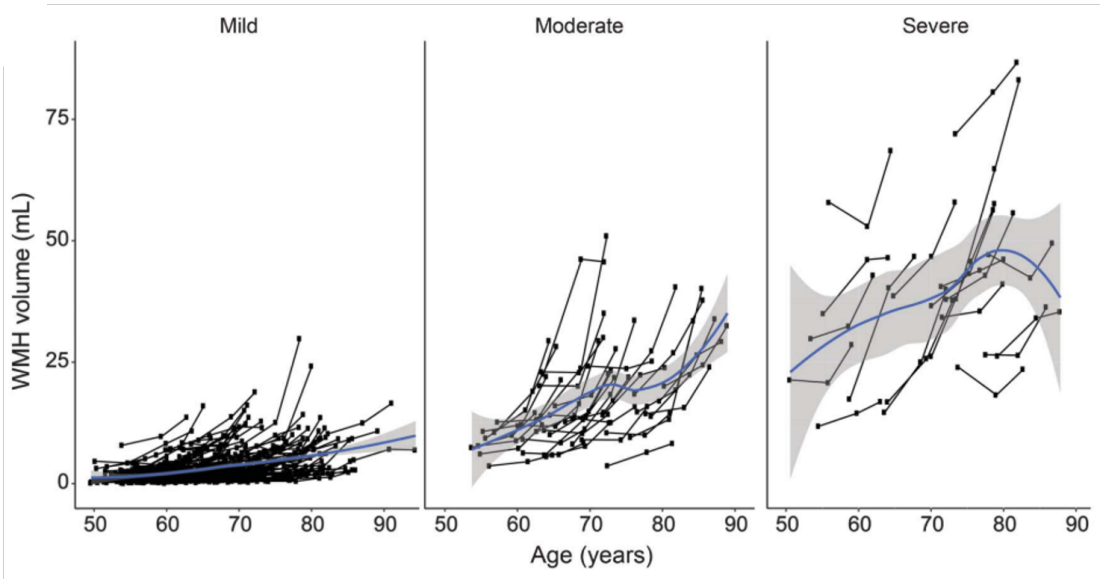


Figure 2.2: Temporal dynamic change of WMH over three time points by age at individual level classified by baseline WMH severity using Fazekas visual rating scale (mild: 0-1, moderate: 2, and severe: 3). See Section 2.2.1 for explanation of Fazekas visual rating scale. This figure is modified from (Van Leijsen et al., 2017).

Extensive studies on longitudinal data have shown that the progress of prevalence,

volume, and severity of WMH over time vary (Veldink et al., 1998; Schmidt et al., 1999, 2003, 2005; Pantoni, 2010). The rate of WMH progression itself varies considerably across the studies (Schmidt et al., 2016; van Leijsen et al., 2017). Some of the most common risk factors and predictors associated with WMH progression are baseline WMH volume (Schmidt et al., 1999, 2002b,a, 2003; Sachdev et al., 2007; Van Dijk et al., 2008; Wardlaw et al., 2017; Chappell et al., 2017), high blood pressure (i.e., hypertension) (Veldink et al., 1998; Schmidt et al., 1999, 2002b; Van Dijk et al., 2008; Godin et al., 2011; Verhaaren et al., 2013), age (Van Dijk et al., 2008), current smoking status (Power C et al., 2015), previous stroke and diabetes (Gouw et al., 2008a), and genetic properties (Schmidt et al., 2000, 2002a, 2011; Godin et al., 2009; Luo et al., 2017). Furthermore, the surrounding region of WMH, which appears like normal appearing white matter (NAWM) with less structural integrity, usually called the "penumbra of WMH" (Maillard et al., 2011), has been reported as having a high risk of becoming WMH over time (Maillard et al., 2014; Pasi et al., 2016). Nevertheless, baseline WMH volume is the strongest predictor and risk factor of WMH progression (Wardlaw et al., 2015).

In the early longitudinal studies of WMH, reduction (i.e., regression) of WMH volume was only observed in a small number of patients (Schmidt et al., 2003, 2005; Sachdev et al., 2007; Gouw et al., 2008b; Maillard et al., 2009; Prins et al., 2004; Rovira Cañellas et al., 2007). Because of that, most earlier studies regarded the regression of WMH as a measurement error (Sachdev et al., 2007; Maillard et al., 2009; Schmidt et al., 2003, 2005) or "no progression" with no further explanation (Prins et al., 2004; Van Dijk et al., 2008; Gouw et al., 2008b). Furthermore, the bias in manual delineation of WMH towards progression when the raters are aware of the scans' time sequence cannot be overlooked (Schmidt et al., 1999, 2005). It is worth to mention that Sachdev et al. (2007) did investigate the possibility of WMH regression in some patients but did not find any significant association to the evaluated risk factors, including the strongest risk factor, baseline WMH volume. One possible explanation of this is that WMH regression could be missed when using two neuroimaging assessments with a long interval where WMH decline within a certain time window is compensated by WMH progression thereafter in a cohort that, on average, showed progression (van Leijsen et al., 2017). Thus, it is important to take into account the time window of assessment when performing longitudinal study of WMH. On the other hand, recent studies have reported the regression of WMH in several radiological observations, especially after some clinical conditions or interventions. For example, WMH regression

was observed on MRI after cerebral infraction (Moriya et al., 2009), strokes (either minor, lacunar, or ischaemic) (Durand-Birchenall et al., 2012; Cho et al., 2015; Wardlaw et al., 2017), improved hepatic encephalopathy due to treatment (Mínguez et al., 2007), lower blood pressure due to treatment (Wardlaw et al., 2017), liver transplantation (Rovira Cañellas et al., 2007), and carotid artery stenting (Yamada et al., 2010).

Many aspects of WMH are still not yet clear given current results from longitudinal studies, especially the regression of WMH. One study suggested that WMH should not be viewed only as "untreatable" or "permanent" because in vivo imaging indicates that water shifts and water content are prominent and could be used to representing and detecting early changes in WMH (Wardlaw et al., 2015). Note that MRI is known to rely on natural properties of the hydrogen molecules that form part of fluids (i.e., water) or lipids, and WMH are water-based tissues. There is also strong evidence that novel imaging techniques, such as DTI, can detect subtle impairments in white matter tract integrity before they can be seen on conventional MRI (Prins and Scheltens, 2015). These findings suggest that WMH might represent only the extreme end of a continuous spectrum of white matter injury, i.e., the WMH are probably only the "tip of the iceberg" (Zhang et al., 2013; Lockhart et al., 2012; Wardlaw et al., 2015). Other studies have shown that there was a strong association of the deteriorating microstructural integrity observed in DTI with WMH progression (Jiaerken et al., 2019; van Leijsen et al., 2018). Jiaerken et al. (2019) reported that growing WMH had significantly lower mean diffusivity and higher fractional anisotropy of DTI compared to constant WMH. Interestingly, there was no significant difference of either metabolism or micro-structure between shrinking WMH and constant WMH regions, either before or after the regression from shrinking WMH to normal white matter. This finding suggests that regions of shrinking WMH which appear to be normal white matter are actually still damaged (Jiaerken et al., 2019). However, a most recent study showed that SVD regression, including WMH regression, did not accompany brain atrophy, which suggests that WMH regression follows a relatively benign clinical course (van Leijsen et al., 2019). Therefore, there might be a possibility to detect WMH at an early stage, predict WMH evolution (i.e., growth and shrinkage), and hold back WMH progression by using cutting-edge imaging technologies in the future.

2.2 Quantitative Assessment of WMH

In this section, different means to quantitatively assess WMH are described. This section is divided into two subsections, which are about cross-sectional assessment of WMH and longitudinal assessment of WMH.

2.2.1 Cross-sectional assessment of WMH

Cross-sectional assessment of WMH refers to the assessment of WMH at one time point, independent of any previous or follow-up assessment. Thus, cross-sectional assessment is usually performed in most cases to observe WMH on a patient at a specific time point. Depending on the methods, cross-sectional assessment of WMH usually produces either location, volume (load), severity level, type, or all of them at the end of the assessment. Clinically, it is often challenging to assess the extent of the WMH contribution to the patient's cognitive level (Prins and Scheltens, 2015). However, some studies have shown that total volume and location of WMH are important determinants for clinical studies (Biesbroek et al., 2013; Yoshita et al., 2006). Location and volume of WMH can be manually produced by clinicians by delineating regions indicated as WMH or automatically produced using CAD intelligent systems. Generally speaking, however, manual delineation of WMH is not widely applicable and can be time-consuming (Gouw et al., 2008b).

The severity of WMH can also be assessed using *visual rating scales*. Visual rating refers to an assessment done by radiologists by looking at the MRI scan and rating the severity of WMH. Some examples of visual rating scales are Fazekas (Fazekas et al., 1987), Scheltens (Scheltens et al., 1993), Longstreth (Longstreth et al., 1996), and Age-Related White Matter Change (ARWMC) (Wahlund et al., 2001). Visual ratings are widely used clinically for describing severity of white matter disease (Scheltens et al., 1993) especially before the wide use of CAD. Note that assessment of WMH using a visual rating scale is faster and more applicable than manually delineating all of WMH in a patient. Nevertheless, studies have shown that WMH volume and WMH clinical scores are very highly correlated (Valdés Hernández et al., 2013). A widely applied visual rating scale, which is used in the validation of the computational methods developed throughout this PhD, are Fazekas and Longstreth visual rating scales.

Fazekas visual rating subdivides WMH based on their location in relation to the brain ventricles, namely PVWMH and DWMH, and rates each "subtype" according to the size and confluence. PVWMH's ratings of Fazekas are:

1. **PVWMH Fazekas 0**: Absent.
2. **PVWMH Fazekas 1**: “Caps” or pencil-thin lining around ventricle.
3. **PVWMH Fazekas 2**: Smooth “halo”.
4. **PVWMH Fazekas 3**: Irregular periventricular (PV) signal extending into the deep white matter.

Whereas, DWMH’s ratings of Fazekas are:

1. **DWMH Fazekas 0**: Absent.
2. **DWMH Fazekas 1**: Punctate foci.
3. **DWMH Fazekas 2**: Beginning confluence.
4. **DWMH Fazekas 3**: Large confluent areas.

On the other hand, Longstreth grades one slice of MRI scan at the level of the body of the lateral ventricles, without distinguishing between PVWMH and DWMH, from 0 to 8 grades. The Longstreth’s grades are shown on list below.

1. **Longstreth 0**: Absent.
2. **Longstreth 1**: Discontinuous PV rim with minimal dots of subcortical disease.
3. **Longstreth 2**: Thin continuous PV rim with a few patches of subcortical disease.
4. **Longstreth 3**: Thicker continuous PV rim with scattered patches of subcortical disease.
5. **Longstreth 4**: More irregular PV rim with mild subcortical disease; may have minimal confluent PV hyperintensities.
6. **Longstreth 5**: Mild PV confluence surrounding the frontal and occipital horns.
7. **Longstreth 6**: Moderate PV confluence surrounding the frontal and occipital horns.
8. **Longstreth 7**: PV confluence with moderate involvement of the centrum semiovale.
9. **Longstreth 8**: PV confluence involving most of the centrum semiovale.

The illustration of Fazekas and Longstreth visual rating scales on brain MRI is depicted in Figure 2.3.

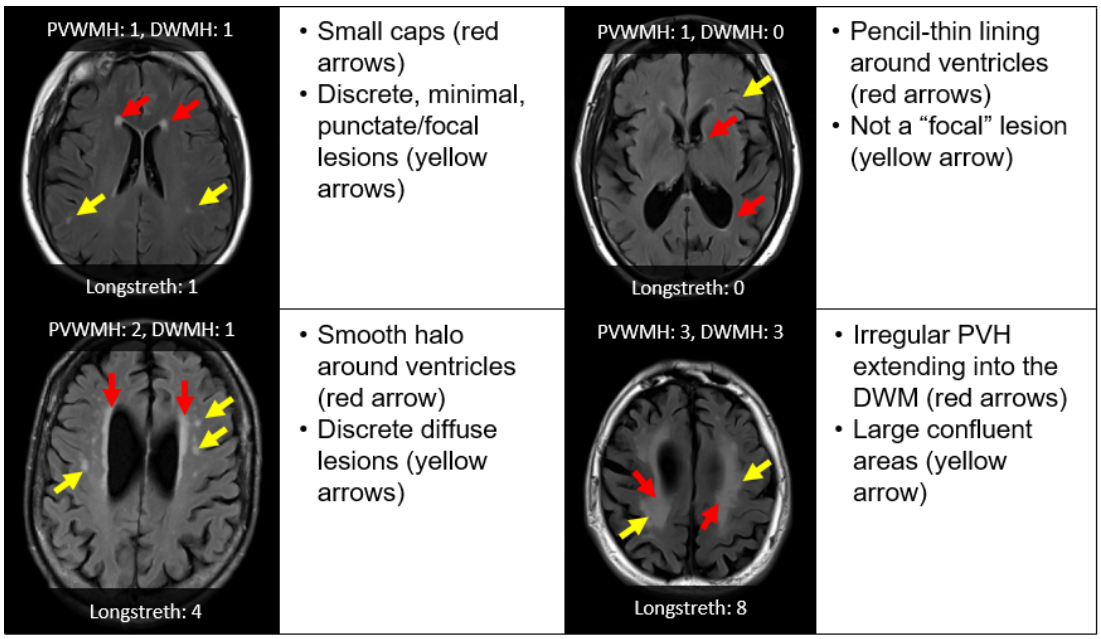


Figure 2.3: Illustration of Fazekas and Longstreth visual rating scales. Fazekas scores are grouped into two, periventricular and deep white matter hyperintensities (i.e., PVWMH and DWMH respectively). In the figure, they are shown as red and yellow arrows respectively. On the other hand, Longstreth evaluates one slice of MRI without distinguishing between PVWMH and DWMH, from 0 to 8 grades.

2.2.2 Longitudinal assessment of WMH

Longitudinal assessment of WMH refers to multiple assessments of brain magnetic resonance (MR) images over time to know how WMH change (i.e., quantification of WMH change/evolution). Thus, the current assessment is dependent from the previous assessments. The most common approach to present the evolution of WMH is using volumetric changes between two or more MRI assessments (i.e., longitudinal global assessments over period of time) (Schmidt et al., 2012b; Van Leijsen et al., 2017). However, it is worth to mention that some early studies used visual ratings of MRI lesions to describe the progression of WMH by their severity (Veldink et al., 1998; Schmidt et al., 2003). Clinically, a longitudinal study of WMH is important to determine the natural course of WMH and may be used to study the effect of clinical interventions (Prins et al., 2004).

Earlier longitudinal studies of WMH used visual ratings of MRI lesions as it was less time consuming than manually delineating all WMH in longitudinal data. For example, Veldink et al. (1998) used an adapted version of Schelten's scale (Scheltens et al., 1993) where a linear scale ranges from 0 to 4, depending on both size and number of lesions

in each brain's regions (i.e., frontal, temporal, parietal, and occipital), and summed up together for a total score of WMH visual score. However, visual rating scales are designed for cross-sectional assessment of WMH and have been indicated as not suited for measuring change in WMH severity (Prins et al., 2004). One of the reasons is due to a ceiling effect: a baseline scan that has the highest rating cannot be properly measured in the follow-up assessment if the volume of WMH increases. Thus, a different study used both volumetric changes and Fazekas visual rating scores (Fazekas et al., 1987) to measure the WMH changes to get more reliable results (Chappell et al., 2017).

Because the visual rating scores used for cross-sectional assessment are not suitable for measuring WMH changes, several attempts have been made to develop visual rating scales to measure changes in WMH such as the Schmidt Progression Scale (SPC) (Schmidt et al., 1999) and the Rotterdam Progression Scale (RPS) (Prins et al., 2004). The SPC measures WMH changes as categories reflecting the number of WMH, that is 0, 1 to 4, 5 to 9, and more than 9 lesions. On the other hand, using RPS, change in WMH is scored by three grades (i.e., -1 for decrease, 0 for no change, and +1 for increase) in three periventricular locations (i.e., frontal caps, lateral bands, and occipital caps) and in four subcortical locations (i.e., frontal, parieatal, temporal, and occipital) resulting in a total scale of -7 to +7. In a follow-up study, Gouw et al. concluded that dedicated progression scales of SPC and RPS are more sensitive, reliable, and correlate better with volumetric changes than cross-sectional visual rating scales of Fazekas, Scheltens, and ARWMC visual rating scales (Gouw et al., 2008b).

In recent years, several studies have proposed the use of spatial dynamic change of WMH as a complementary metric to the volumetric change of WMH. Spatial dynamic change of WMH are performed by separating WMH into three categories, which are growing WMH, shrinking WMH, and stable WMH (Ramirez et al., 2016; Jiaerken et al., 2019; van Leijsen et al., 2018). WMH are labelled as growing if they are absent at baseline but present at the follow-up, shrinking if WMH are present at baseline but absent at the follow-up, and stable if WMH are present at both baseline and follow-up assessments. The illustration of these categories can be seen in Figure 2.4. Using these categories, the evolution of WMH is not only focused on the size of WMH in a patient but also on the position of the changes. These previous studies suggested that the progression of WMH may be more dynamic than previously thought (Ramirez et al., 2016) and followed by dynamic changes in microstructural and metabolism in WMH (Jiaerken et al., 2019).

Nevertheless, it has to be mentioned that there is no gold standard for the assessment

of WMH changes. Volumetric change may provide the most objective assessment method, but it cannot be interpreted as gold standard (Prins et al., 2004). The reason is because the volume of WMH itself is an estimation of the real WMH volume, regardless of manual or automated (computer-aided) assessment. Similarly, spatial dynamic change of WMH are also subject to the expertise of the raters. However, because it is difficult to assess WMH in noninvasive manner (Schmidt et al., 2004), the assessment using medical images is still the preferred course of quantitative assessment of WMH change.

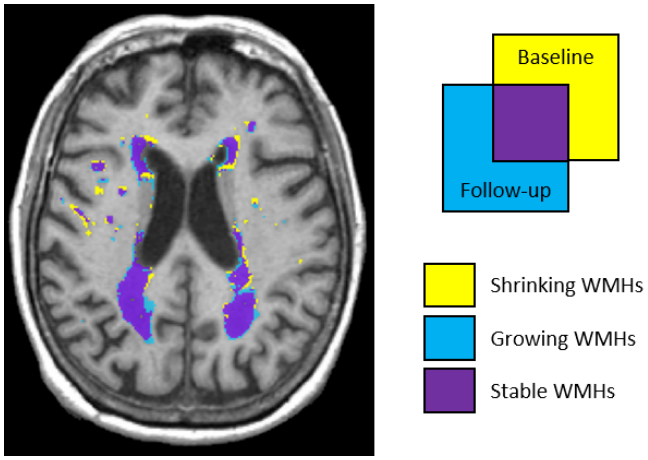


Figure 2.4: Visualisation of dynamic change of WMH categorised into three categories; growing WMH, shrinking WMH, and stable WMH. This figure is modified from (Ramirez et al., 2016).

2.3 Computer-Aided Detection and Diagnosis for WMH

In recent years, CAD system has become mainstream in radiology and clinical work. A CAD system is a class of computer systems that aim to assist in the detection and/or diagnosis of diseases through a “second opinion” (Doi, 2007; Suzuki, 2012; Shiraishi et al., 2011). The goal of CAD systems is to improve the accuracy of radiologists by decreasing false negatives, usually due to observational oversights (Castellino, 2005; Doi, 2007), with a reduction of time in the interpretation of images (Firmino et al., 2016). With the increasing number of accuracy and reliability of CAD results due to the rapid development of artificial intelligence and deep learning, CAD has been commonly used in routine clinical use (Shiraishi et al., 2011) and proposed to perform independent diagnosis in recent years (Litjens et al., 2017).

In general, CAD systems can be classified into two groups: computer-aided detection (CADe) and computer-aided diagnosis (CADx) systems (Firmino et al., 2016). CADe are systems geared for the location of lesions in medical images whereas CADx systems perform the characterisation of the lesions. For example, CADe is designed and proposed for segmenting breast cancer (Dheeba et al., 2014) while CADx is used for differentiating benign and malignant lesions in breast MRI (Newell et al., 2010). In WMH case, CADe systems is used for WMH segmentation using well-known machine learning algorithms (Klöppel et al., 2011) whereas CADx is used for WMH characterisation based on etiology (i.e., demyelinating WMH and ischemic WMH (Leite et al., 2015)), anatomical mapping (i.e., PVWMH and DWMH (DeCarli et al., 2005)), blood flow (Promjunyakul et al., 2015), potential growth (Gwo et al., 2019), or other WMH characteristics.

In this section, the basics of the machine learning techniques commonly used nowadays for CAD of WMH is explained. Machine learning algorithms for assessment of WMH proposed in previous studies are also presented. All machine learning algorithms discussed in this section are divided into two groups, which are conventional machine learning algorithms and deep learning algorithms.

2.3.1 Conventional machine learning algorithm

Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated features (Goodfellow et al., 2016). Most machine learning algorithms can be broadly categorised into *unsupervised* learning and *supervised* learning. These categories are based on how the machine learning system should observe a dataset. Unsupervised machine learning algorithms observe a dataset containing features and learn useful properties of the dataset (e.g., distribution of the features) without corresponding labels of the data (i.e., unlabelled data). Unsupervised machine learning algorithms is usually done by performing clustering (grouping) which groups unlabelled data in such a way that objects in the same group are more similar to each other than the other objects in different groups. On the other hand, supervised machine learning algorithms observe a dataset containing features and associated labels or targets (i.e., labelled data). Thus, the supervised machine learning algorithms learn a function that maps an input data to the associated output label from a set of paired input-output training dataset in a training process.

2.3.1.1 Unsupervised learning algorithms

Several unsupervised CAD methods for WMH have been proposed in the past few years, and most of them perform a kind of clustering based on a map (atlas) of the brain or intensity distributions. Some examples of unsupervised methods for WMH segmentation are Lesion-TOADS (Shiee et al., 2010) and Lesion Growth Algorithm from Lesion Segmentation Tool (LST-LGA) (Schmidt et al., 2012a). Lesion-TOADS uses atlas of the healthy brain to produce a belief map of outliers or irregular intensities (i.e., WMH). To perform the segmentation of WMH, the input MR images need to be registered to the atlas so that outliers can be detected based on the topology of human brain. On the other hand, LST-LGA creates intensity distributions of three classes (i.e., grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF)), and a voxel is deemed as WMH if its intensity is located outside the distribution of these three classes. The results is then refined by using lesion growth model to include more subtle WMH in the neighbourhood of initial WMH. See Section 3.3 for further explanation of LST-LGA. Other unsupervised methods that have been proposed for WMH segmentation are fuzzy C-means methods (Gibson et al., 2010), Expectation–Maximization (EM) based algorithms (Dugas-Phocion et al., 2004; Forbes et al., 2010; Kikinis et al., 1999), and Gaussian mixture models (Freifeld et al., 2009; Khayati et al., 2008).

2.3.1.2 Supervised learning algorithms

The most common supervised machine learning algorithms used for WMH segmentation are Support Vector Machine (SVM) and Random Forest (RF). SVM is a supervised machine learning algorithm that separates data points projected to a high-dimensional feature space by using a hyperplane (Cortes and Vapnik, 1995). SVM is a popular supervised (conventional) machine learning algorithm for classification especially when there are only two classes involved. SVM is optimised by maximising its margin, which is the smallest distance between the hyperplane and the closest samples from each class. These two closest samples from the separating hyperplane are usually called *support vectors*. SVM has a property that corresponds to a convex optimisation problem in its model determination, which is important to get the optimum hyperplane parameters (Bishop, 2006). The SVM can be modelled as Equation (2.1) below

$$\min \frac{1}{2} \|\mathbf{w}\|^2 \text{ subject to } r_t(\mathbf{x}_t^T \mathbf{w} + \beta) \geq +1, \forall t \in T \tag{2.1}$$

where t is a sample from dataset T, $\mathbf{x}_t$ is a feature vector of sample-t, r_t is the label of sample-t where it has value either +1 or -1 to describe underlying classes for each data,

β is a bias, and $\mathbf{w}$ are the parameters of the hyperplane that will maximise the margin between the support vectors and produce the optimal separating hyperplane.

An important aspect of SVM is that it can be modified for handling more complex dataset such as non-separable and non-linear datasets (Lyu and Farid, 2003). In the non-separable dataset, some samples are either located not far enough from the hyperplane or on the wrong side of the hyperplane (i.e., misclassified). In this case, SVM with *soft margin hyperplane*, where a *slack variables* of ξ_t is employed, can be used. Slack variable refers to the deviation of SVM's margin, and it is defined as $\xi_t \geq 0$ where $\xi_t = 0$ if data x_t is correctly classified, $0 < \xi_t < 1$ if data x_t is correctly classified but it is not far enough from the hyperplane, and $\xi_t \geq 1$ if data x_t is misclassified. In the non-linear dataset, transformation to a new space by using a non-linear transformation is needed to solve the problem linearly in the new space. One of the most commonly used transformation function for SVM is radial basis function (RBF) (Alpaydin, 2010).

RF is a collection of decision trees trained individually to produce one combined output (Tin Kam Ho, 1995; Opitz and Maclin, 1999; Criminisi and Shotton, 2013). Collection of RF's trees are created by using bootstrap sample data where a few sets of small training data are used to train the trees independently. Bootstrap sample data is generated by creating m sets of sample data in which every of them contains of n' samples from a training dataset with n samples. Sampling is performed using a uniform random generator for all training samples with replacement (i.e., a sample can be selected multiple times). Some advantages of using bootstrap sample data are improving accuracy and stability, avoiding overfitting, and reducing variance. In addition to using bootstrap sample data, RF also uses unique method to construct a tree where best splits are performed based on a set of randomly chosen features at each node. In other words, each three of RF will be constructed based on different best splits of features. It is said that these approaches performed better than any other supervised conventional machine learning algorithms such as SVM.

Unlike unsupervised CAD algorithms where most of them use MRI's intensity as input, supervised CAD algorithms usually use hand-designed features as the input. Some feature extraction methods that have been proposed for CAD WMH are statistical histogram analysis, grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (GLRLM) (Leite et al., 2015), Gabor filters (Klöppel et al., 2011), and texton-based features that consist of low-pass, high-pass, band-pass, and edge filters (Ithapu et al., 2014). Detailed explanation of these previous studies mentioned above can be found in Section 3.1.1.

2.3.2 Deep learning algorithm

Deep learning is a kind of machine learning which learn the representation of the data (i.e., features) automatically without using any feature extraction methods. The most crucial step in machine learning, including in a CAD system, is the representation of data. In conventional machine learning, various features (or data representation) are defined by a designer after inspecting the data (i.e., hand-designed features). If the data representation is unsuitable for the objective task (e.g., classification or segmentation), the machine learning algorithm will struggle to find the optimal solution of the objective task. On the other hand, a deep learning model learns not only the association between representation and output but also the best possible representation of the data. To do so, the deep learning model relies on one important principle called *hierarchical feature representation* where multiple hidden layers are used to learn different levels of the data representation. In the hierarchical feature representation, shallower hidden layers learn low-level features (e.g., edges) while deeper hidden layers learn high-level features (e.g., context of the image). Because of the reasons described above, deep learning can be categorised as a kind of representation learning (Goodfellow et al., 2016). Flowcharts showing how conventional machine learning and deep learning differs can be seen in Figure 2.5.

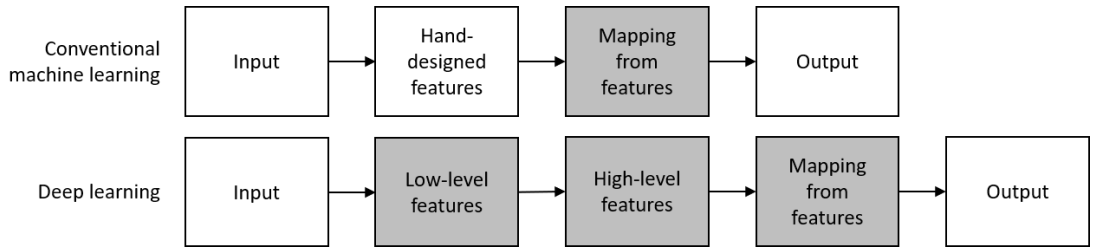


Figure 2.5: Flowcharts showing how conventional machine learning and deep learning differs in learning process. Learning processes are performed in the shaded boxes. This figure is modified from (Goodfellow et al., 2016).

Similar to conventional machine learning, deep learning algorithms can be generally divided into supervised deep learning algorithms and unsupervised deep learning algorithms. In this subsection, the most common examples of supervised and unsupervised deep learning models, named CNNs and GANs, are described. Some previous studies of CAD systems using deep learning for the assessment of WMH are also briefly introduced.

2.3.2.1 Convolutional Neural Networks

CNNs (LeCun et al., 1989), also known as convolutional networks, are a specific kind of neural network that is suitable for grid-like topology data such as images (Goodfellow et al., 2016). CNNs rely heavily on convolution, a mathematical linear operation. Typically, CNNs consist of convolutions, non-linear operations, and pooling operations stacked together as convolutional layer. Depiction of convolutional layer can be seen in Figure 2.6.

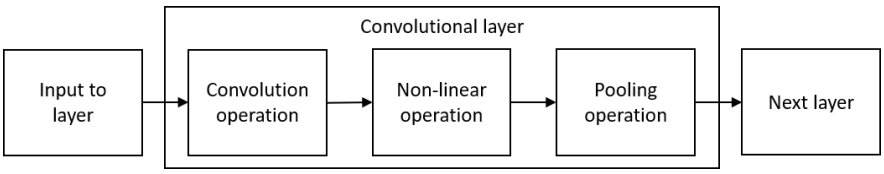


Figure 2.6: Illustration of convolutional layer which is formed by convolution, non-linear operation, and pooling operation. This figure is modified from (Goodfellow et al., 2016).

Convolution refers to an operation on two functions of a real valued argument (Goodfellow et al., 2016). With two-dimensional numerical images, convolution is defined as:

$$S(i,j) = (I * K)(i,j) = \sum_{m=-a}^{a} \sum_{n=-b}^{b} I(i-m, j-n)K(m,n) \tag{2.2}$$

where I is the input image, K is the convolutional kernel, $*$ is the notation for convolution, and S is the output called *feature map*. As for the indices, i and j are the two-dimensional indices of the output S while a and b are the ranges of valid values in the two-dimensional kernel. Convolution leverages three important ideas that improve learning capability of machine learning system, which are sparse interactions, parameter sharing, and equivariant representation (Goodfellow et al., 2016).

Sparse connectivity refers to limited interaction between output and input units in local space. In conventional neural networks, all input units interact with each output unit using large matrix multiplication. In CNNs, on the other hand, only small numbers of inputs unit interact with output unit using convolution. The advantages of using sparse connectivity are fewer parameters, better statistical efficiency, and lower computational costs.

Parameter sharing refers to using the same parameter for more than one output unit. In conventional neural networks, a connection (weight) between an input unit and an output unit is used only once (i.e., tied to specific input and output). In CNNs, a weight

between an input unit and output unit is used and shared by other input and output units. This implies that a set of weights or feature detector can be used in different locations instead of at a specific location.

Equivariant representation to translation means that convolution produces the same result even if the image is transformed by a shift operation. This implies that a feature still can be detected even if it is moved to a different location. This property is tied to parameter sharing property of the convolution. It should be noted that convolution is naturally not equivariant to other transformation such as rotation and scaling.

Non-linear operation is inherited from conventional neural networks where it transforms the output of linear operation, such as convolution, using a non-linear function. In neural networks, the non-linear function is often called *activation function* because it restricts values that can activate the output unit. Non-linear functions such as sigmoid (Equation (2.3)) and tanh (Equation (2.4)) are commonly used in conventional neural networks, but they suffer from a problem called "vanishing gradient" where gradients in the shallower layers vanish (Hochreiter, 1998). In deep learning, more effective non-linear functions, such as Rectified Linear Unit (ReLU) (Equation (2.5)), are commonly used in hidden layers because they do not suffer from the vanishing gradient problem. It should be noted that sigmoid and tanh are still used in deep learning but restricted to the last non-linear layer (i.e., final output) to produce real values from 0 to 1 (i.e., probability-like values) and real values from -1 to 1 respectively. Depictions of sigmoid, tanh, and ReLU can be seen in Figure 2.7.

$$\text{sig}(x) = \frac{1}{1+\exp(-x)} \tag{2.3}$$

$$\tanh(x) = \frac{\sinh x}{\cosh x} \tag{2.4}$$

$$\text{ReLU}(x) = \max(0,x) \tag{2.5}$$

Pooling is another important concept used in CNNs, where it replaces the output of previous operations with a summary statistic of a rectangular neighbourhood outputs. For example, max pooling summarises a neighbourhood with the largest value in the neighbourhood. Figure 2.8 provides an illustration for max pooling in two-dimensional data. Pooling introduces a more compact representation that is approximately invariant to small translations.

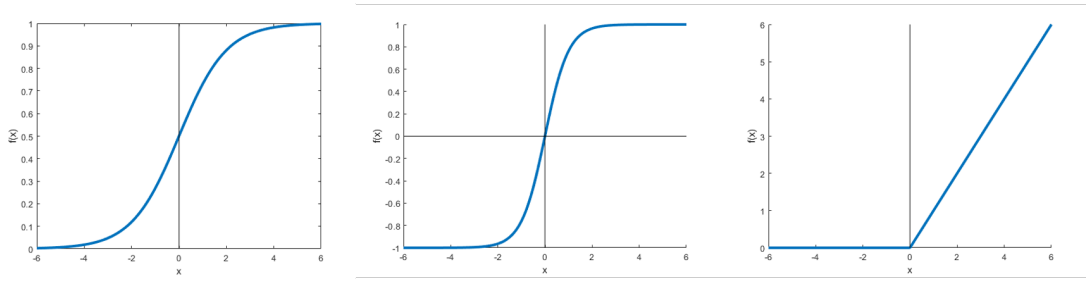


Figure 2.7: Illustrations of non-linear functions ($f(x)$) of sigmoid, tanh, and ReLU respectively from left to right. These illustrations are modified from (Ghafoorian, 2018).

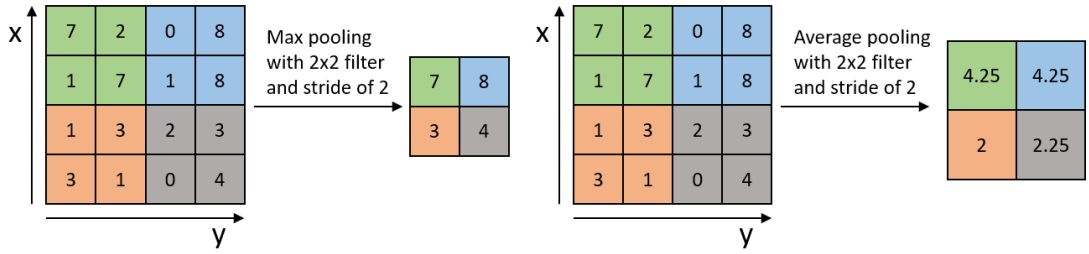


Figure 2.8: Illustrations of maximum pooling (left) and average pooling (right) operations on two-dimensional data. These illustrations are modified from (Ghafoorian, 2018).

In the past few years, several architectures of deep learning have been proposed for WMH segmentation. For example, one study proposed the use of parallel convolutional layers with different size of input patches and additional hand-designed spatial features to provide location information for the CNNs (Ghafoorian et al., 2017b). Moeskops et al. (2018) proposed a multi-scale CNN with different resolution images of a T1-weighted (T1-W), a T2-FLAIR, and a T1-W inversion recovery image as input for automatically segmenting WMH and other brain regions (e.g., cortical grey matter and cerebrospinal fluid). Guerrero et al. (2018) proposed simultaneous segmentation of WMH and SL from T2-FLAIR using combination of U-Network (UNet) (Ronneberger et al., 2015; Çiçek et al., 2016) and residual network (He et al., 2016) named U-Residual Network (UResNet). Li et al. (2018) proposed a Fully Connected Network (FCN) ensembles based on UNet which combine several models with same architecture to reduce over-fitting problems of a complex model for WHM segmentation. In a more recent study, Jeong et al. (2019) proposed the use of transfer learning to improve the performance of UNet for WMH segmentation. Detailed explanation of the deep learning methods for WMH segmentation mentioned above can be found in Section 3.1.1.

2.3.2.2 Generative Adversarial Networks

GANs consist of two networks, generator and discriminator, which are trained based on game theory scenarios in which the generator must compete against the discriminator (Goodfellow et al., 2016). GANs (Goodfellow et al., 2014) are generally categorised as unsupervised deep learning models because there is no label or target associated with the input needed. However, some models of GANs, such as Conditional GAN (C-GAN) (Mirza and Osindero, 2014), use valuable information like labels as additional input parameter to generate meaningful outputs.

One of the most recent GANs model is the Wasserstein GAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017). WGAN-GP was proposed as an improved training scheme for GANs as the original one is notoriously unstable. In the training, the discriminator (critic) network attempts to distinguish between real (desired) data sampled from training dataset and fake (generated) data sampled from the generator network. Let us assume $\mathbf{x}$ be the real image and a generator network (g_θ) generates a fake image $\mathbf{x}'$ from vector $\mathbf{z} \sim \mathcal{N}(0,1)$ (i.e., $\mathbf{x}' = g_\theta(\mathbf{z})$). Once g_θ is fully trained, $\mathbf{x}'$ (fake image) and $\mathbf{x}$ (real image) should be indistinguishable by a critic/discriminator function (f_w). The f_w returns real values bigger than 0 if real image is inputted (i.e., $f_w(\mathbf{x})$) while it returns real values lower than 0 if fake image is inputted (i.e., $f_w(\mathbf{x}')$). To optimise both generator g_θ and critic f_w, the following minmax objective is used:

$$\arg\min_{g} \max_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_r}[f_w(\mathbf{x})] - \mathbb{E}_{\mathbf{x}' \sim \mathbb{P}_g}[f_w(g_\theta(\mathbf{z}))] \tag{2.6}$$

where $\mathbf{x}$ is a real image sampled from an underlying distribution $\mathbb{P}_r$, $\mathbf{x}'$ is a fake image sampled from an underlying distribution $\mathbb{P}_g$, $\mathbb{E}$ is the expected value (expectation), and $\mathcal{F}$ is the set of 1-Lipschitz functions (Gulrajani et al., 2017).

In most cases, WMH segmentation can be performed by using GANs by learning the manifold (latent) representation of brain's normal tissues through disentanglement of brain's normal (i.e., non-WMH) and abnormal (i.e., WMH) tissues. Disentanglement refers to a process of separating salient factors in the high-dimensional space of data (Bengio et al., 2009). In WMH segmentation, the most important factors are the non-WMH and WMH tissues. The idea behind this approach is that WMH can be detected and then replaced by generated normal brain tissue to produce a "pseudo-healthy" brain image if a latent representation of brain's normal tissue is successfully learned. This has been demonstrated just recently by Xia et al. (2019) where disentanglement of WMH and other brain tissues is performed by using Cycle GANs (Zhu et al., 2017) and pathology factorisation. Unfortunately, the use of GANs for WMH segmentation

is still limited to this day as disentanglement of WMH and non-WMH regions is very challenging.

2.4 Discussion

In this chapter, basic knowledge of WMH has been described. Different ways of quantitative assessment of WMH in cross-sectional and longitudinal have also been introduced and discussed. Finally, several kinds of automatic CAD system for WMH using both conventional machine learning and deep learning algorithms have been introduced.

In the next chapters, the development of machine learning algorithms for segmentation, characterisation, and evolution prediction of WMH is explored. In Chapter 3, the use of CNNs for segmentation of early, small, and subtle WMH is proposed and discussed. Previous studies on WMH segmentation using both machine learning and deep learning algorithms, which have been introduced in Sections 2.3.1 and 2.3.2 respectively, mainly used old WMH which has relatively large volume. These previous methods reported either no evaluation or low performance of small and early WMH segmentation. Note that the segmentation of early WMH is clinically important for preventing the progression of WMH.

In Chapter 4, a novel unsupervised quantitative characterisation method for WMH from T2-FLAIR brain MRI, called Limited One-time Sampling Irregularity Map (LOTS-IM), is proposed and described. The LOTS-IM produces irregularity map (IM) which has higher level of WMH granularity than probability map, produced by machine learning algorithm, and binary mask, produced by human expert. In this chapter, the use of IM for segmentation of WMH and simulating the progression and regression of WMH are described.

In Chapter 5, a novel deep learning model, named Disease Evolution Predictor (DEP), for automatic prediction and estimation of WMH evolution is described. Two DEP models are proposed, DEP based on U-Residual Network (DEP-UResNet) and DEP based on Generative Adversarial Network (DEP-GAN). DEP-UResNet performs prediction and estimation of WMH evolution by segmenting the WMH into three classes: growing, shrinking, and stable WMH. Whereas, DEP-GAN performs prediction and estimation of WMH evolution by regressing the real values of Disease Evolution Map (DEM) (described in Section 5.2). To the best of our knowledge, this is the first extensive study on modelling WMH evolution using deep learning algorithms.

Chapter 3

WMH Segmentation using CNNs with Global Spatial Information

In this chapter, various algorithms from both conventional and deep learning are described, tested, and evaluated for WMH segmentation. Furthermore, the use of Global Spatial Information (GSI) to improve the performance of CNNs on segmenting small and subtle WMH is also proposed. This chapter is based on the following publications:

1. Rachmadi, M., Valdés-Hernández, M., Agan, M., and Komura, T. (2017a). Deep learning vs. conventional machine learning: Pilot study of WMH segmentation in brain MRI with absence or mild vascular pathology. *Journal of Imaging*, 3(4):66.

2. Rachmadi, M., Valdés-Hernández, M., Agan, M., Di Perri, C., and Komura, T. (2018b). Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Computerized Medical Imaging and Graphics*, 66, 28-43.

3.1 Motivation

In this section, previous studies that evaluate automatic methods for segmentation of WMH, challenges of developing WMH segmentation, and contributions of this study are presented.

3.1.1 Existing methods for automatic WMH segmentation

Due to clinical importance of WMH (discussed in Section 2.1) and increasingly large sample sizes of clinical trials and observational studies, considerable efforts have been made to develop automatic assessment of WMH from brain MRI (Caligiuri et al., 2015; García-Lorenzo et al., 2013; Wardlaw et al., 2015). Amongst several attempts to automatically segment WMH from brain MRI (Lao et al., 2008; Schmidt et al., 2012a; Steenwijk et al., 2013; Roy et al., 2015; Yu et al., 2015; Khademi et al., 2012), few methods have shown promising results. One of these works, done by Ithapu et al. (2014), evaluated the application of supervised machine learning algorithms, namely SVM and RF, on WMH segmentation using brain MRI from AD patients. The SVM and RF were tested on 251 subjects, which come from one of the several studies conducted at Wisconsin Alzheimer's Disease Research Centre (WADRC). All scans were acquired on a GE 3T scanner. WMH labels were produced by an expert who scanned through all images and marked out all the WMH regions by using a semi-supervised Random Walker based segmentation method (Grady, 2006) where the expert marked seed points of WMH, traced the segmentation incrementally, and checked for accuracy in a second session to ensure no WMH are missed. For predictors or features that characterise WMH, Ithapu et al. used three-dimensional region of interests (ROIs) with size of $5 \times 5 \times 5$ to extract greyscale values and feed them to a texton-based feature extraction space (Malik et al., 1999). In their study, T2-FLAIR was used as the source for feature extraction and T1-W was used for co-registration and preprocessing. From precision, recall, and Dice similarity coefficient (DSC) measurements obtained for each algorithm, Ithapu et al. concluded that RF was the best machine learning algorithm to do automatic WMH segmentation on their sample.

Another work was done by Leite et al. (2015). They used manually segmented regions from human brain images to train their automatic classifiers, namely SVM, k-nearest neighbour (k-NN), Optimum Path Forest (OPF), and Linear Discriminant Analysis (LDA). The classifiers were tested on 19 healthy volunteers and 54 patients of MS and stroke. The manual region of interest were manually extracted by an expert from two-dimensional slices of the T2-W MRI images and annotated based on the clinical data of the patients. In their study, T2-FLAIR was used as the main source for feature extraction. Features from T2-FLAIR were extracted using statistical analyses based on grey-level histogram, GLCM, GLRLM, and image gradients. Principal component analysis (PCA) was used to reduce the dimension of the feature vector. For evaluation,

accuracy and confusion matrix measurements were used for analysing the performance of each classifier. Leite et al. concluded that SVM was the best classifier in terms of accuracy.

Klöppel et al. (2011) also investigated different methods for WMH segmentation such as greyscale thresholding based on Otsu's method (Otsu, 1975) (thresholding method), k-NN (unsupervised method), and SVM (supervised method). Data for evaluation were collected from 10 subjects with MCI and another set of 10 individuals with dementia. An expert manually outlined WMH based on the T2-FLAIR image from all 20 subjects, and a second expert outlined a subset of 10 randomly chosen images. Their agreement was compared using area under Precision-Recall curve (AUC-PR). Both T2-FLAIR and T1-W were used as sources in feature extraction, and the features were formed by three-dimensional spherical ROI of image intensity values, probability distribution of WMH based on their anatomical location in the brain, and Gabor filters in $1 \times 1 \times 3$ three-dimensional ROIs. The best algorithm in this study in terms of AUC-PR was SVM.

While SVM and RF work well on WMH segmentation according to previous papers, they have a major drawback as conventional machine learning algorithms: hand-crafted features are always needed. This major drawback is eliminated in the current state-of-the-art approach, deep learning using CNN. CNNs (LeCun et al., 1995) are known as the state-of-the-art approach for object recognition in natural images. In a recent study, Moeskops et al. (2018) proposed a multi-scale CNN with different resolution images of a T1-W, a T2-FLAIR, and a T1-W inversion recovery image as input. The method automatically segment WMH and other brain regions (e.g., cortical GM and CSF). The method was evaluated quantitatively with images publicly available from the MRBrainS13 challenge[1] (Mendrik et al., 2015) ($n = 20$) and produced high values of DSC for WMH and other brain regions. The proposed method also produced high correlation of automatic and manual WMH volumes with Spearman's $\rho = 0.83$ for relatively healthy older subjects ($n = 96$) from the Utrecht Diabetic Encephalopathy Study part 2 (UDES2) (Reijmer et al., 2013).

In another study, Li et al. (2018) proposed an FCN ensembles for WHM segmentation which was evaluated and ranked 1st in the WMH Segmentation Challenge 2017[2] at Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017. The proposed method is a variant of FCN architecture based on UNet (Ronneberger et al.,

[1] https://mrbrains13.isi.uu.nl/

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

2015), which takes as input the 2D axial slices of T2-FLAIR and T1-W modalities, and trained in ensemble technique which combine several models with same architecture and is helpful to reduce over-fitting problems of a complex model (Opitz and Maclin, 1999). The proposed method was trained on 60 subjects from 3 different scanners and evaluated on 110 subjects from 5 different scanners (i.e., 2 of them are not represented in the training set) by using five different measurements: DSC, Hausdorff distance (95th percentile) (Huttenlocher et al., 1993), average volume difference (in percentage), sensitivity for individual lesions, and F1-score for individual lesions. Li et al. (2018) reported that the ensemble with 3 or more models clearly outperformed the ensemble of only one model on all of the five measurements.

There is also anther study which proposed the use of CNN for segmenting hyper-intensities and differentiating between WMH and SL (Guerrero et al., 2018). The proposed method is called UResNet which combines UNet (Ronneberger et al., 2015) with residual elements (He et al., 2016) to reduce model complexity. The proposed UResNet was evaluated using 167 MRI data acquired at the Brain Research Imaging Centre of Edinburgh[3] on a GE Signa Horizon HDx 1.5 T clinical scanner (General Electric, Milwaukee, WI). WMH were delineated using Multispectral Coloring Modulation and Variance Identification (MCMxxxVI) while SL were extracted semi-automatically by thresholding and interactive region-growing method, guided by radiological knowledge, on T2-FLAIR and T2-star-weighted (Valdés Hernández et al., 2015a,b). Guerrero et al. (2018) reported that the proposed UResNet outperformed DeepMedic (Kamnitsas et al., 2017) and algorithms from the lesion segmentation toolbox (Schmidt et al., 2012a) where DSC was used as the main evaluation measurement.

3.1.2 Challenges and contributions

WMH at early stages of several neurodegenerative diseases are difficult to assess for two main reasons. The first is their subtlety (i.e., the intensities of WMH are close to the normal tissues), which makes WMH hard to identify, even by human eyes, and easily mistaken by imaging artefacts (Valdés Hernández et al., 2014). The second is their small size which makes WMH hard to detect by automatic WMH segmentation methods. These two facts make the development of automatic WMH segmentation methods for brains with mild or none vascular pathology challenging.

The success of deep learning algorithms in pattern recognition have made them a

[3] http://www.bric.ed.ac.uk/

good candidate for the automatic identification of WMH. For example, Lyksborg et al. (2015), Havaei et al. (2017), and Pereira et al. (2016) used CNNs for segmenting brain tumours; Kleesiek et al. (2016) and Stollenga et al. (2015) also used CNNs for brain extraction and segmenting conventional tissues in general, respectively. Liu et al. (2012) classified MRI data into AD vs. non-AD using Deep Boltzmann Machine (DBM). These works obtained better results from deep learning methods than from classical feature extraction methods, suggesting that the use of deep learning can significantly improve the precision of automatic segmentation of brain MRI features.

In this chapter, a novel way to incorporate spatial information into CNNs for segmenting WMH in the first convolutional layer is proposed and evaluated. This approach is called CNN with GSI (CNN-GSI), where GSI stands for "Global Spatial Information". Spatial information becomes important in WMH segmentation because appearance of the WMH partly depends on their location in the brain; there are regions reported to have higher incidence of WMH (Valdés Hernández et al., 2015b, 2017). These indicate that WMH have different characteristics, given their diverse aetiology, in different locations. Their appearances also depend on clinical factors like blood pressure, type of pathology, disease stage, etc. Therefore not only local and contextual, but also global information are necessary for accurate WMH segmentation.

The most common strategy for incorporating GSI to WMH segmentation schemes consists in masking or weighting the region where the segmentation is applied, either before or after applying the segmentation technique *per se*, using template expressing the probability of each voxel to be WMH (Schmidt et al., 2012a; Shiee et al., 2010). This template is usually a result of averaging and rescaling multiple co-registered WMH segmentation from cohorts of similar clinical characteristics to the one studied (Caligiuri et al., 2015; García-Lorenzo et al., 2013).

Specifically in the case of deep neural networks, Van Nguyen et al. (2015) introduced three-dimensional Cartesian coordinates (i.e., x, y, and z) fused together with other input features using a function for improving results of brain synthesis. In another study, de Brébisson and Montana (2015) explored adding relative distances of each voxel to the centroids of each brain's regions for improving brain segmentation result. Ghafoorian et al. (2017a) also proposed adding eight hand-crafted spatial location features to segmentation layer of CNNs to improve the results. While these approaches have been shown to be useful, relying to spatial information that are hand-crafted produced by some specific functions may result in ignoring the scarce subtle WMH due to inconsistencies. Hence, incorporating spatial information in the form of a

synthetic volume (Steenwijk et al., 2013; Roy et al., 2015) as an input to CNNs through additional channels is proposed. In this way, convolutional layers learn automatically the representation of spatial information for all types of WMH (i.e., subtle and non-subtle WMH) without hand-crafted features.

The performance of the proposed CNN-GSI framework is compared with those of existing CNN (i.e., CNN without GSI), SVM, RF, and DBM. Both SVM and RF have been reported to work well for WMH segmentation (Ithapu et al., 2014; Klöppel et al., 2011). DBM is another model of supervised deep neural network which reportedly works well for feature extraction of MRI (Liu et al., 2012). In this study, greyscale value and texton features are used as features for SVM and RF, as per (Ithapu et al., 2014). Whereas, only greyscale value of the voxels is used for DBM. The results obtained by the proposed deep learning schemes are also compared against those obtained from a popular public tool, namely LST-LGA (Schmidt et al., 2012a). The results of all methods are compared and analysed. Finally, the results from six schemes that performed best against the performance of trained human observers are evaluated in neuroradiological clinical assessments.

In summary, the contributions in this study are comparing the use of CNNs with the other algorithms (i.e., LST-LGA, SVM, RF, and DBM) for automatic WMH segmentation in routine clinical brain MRI of individuals with none or mild vascular pathology and proposing a way for incorporating spatial information into CNNs in the first convolutional layer by creating an artificial volume information named GSI.

3.2 Materials and data processing

In this section, the MRI data samples, preprocessing steps, and postprocessing steps used in this study are described. All preprocessing and postprocessing steps are used in both conventional machine learning and deep learning algorithms.

3.2.1 Subjects and MRI data

The data used in this study is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [4] public database (Mueller et al., 2005; Weiner et al., 2013). ADNI

[4]Data used in preparation of this article were obtained from the ADNI database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Note that ADNI did not provide any labels of WMH in their database.

- The first dataset used in this study contains MRI data from 20 ADNI participants (12 men and 8 women, mean age at baseline 71.7 (standard deviation (SD) 7.18) years), which are the first 20 patients from ADNI-GO study imaged in 3 consecutive years, resulting in data from a total of 60 MRI scans. Three of them were cognitive normal (CN), 12 had early MCI, and 5 had late MCI. But the Mini Mental State Examination scores did not differ considerably between these 3 cognitive groups of individuals: mean values were 28.5 (SD 2.12) for the CN, 27.83 (SD 1.75) for early MCI and 27.67 (SD 2.08) for late MCI. The cognitive status of the individuals that provided data for this study did not change across the 3 visits. Other than the availability of WMH labels (discussed in Section 3.2.2) and measurements for inter-/intra-observer reliability analysis (discussed in Section 3.2.3), no other criteria (e.g., clinical, imaging, or demographic information) were used for data/subject selection. The distribution of WMH size (volume) of this dataset is depicted in Figure 3.1.

- The second dataset used in this study contains 268 MRI data from 268 different ADNI participants, for which WMH reference masks are unavailable. The only labels available for each MRI data from this second dataset are Fazekas scores (described in Section 2.2.1) consisting of visual ratings of WMH burden in the PVWMH and DWMH (Fazekas et al., 1987). In this study, paired Spearman's correlation is used to assess monotonic correlation between the total Fazekas score (i.e., the sum of PVWMH and DWMH scores) and the WMH volume estimated automatically by CNNs. This dataset was chosen to evaluate the performance of different machine learning algorithms in a bigger dataset with different WMH severity.

The mean and SD of the clinical data that has been reported to be relevant to WMH burden and progression (Wardlaw et al., 2013) and acquired at each MRI visit (i.e., diastolic blood pressure, systolic blood pressure, and pulse rate) for the first dataset are summarised in Table 3.1. To evaluate the clinical relevance of the results, the serum

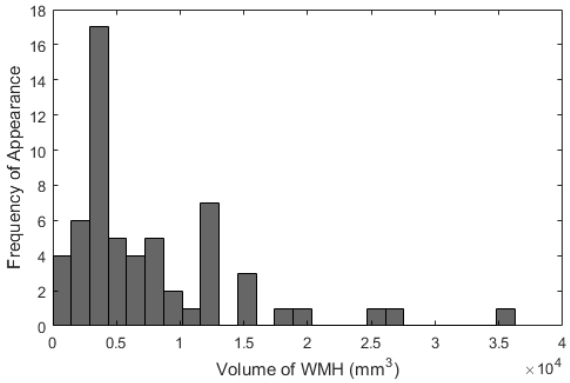


Figure 3.1: Individuals with mild or no vascular pathology have, in general, a small WMH volumetric burden. Histogram showing the volumetric burden of WMH, represented by their volume (mm^3) in the first dataset (see Section 3.2.1).

cholesterol and glucose levels obtained on visit 1 are also used. Studies have shown these factors could play a role in WMH progression (Dickie et al., 2016). The mean (SD) values were 206.2 (35.38) mg/dL for cholesterol and 96.4 (11.35) mg/dL for glucose. Finally, MRI data acquisition parameters of T1-W and T2-FLAIR for both datasets are shown in Table 3.2.

3.2.2 Ground truth

Ground truth (GT) WMH labels of the first dataset were produced by an experienced image analyst (Observer #1), semi-automatically by delineating the contours of WMH in T2-FLAIR images using the region-growing algorithm in the Object Extractor tool of Analyze^{TM} software, simultaneously guided by the co-registered T1-W and T2-W sequences. Each brain scan was processed independently, blind to any clinical, cognitive or demographic information, and to the results of the WMH segmentation from the

Table 3.1: Mean and SD of the clinical data (diastolic blood pressure, systolic blood pressure and pulse) of the individuals in the first dataset.

Parameter	Year 1 *mean (SD)*	Year 2 *mean (SD)*	Year 3 *mean (SD)*
Diastolic BP (mmHg)	72.60 (8.95)	73.25 (11.01)	73.80 (11.81)
Systolic BP (mmHg)	125.55 (12.56)	127.00 (12.94)	128.70 (13.97)
Pulse rate (bpm)	65.10 (10.78)	61.00 (9.53)	62.45 (13.82)

Table 3.2: Data acquisition protocol parameters of both datasets.

Parameter	T1-W	T2-FLAIR
In-plane matrix (pixels)	256 × 256	256 × 256
Number of slices	256	35
Thickness (mm)	1.2	5
In-plane resolution (mm)	1.0 x 1.0	0.8594 x 0.8594
Repetition time (ms)	2300	9000
Echo time (ms)	2.98	90 or 91
Flip Angle	9.0	90 or 150
Pulse Sequence	GR/IR	SE/IR

same individual at different time points. The resultant mean WMH volume of the GT labels for Year 1 was 6002.1 (mm^3) (SD 4112.7), for Year 2 5794.9 (mm^3) (SD 4281.6), and for Year 3 7004.2 (mm^3) (SD 5274.7). For more details and to access these labels, please refer to the datashare Uniform Resource Locator (URL)[5]. Visualisation of WMH label produced by Observer #1 is depicted in Figure 3.2 (middle).

3.2.3 Measurements for inter-/intra-observer reliability analyses

It is worth mentioning that relying on one assessment from one rater is often biased towards the rater's expertise and experience. Thus, a second image analyst (Observer #2) generated two sets of longitudinal WMH binary masks for 7/20 subjects (i.e., 42 measurements in total), blind to the GT measurements and to previous assessments for measurements for inter-/intra-observer reliability analyses. These were done semi-automatically using Mango[6], individually thresholding each WMH 3D cluster in the original T2-FLAIR images. Note that Observer #1 and Observer #2 used different tools based on their experience for creating the labels. Visualisation of WMH label produced by Observer #2 is depicted in Figure 3.2 (right). As shown in Figure 3.2, there are some differences between WMH labels produced by Observer #1 and Observer #2, largely due to different experience on detecting early and subtle WMH. Information and segmentation results of the 7 subjects for intra-/inter-observer reliability evaluation can be accessed in another datashare URL[7].

[5] http://hdl.handle.net/10283/2214
[6] http://ric.uthscsa.edu/mango/
[7] http://hdl.handle.net/10283/2706

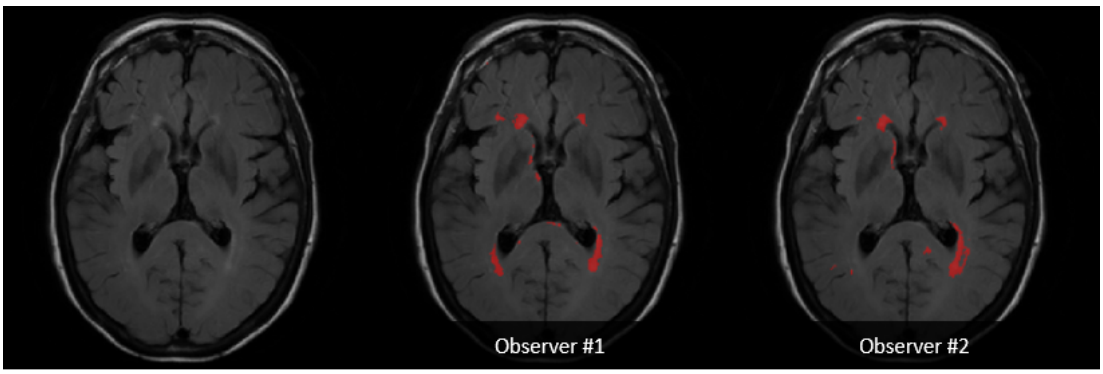


Figure 3.2: Visualisation of T2-FLAIR MRI (left) and corresponding WMH labels produced by Observer #1 (middle) and Observer #2 (right). The difference between two WMH labels produced by the two observers is measured in inter-observer analysis using Equation (3.12).

3.2.4 Preprocessing

The preprocessing steps of the data comprise co-registration of the MRI sequences on each scanning session, skull stripping and intracranial volume mask generation, cortical GM/CSF/brain ventricle extraction, and intensity value normalisation. Rigid-body linear registration of the T1-W to the T2-FLAIR image, as T2-FLAIR is the base sequence for identifying WMH, is achieved using FSL-FLIRT (Jenkinson et al., 2002). Note that rigid-body linear registration could be performed because there are no apparent deformations to the brain's regions (e.g., ventricle) in the both datasets. Skull stripping and generation of the intracranial volume mask are done using optiBET (Lutkenhoff et al., 2014). OptiBET, while attempting to extract the brain, also excludes parts of the brain ventricles from the intracranial volume. Therefore, morphological operation of fill holes is performed to the binary mask created by optiBET to obtain the intracranial volume.

Cortical GM, CSF, and brain ventricles are three brain regions where WMH do not appear and can present artefacts often wrongly mislabelled as WMH (Wardlaw et al., 2015). Because of that, these regions are excluded by masking them out as follows. Binary masks of NAWM and cerebrospinal fluid are obtained using FSL-FAST (Zhang et al., 2001). The holes in the obtained white matter mask are filled using morphological operation of "closing". Subsequently, the ventricles (and possible lacunes) are removed from it by subtracting the results of a logical "and" operation between the "filled white matter" mask and the mask of cerebrospinal fluid.

Intensity value normalisation is done in two steps. The first step is adjusting the

maximum grey scale value of the brain without skull to 10 percent of the maximum T2-FLAIR intensity value so that each data has the same maximum intensity value while not stretching the values too far. The second step is adjusting the contrast and brightness of the MR images such that their histograms are consistent. To equalise contrast and brightness, a histogram matching algorithm for MR images developed by Nyúl et al. (2000) is performed where an MR image is used as a reference image. The approach of using histogram matching for preprocessing images with non-healthy tissue has been reported to be promising (Shah et al., 2011) and previously used for preprocessing in CNNs approaches (Pereira et al., 2016). Furthermore, normalisation of the intensities into zero-mean and unit-variance were also necessary so that the modifications implemented to optimise the CNNs can run smoothly.

3.2.5 Postprocessing

Results from all segmentation schemes are expressed as probability maps of voxels being WMH. To make a clear-cut segmentation, cutting of the probability map's values using a threshold value of $t \geq 0.5$ (chosen as being the standard for two-class segmentation) is performed, Then, the voxels that belong to 3D clusters smaller than 3 mm^3 maximum in-plane diameter (as per definition of WMH in Wardlaw et al. (2013)) are also removed. Furthermore, the NAWM mask is used to eliminate the spurious false positives that may appear in the cortical brain region. In the evaluation, both probability maps and clear-cut segmentation results are used.

3.3 Conventional Machine Learning Algorithms, Feature Extraction, and Public Toolbox

The performance of the CNNs is compared against the output from two conventional machine learning algorithms, SVM (Cortes and Vapnik, 1995) and RF (Tin Kam Ho, 1995), and LST-LGA (Schmidt et al., 2012a) commonly used in medical image analysis for WMH segmentation. SVM is a supervised machine learning algorithm that separates data points projected to a high-dimensional feature space by using a hyperplane (Cortes and Vapnik, 1995). RF is a collection of decision trees trained individually to produce outputs that are collected and combined together (Tin Kam Ho, 1995; Criminisi and Shotton, 2013). Detailed explanation of SVM and RF are described in Section 2.3.1.2.

A public toolbox named W2MHS[8], developed by Ithapu et al. (2014), was modified so that the desired conventional machine learning algorithms, SVM and RF, could be trained using the available GT whilst using the same feature extraction methods for repeatability and reproducibility reasons. The modified version extracts greyscale values and texton based features from either T2-FLAIR or T1-W MRI sequences on $5 \times 5 \times 5$ regions of interest as per (Ithapu et al., 2014). Texton-based features are formed by concatenating all responses from low-pass, high-pass, band-pass, and edge filters (full explanation in Ithapu et al. (2014)). An array of 2000 values were generated by the texton feature extraction and used for SVM and RF.

The results is also compared against LST-LGA (Schmidt et al., 2012a) version 2.0.15[9]. LST-LGA performs lesion segmentation by producing intensity distributions and belief classes of CSF, GM, and WM. The assumption is that lesions behave as hyperintense outliers from these distributions. Afterwards, the proposed lesion growth algorithm performs expansion of lesion seeds (i.e., hyperintense outliers), deemed as conservative assumption for lesions, using approximation of gamma distribution while the distributions of GM, WM, and CSF are approximated by a mixture of three Gaussian. If a voxel is more likely to be part of other classes while completely surrounded by lesion voxels, Markov random field is utilised for computing the final probability. The threshold parameter κ is then used to cutoff the belief map for final segmentation. In the original study, the LST-LGA was tested on 18 control patients and 52 MS patients. The manual segmentation was independently performed by two investigators, who were blinded to the study group, by applying a semi-automatic manual tracing pipeline using commercially available software (Amira 5.3.3, Visage Imaging, Inc). A difference image of the two labels was generated and both experts together decided which differences were assigned to lesions or not. The performance of LST-LGA was then evaluated using volumetric agreement (i.e., correlation and regression) and spatial agreement (i.e., DSC (Dice, 1945)) measurements. Unfortunately, the original study did not compare with any previous methods for lesion segmentation and did not test the LST-LGA for small lesion segmentation (i.e., lesion volume $\leq$ 2 ml). In this study, LST-LGA with kappa-value $\kappa = 0.05$, the lowest recommended kappa-value from LST-LGA, was used to increase sensitivity to the subtle WMH.

[8] https://www.nitrc.org/projects/w2mhs/
[9] www.statisticalmodelling.de/lst.html

3.4 Deep Learning Algorithms

In this section, a supervised deep learning algorithm named DBM is explained briefly. Then, the setup of CNNs scheme using DeepMedic (Kamnitsas et al., 2017) and how GSI is encoded into the CNNs are described in details.

3.4.1 Deep Boltzmann Machine

DBM is a variant of the Restricted Boltzmann Machine (RBM), a generative neural network that works by minimizing its energy function, and uses multiple layers of RBM instead of only one layer. Each hidden layer captures more complex high-order correlations between activities of hidden units than the layer below (Salakhutdinov and Hinton, 2009). *Pre-training* can be done independently in each layer to get better better initialization of the weight matrix. In this study, a simple DBM with two hidden layers is used (Figure 3.3). The energy function of DBM (E) is defined by Equation (3.1):

$$E\left(\mathbf{v},\mathbf{h}^1,\mathbf{h}^2;\Theta\right) = -\mathbf{v}^\top \mathbf{W}^1\mathbf{h}^1 - (\mathbf{h}^{1\top})\mathbf{W}^2\mathbf{h}^2 \tag{3.1}$$

where $\mathbf{v}$ is the vector of visible layer (i.e., voxel intensity values), $\mathbf{h}^1$ and $\mathbf{h}^2$ are vectors for the first and second hidden layers respectively (i.e., feature maps), and $\Theta = \left\{\mathbf{W}^1,\mathbf{W}^2\right\}$ encloses the model's parameters where $\mathbf{W}^1$ and $\mathbf{W}^2$ are symmetric matrices (i.e., weights) that connect visible-hidden and hidden-hidden layers respectively. The DBM's objective function is the probability that the DBM model generates back the visible variables of $\mathbf{v}$ using the DBM model's parameter Θ, as per Equation (3.2):

$$p(\mathbf{v};\Theta) = \frac{1}{Z(\Theta)} \sum_{\mathbf{h}^1,\mathbf{h}^2} \exp\left[-E\left(\mathbf{v},\mathbf{h}^1,\mathbf{h}^2;\Theta\right)\right]. \tag{3.2}$$

where E is the energy function as per Equation (3.1), exp is the exponential function, and $Z(\Theta)$ is a partition function over all possible configurations (i.e., a normalising constant to ensure the probability distribution sums to 1). Given a restricted structure where each layer units are conditionally independent from each other, the conditional distribution of the probability for a unit in a layer given other layers can be computed as in Equations (3.3), (3.4), and (3.5) as follows:

$$p\left(h_p^2 = 1|\mathbf{h}^1\right) = \sigma\left(\sum_n W_{np}^2 h_n^1\right) \tag{3.3}$$

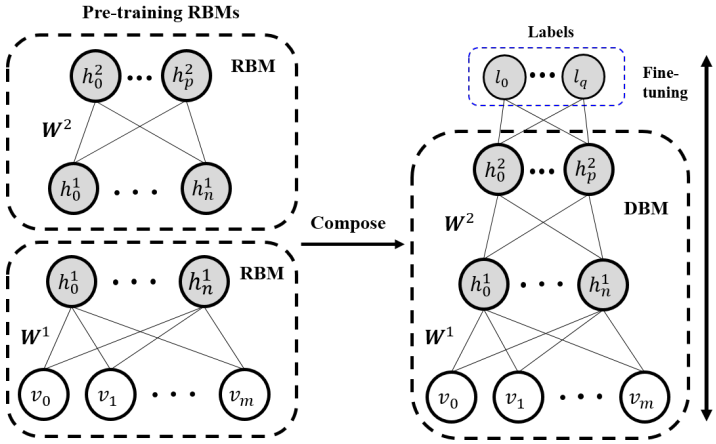


Figure 3.3: Illustrations of DBM used in this study. Two RBMs (left) are stacked together for pre-training to form a DBM (right).

$$p\left(h_n^1 = 1|\mathbf{v},\mathbf{h}^2\right) = \sigma\left(\sum_m W_{mn}^1 v_m + \sum_p W_{np}^2 h_p^2\right) \tag{3.4}$$

$$p\left(v_m = 1|\mathbf{h}^1\right) = \sigma\left(\sum_n W_{mn}^1 h_n\right) \tag{3.5}$$

where σ is a sigmoid non-linear function and m, n and p are neuron's index of input layer vector, hidden layer vector, and output layer vector (please look at Figure 3.3 for visual explanation). Full mathematical derivation of RBM and its learning algorithm can be read in Hinton (2010) and the derivation of DBM and its learning algorithm in Salakhutdinov and Hinton (2009).

In this study, 3D ROIs of $5 \times 5 \times 5$ are used to get grayscale intensity values from the T2-FLAIR MRI for DBM's training process. The intensity values are feed-forwarded into a 2-layer DBM with 125-50-50 structure where 125 is the number of units of the input layer and 50 is the number of units of both hidden layers. Each RBM layer is pre-trained for 200 epochs, and the whole DBM is trained for 500 epochs. After the DBM training process is finished, a label layer is added on top of the DBM's structure and *fine-tuning* is done using gradient descent for supervised learning of WMH segmentation. Salakhutdinov's DBM public code[10] was modified and used for this study.

[10] http://www.cs.toronto.edu/~rsalakhu/DBM.html

3.4.2 Convolutional Neural Network

CNN (LeCun et al., 1995) has emerged as a powerful supervised learning scheme on natural images that can learn highly discriminative features from a given dataset (Kamnitsas et al., 2017). Unlike fully connected neural networks, CNN uses sparse local connections instead of dense, which is realized by the *convolutional layers* that apply local filters to a portion of input image called *receptive field*. Multiple filters are used to learn more variants of object's features in each convolutional layer where their activations generate multiple number of *feature maps*. Because of the sparse local connection, the convolutional layers of CNN have fewer parameters to train than the fully connected neural networks, and it can naturally learn contextual information from the data which is important in object detection and recognition (LeCun et al., 2015). Several number of convolutional layers can also be stacked together to capture more complex feature representations of the input image.

In this study, a CNN framework named "DeepMedic" proposed by Kamnitsas et al. (2017), which efficiently implements a dual-pathway scheme for CNN (discussed in Section 3.4.2.4), is used. The publicly available DeepMedic is chosen for reproducibility and repeatability reasons. Also, 2D CNN is used instead of 3D CNN like in the original study due to the anisotropy of the MR images used in this study (i.e., the T2-FLAIR MRI from ADNI database have dimensions of $256 \times 256 \times 35$ and voxel size of $0.86 \times 0.86 \times 5$ mm^3). Note that the T2-FLAIR sequence is usually anisotropic due to the acquisition time required and the limited practical use that clinically poses to acquire it isotropically (i.e., a clinician does not need isotropic voxels to estimate the burden of WMH, and by definition WMH are minimum 3 mm $\times$ 3 mm $\times$ 3 mm in diameter for being considered relevant), meaning that in clinical practice there will be either an inter-slice "gap" or a wider "thickness" between each T2-FLAIR slice. In such case, a deep learning scheme with 3D operations would not be as effective as when isotropic data are used.

3.4.2.1 Global Spatial Information for CNNs

GSI in this study refers to a set of synthetic images that encode spatial information of brain in MRI. CNNs are powerful methods to extract features from a set of images when these are local features of an object. However, CNNs are not designed to learn global spatial information of some specific features, especially when patch-based CNNs is performed. As spatiality of features is an important information for WMH segmentation

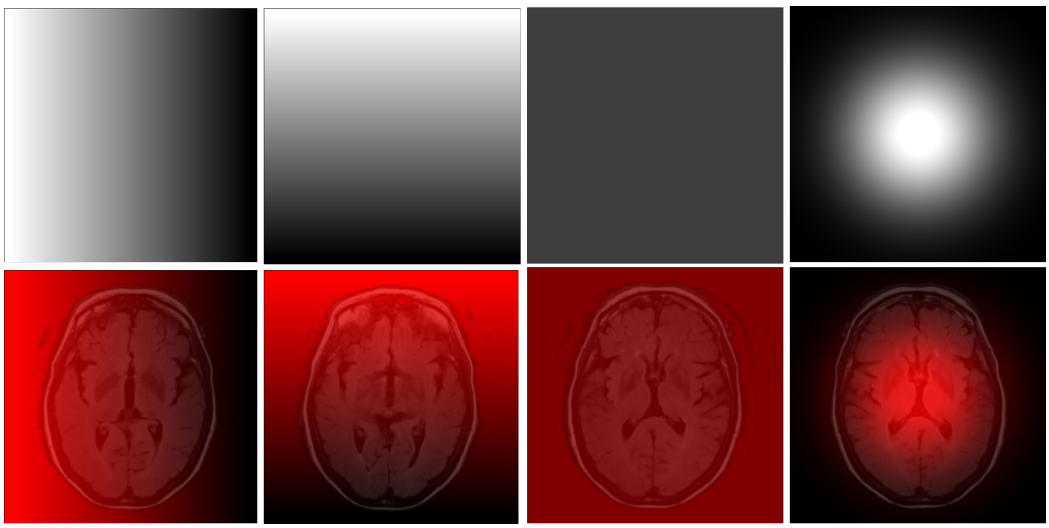


Figure 3.4: Illustration of four different types of GSI of MRI proposed in this study, which are x axis, y axis, z axis, and radial. Upper ones are the synthesised images of spatial information, while the lower ones are MRI overlaid by spatial information.

(Kim et al., 2008), the proposed GSI is designed to augment the performance of CNNs for this task.

In this study, GSI is a set of four different spatial information from the three MRI axes (i.e., x, y, and z axes) and a radial filter that encodes the distance from the centre of the MR image. In each axis, numbers in the range of 0 to 1 are generated sequentially to realise a *spatial information slide* for each axis. The radial filter is generated using a 2D Gaussian function where $x = y = 256$, $\mu = 51$, and $\sigma = 51$ (i.e., an arbitrary value that generates a nice cover of the 2D Gaussian function to an MRI slice sized 256×256). The illustration of GSI can be seen in Figure 3.4. An illustration of CNN-GSI is depicted in Figure 3.5.

3.4.2.2 Network architecture

Small-sized and single stride kernels, preferred for MR images (Simonyan and Zisserman, 2014), are used in all convolutional layers. Two different CNN architectures, which are 5 convolutional layers of 2D CNN and 8 convolutional layers of 2D CNN, are implemented by using the DeepMedic framework (Kamnitsas et al., 2017). Two different architectures are used to see different impacts of spatial information (i.e., GSI) in different CNN architectures. The first network has a receptive field of 15×15 while the second one has 17×17. The performance of the two architectures are compared with each other and other conventional and deep machine learning algorithms in the

evaluation (see Table 3.5).

To make the comparison feasible between schemes (and with other works that may use DeepMedic for the same purpose), only the number of convolutional layers and their kernel size are changed. The original 3D CNN of DeepMedic is formed by 8 convolutional layers, 2 fully connected layers and 1 segmentation layer. Fully connected layers are used to combine normal and sub-sampled pathways (will be explained in the next sub-section) whereas the segmentation layer is an output layer for voxel classification. There is a naive up-sampling operation layer in the sub-sampled pathway to make sure that the size of input segment for fully connected layers from both pathways are the same. For regularisation, DeepMedic uses *dropout* (Srivastava et al., 2014; Hinton et al., 2012) in the two last layers (i.e., the second fully connected layer and the classification layer), where some nodes from fully connected layers are removed with some probability p thus forcing the network to learn better representations of the data. In this study, the dropout probability is set to $p = 0.5$. *Data augmentation*, which is useful for reducing overfitting (Krizhevsky et al., 2012), is also used with some variations in rotation space (i.e., the original training data are rotated by the x axis with probability $p = 0.5$). *Pooling layer* is not used because, while pooling is usually used to make feature representation invariant to small changes and more compact (LeCun et al., 2015), it might introduce some spatial invariances undesirable for lesion segmentation (Kamnitsas et al., 2017). A diagram of the CNN architecture used in this study can be seen in Figure 3.5.

3.4.2.3 Kernel function and loss function

Transformation in convolutional layers is achieved by convolving kernels to the input image segments and applying the output to an activation function. Each convolution computes a linear transformation between input values and weight values of kernels whereas the activation function applies a non-linear transformation to its input. The calculation of linear transformation between input values and kernel weight values can be written as in Equation (3.6) where h is output to the neuron, $\mathbf{x}$ is a one-dimensional input vector, $\mathbf{w}$ is a one-dimensional vector of kernel weight values, β is a bias value, and σ is a non-linear activation function. In this study, Parametric Rectifier Linear Units (PreLU) activation function (Equation (3.7)) is used where α is a trainable parameter

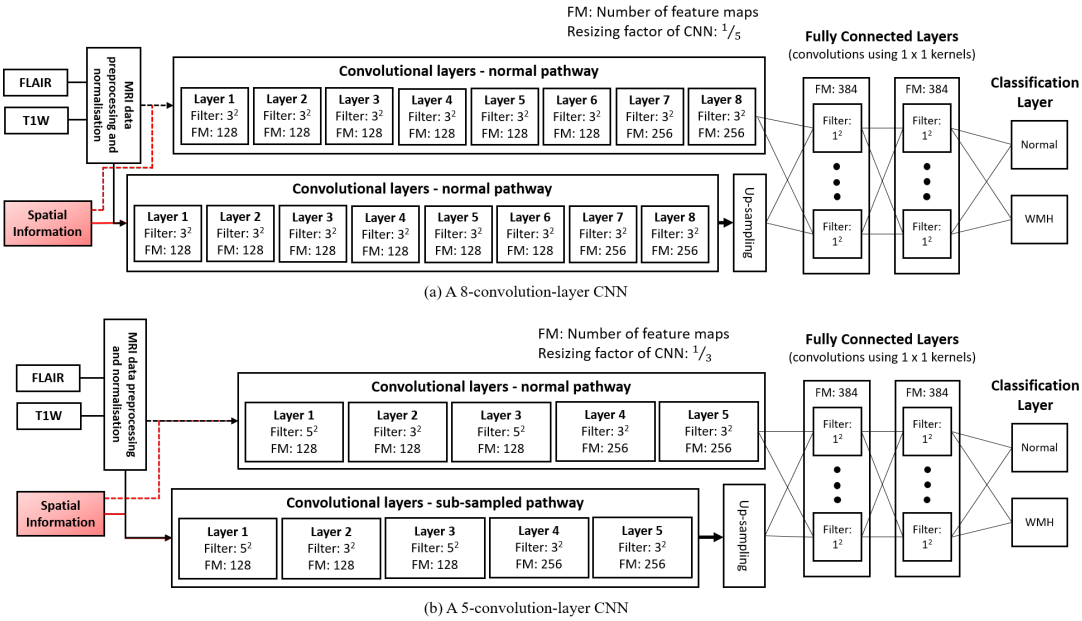


Figure 3.5: A diagram of two CNN architectures used in this study, which are based from 3D CNN DeepMedic framework (Kamnitsas et al., 2017). The upper one, (a), is formed of 8 convolution layers whereas the lower one, (b), is formed of 5 convolution layers. Black dashed arrows refer to the *normal* pathways whereas non-dash arrows refer to the *sub-sampled* pathways. Red boxes and arrows refer to the GSI and its path to the network respectively.

(He et al., 2015).

$$h = \sigma\left(\mathbf{x}^\top \mathbf{w} + \beta\right) \tag{3.6}$$

$$\sigma(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{otherwise} \end{cases} \tag{3.7}$$

Voxels in the WMH segmentation scheme will be of two classes: WMH (i.e., y) and non-WMH (i.e., $(1-y)$). Hence, a binary cross-entropy loss (BCEL) function, written in Equation (3.8), is used, where q is the predicted (class) probability of the voxel, y is the target (class) label of the voxel, i is the voxel index, and N is the number of all voxels.

$$BCEL = -\frac{1}{N}\sum_{i=1}^{N} y_i \log q_i + (1-y_i)\log(1-q_i) \tag{3.8}$$

3.4.2.4 Multiple-pathway architecture of CNNs

Multiple-pathway architecture refers to the use of additional paths to extract more contextual information. Approaches of multiple-pathway CNNs have been previously studied by Havaei et al. (2017), Moeskops et al. (2016), and Kamnitsas et al. (2017). By applying multiple-pathway architectures, different amounts of contextual information can be extracted and used simultaneously. In Moeskops et al. (2016), for example, the authors use three paths of CNN where the second and third paths use twice and thrice the size of the first path's receptive field. Note that the amount of contextual information is decided by the size of the receptive field.

Multiple-pathway structures introduce more parameters and thus results in larger memory usage and computation time. To avoid the explosion of memory usage and processing time, Kamnitsas et al. (2017) introduced a new scheme of multiple-pathway (i.e., two-pathway) where different resolutions of input images are fed into two different networks and then merged together at the end. For example, by resizing MR images to be $1/3$ of the original size, three times bigger receptive field of MR images can be obtained without increasing the number of parameters. Full reports on its application can be read in (Kamnitsas et al., 2017). In this study, the resizing factor of either $1/3$ (default) or $1/5$ is used to see whether different resizing factor affects performance of CNN-GSI. For the rest of this chapter, the original and resized paths will be referred as *normal* and *sub-sampled* pathways respectively. The illustration of the dual-pathway architecture of CNNs proposed by Kamnitsas et al. (2017) and used in this study can be seen in Figure 3.5.

3.4.2.5 Image segments and training

Image segments are image patches used as input to the CNNs. As WMH segmentation is performed on a voxel basis, a full MR image does not have to be loaded into the CNNs. Image segments used in the training process are selected using the scheme developed in DeepMedic framework where probability of 50% is used to extract an image segment centred on a non-WMH or WMH (Kamnitsas et al., 2017). Root Mean Square propagation (RMSprop) optimiser (Dauphin et al., 2015) is used to minimise the binary cross-entropy loss function. RMSprop's main idea is to divide the gradient by a running average of its recent magnitude. This way, RMSprop can be used in mini-batch training unlike its predecessor resilient propagation (rprop) (Riedmiller and Braun, 1992). To speed up the training process, Nesterov's Accelerated Momentum

(Sutskever et al., 2013) is also used where the momentum value is kept constant to 0.6 while learning rate decreases linearly from its initial value of 0.001.

3.5 Experimental Setup

In this section, training and testing processes, parameter setup of machine learning methods, and evaluation methods used in this study are presented.

3.5.1 Training and testing processes

Due to the limited number of data available in the first dataset (i.e., 60 MRI scans), a 5-fold cross validation is used across the dataset, where 48 MRI scans from 16 individuals are used as training samples and 12 MRI scans from 4 individuals are used for testing. The selection of individuals/subjects for training and testing in each cross validation was done randomly. All MRI scans from the first dataset are used as training samples for generating the WMH segmentation of the second dataset (i.e., from 268 MRI scans), which is used as testing sample. Performance is evaluated using the Fazekas scores.

Class balancing (i.e., WMH and non-WMH) from training datasets is done differently depending on the machine learning algorithm used. For SVM and RF algorithms, the same sampling scheme as in (Ithapu et al., 2014) is performed, which is to equally sample WMH and non-WMH data from the training dataset. For DBM, weighted sampling method of WMH and non-WMH is performed, where the number of non-WMH data are four times more than the WMH data. For CNNs, dense training on image segments that adjusts to the true distribution of non-WMH and WMH provided in DeepMedic framework (Kamnitsas et al., 2017) is used.

3.5.2 Parameter setup

There are some parameters for each machine learning method that need to be set before starting the training process. In this study, for each machine learning methods, the sets of parameters that previous studies reported to give the best results are used, verified in the preliminary experiments (Rachmadi et al., 2017b). RBF is used for SVM classifier and extracted features for conventional machine learning, discussed in Section 3.3, is reduced to 10 using PCA and then whitened before training. The RF model used in this training is set using the following parameters: 300 trees, 2 minimum samples in a leaf, and 4 minimum samples before splitting. A 2-layer DBM with 125-50-50

Table 3.3: Parameters of Convolutional Neural Network (adopted directly from (Kamnitsas et al., 2017))

Convolutional Neural Network		
Stage	**Parameter**	**Value**
Initialisation	weights	(He et al., 2015)
Regularisation	L1	0.000001
	L2	0.0001
Dropout	p - 2nd last layer	0.5
	p - Last layer	0.5
Training	epochs	35
	momentum	0.5
	Initial LR	0.001

structure is constructed where 125 is the number of units of the input layer and 50 is the number of units of both hidden layers. Each RBM layer is pre-trained for 200 epochs, and the whole DBM is trained for 500 epochs. In the end of the training process, a label layer is added on top of the DBM's structure and *fine-tuning* is done using gradient descent for supervised learning of WMH segmentation. The CNN has many hyper-parameters for constructing the network, so the default parameters provided by DeepMedic framework are used as they have been reported to work well for segmentation and also for reproducibility reason. All parameters of the CNN are listed in Table 3.3.

3.5.3 Evaluation

AUC-PR and DSC measurements, the most commonly used measurements to evaluate medical image segmentation results, are calculated for evaluation. AUC-PR calculates the size of area under the precision (i.e., Positive Predictive Value (PPV)) and recall (i.e., True Positive Rate (TPR)) curve between GT and the automatic segmentation result. DSC (Dice, 1945) measures similarity (i.e., spatial coincidence) between GT and automatic segmentation results. Precision, recall, and DSC are defined in Equations (3.9), (3.10), and (3.11) where True Positive (TP), False Positive (FP), and False Negative (FN). A paired two-sided Wilcoxon signed rank significance test was performed to see

whether the improvements were significant or not.

$$Precision = PPV = \frac{TP}{TP + FP} \tag{3.9}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{3.10}$$

$$DSC = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{3.11}$$

As an additional evaluation, the non-parametric Spearman's correlation coefficient, which is used to assess monotonic correlation between the total Fazekas scores (Fazekas et al., 1987) and the WMH volumes produced by all automatic schemes, is evaluated. It is known that these two measurements are highly correlated (Valdés Hernández et al., 2013). Fazekas scores (described in Section 2.2.1) consider the WMH subdivided into PVWMH and DWMH. For the evaluation, the PVWMH and DWMH ratings are summed for each of the 268 unlabelled MRI data in the second dataset.

Two additional measurements called Volume Difference (VD) and volumetric Disagreement (D) were also calculated for evaluating intra-/inter-observer reliability measurement. VD (Equation (3.12)) evaluates volumetric difference between predicted segmentation (PS) and GT labels using volume (vol) function which computes volumetric measurement by multiplying the number of PS/GT voxels in one patient with the real-world voxel size, $0.8594 \times 0.8594 \times 5$ mm^3 (see data acquisition protocol parameters for T2-FLAIR in Table 3.2). On the other hand, D is used to evaluate volumetric disagreement of intra-/inter-observer reliability. In intra-observer reliability test, D (Equation (3.13)) is used to evaluate disagreement between automated schemes ($GT_1 = PS$) and two manual WMH labels produced by the first observer ($GT_2 =$ Observer #1) on 12 random MRI scans. In inter-reliability test, D is used to evaluate disagreement between automated schemes ($GT_1 = PS$) and two manual WMH labels produced by two different observers (i.e., $GT_2 =$ Observer #1 and $GT_2 =$ Observer #2 in the first and second evaluation respectively) on 20 random MRI scans.

$$VD = \frac{vol(PS) - vol(GT)}{vol(GT)} \tag{3.12}$$

$$D = abs\left(\frac{vol(GT_1) - vol(GT_2)}{mean(vol(GT_1), vol(GT_2))}\right) \times 100\% \tag{3.13}$$

In addition, the outcome of each segmentation method in relation with age, gender, and clinical parameters selected based on clinical plausibility and/or previous research

(i.e., blood pressure parameters (systolic and diastolic), pulse rate, cholesterol, and serum glucose) was evaluated. One-way analysis of covariance (ANCOVA) were performed to evaluate the association of candidate variables (clinical data) with potential change in WMH volume at each time point. Since WMH volumes were obtained at three time points (i.e., Year 1 (Y1), Year 2 (Y2), and Year 3 (Y3)), evaluation was performed for potential change from Y1 to Y2, Y2 to Y3, and Y1 to Y3. Prior to conducting each ANCOVA model, collinearity assessment using Belsley collinearity diagnostics was performed (Belsley et al., 2005), independence between each covariate and the independent variable, and homogeneity of regression slope assumptions, all using MATLAB 2015a.

Finally, the results of the six best-performing schemes were visually evaluated by a neuroradiologist using a form, which records the number of WMHs not identified, missed partially, and misclassified in the following anatomical brain regions: pons, periventricular, corpus striatum, anterior, central, and posterior white matter bundles. Completed forms by the neuroradiologist are given as supplementary material in Appendix A.

3.6 Results and Discussions

In this section, the impact of using multiple MRI sequences for automatic segmentation of WMH, the difference in performance between conventional machine learning algorithms (i.e., SVM and RF) and deep learning algorithms (i.e., DBM and CNN), the differences in performance of the public toolbox (i.e., LST-LGA) versus other algorithms, the impact of using GSI in CNN, the influence of WMH volume in the performance of each algorithm, longitudinal analysis, intra- and inter-observer analyses, the processing time needed for training and testing each algorithm, and the clinical evaluation of automatic WMH segmentation schemes are discussed.

In total, 5 machine learning algorithms with 24 different schemes/settings were tested in this study for automatic WMH segmentation. The list of the machine learning algorithms can be seen in Table 3.4 whereas all schemes/settings and their general evaluation can be seen in Table 3.5.

Table 3.4: List of all machine learning algorithms and their category used in this study. “ML”, “SPV”, “DL”, “NHL”, and “SN” stand respectively for “Machine Learning”, “Supervised”, “Deep Learning”, “Number of Hidden Layer”, and “Scheme Number”.

No.	ML	SPV	DL	NHL	Input(s)	SN
1	LST-LGA	No	No	-	T2-FLAIR	1
2	SVM	Yes	No	-	T2-FLAIR & T1-W	2,3
3	RF	Yes	No	-	T2-FLAIR & T1-W	4,5
4	DBM	Yes	Yes	2	T2-FLAIR	6
5	CNN	Yes	Yes	5 or 8	T2-FLAIR & T1-W	7-24

Table 3.5: Experiment results reporting DSC and AUC-PR measurements. “one” in the scheme’s name refers to one-pathway CNN, and “two” refers to two-pathway CNN. Label “diff” refers to the mean difference between CNN without GSI and CNN-GSI. Automated WMH segmentation is produced by using threshold value of $t = 0.5$. Values in bold are the best score whereas values in italic are the second-best score.

		DSC		DSC postprocessing			AUC-PR	
No.	**Scheme’s Name**	**mean**	**diff.**	**mean**	**diff.**	**SD**	**mean**	**SD**
1	LST-LGA (Schmidt et al., 2012a)	0.2921	-	0.2963	-	0.1620	0.0942	0.0682
2	SVM_FLAIR	0.0855	-	0.0891	-	0.1266	0.1698	0.1203
3	SVM_FLAIR_T1W	0.1148	-	0.1194	-	0.1036	0.1207	0.0958
4	RF_FLAIR	0.1516	-	0.1621	-	0.1464	0.4126	0.1671
5	RF_FLAIR_T1W	0.1589	-	0.1633	-	0.1513	0.3624	0.1767
6	DBM_FLAIR	0.3152	-	0.3264	-	0.1425	0.3188	0.1592
7	CNN_one_FLAIR_T1W (5-layer)	0.4332	-	0.5118	-	0.1519	0.5248	0.1838
8	CNN_one_FLAIR_T1W_GSI-xyz (5-layer)	0.4570	2.36%	0.5125	0.07%	0.1489	0.5498	0.1846
9	CNN_one_FLAIR_T1W_GSI-xyz-rad (5-layer)	0.4524	1.92%	0.5150	0.32%	0.1476	0.5485	0.1795
10	CNN_one_FLAIR_T1W	0.4601	-	0.5178	-	0.1417	0.5418	0.1737
11	CNN_one_FLAIR_T1W_GSI-xyz	0.4789	1.87%	0.5227	0.49%	0.1474	0.5548	0.1777
12	CNN_one_FLAIR_T1W_GSI-xyz-rad	0.4738	1.37%	0.5230	0.52%	0.1508	0.5566	0.1761
13	CNN_two_FLAIR (5-layer)	0.4843	-	0.5226	-	0.1538	0.5673	0.1824
14	CNN_two_FLAIR_GSI-xyz (5-layer)	0.4987	1.45%	0.5268	0.42%	0.1517	0.5738	0.1820
15	CNN_two_FLAIR_GSI-xyz-rad (5-layer)	0.4984	1.41%	0.5273	0.47%	0.1542	0.5767	0.1831
16	CNN_two_FLAIR	0.4842	-	0.5287	-	0.1486	0.5716	0.1724
17	CNN_two_FLAIR_GSI-xyz	0.4856	0.14%	0.5305	0.18%	0.1507	0.5637	0.1770
18	CNN_two_FLAIR_GSI-xyz-rad	*0.5174*	*3.33%*	0.5307	0.20%	0.1485	**0.5872**	**0.1754**
19	CNN_two_FLAIR_T1W (5-layer)	0.5051	-	0.5333	-	0.1505	0.5676	0.1869
20	CNN_two_FLAIR_T1W_GSI-xyz (5-layer)	0.5090	0.39%	0.5348	0.15%	0.1530	0.5768	0.1891
21	CNN_two_FLAIR_T1W_GSI-xyz-rad (5-layer)	0.5129	0.78%	0.5381	0.48%	0.1500	0.5778	0.1869
22	CNN_two_FLAIR_T1W	0.4972	-	0.5359	-	0.1434	0.5764	0.1773
23	CNN_two_FLAIR_T1W_GSI-xyz	0.5147	1.75%	**0.5390**	**0.31%**	**0.1437**	0.5806	0.1796
24	CNN_two_FLAIR_TW1_GSI-xyz-rad	**0.5159**	**1.87%**	*0.5389*	*0.30%*	*0.1436*	*0.5815*	*0.1831*

3.6.1 Conventional machine learning vs. deep learning

Generally, deep learning algorithms (i.e., DBM and CNN) performed better than conventional machine learning algorithms (i.e., SVM and RF). In the experiments, SVM's performance was low in both AUC-PR and DSC while RF's performance was a lot better than SVM in AUC-PR. On the other hand, DBM's performance was a lot better than SVM and RF, especially in DSC, even though DBM used the same ROI with SVM and RF. These results suggest that a simple DBM architecture (i.e., 2-hidden layer) is still more powerful than SVM and RF in WMH segmentation. However, in this study, CNN outperformed all other methods with much better AUC-PR and DSC values.

3.6.2 LST-LGA vs. other methods

Interestingly, the average DSC value for the LST-LGA (with $\kappa = 0.05$) was higher than that for SVM and RF. However, the AUC-PR for LST-LGA was the lowest from all methods. A low value of AUC-PR means that the algorithm failed to detect subtle hyperintensities, even though the κ-value parameter used in the experiment for LST-LGA is recommended as the most sensitive one.

3.6.3 Impact of using multiple MRI sequences

In general, segmentation results improved when additional information (i.e., MRI sequence/channel) was added, especially in DSC. Improvement in AUC-PR was not always seen, as adding T1-W in SVM/RF decreased the value of AUC-PR (Table 3.5 Scheme No. 2-5). However, AUC-PR always increased for CNN when both sequences were used although the improvement was very subtle (i.e., 0.02% and 0.48% in Scheme No. 13 vs. No. 19 and Scheme No. 16 vs. No. 22 respectively).

3.6.4 Impact of incorporating GSI into CNNs

The use of synthetic GSI sequences improved the performance of CNNs in all cases with variations in the level of improvement, both in AUC-PR and DSC (Table 3.5). The least improvement occurred in Scheme No. 17 (i.e., 0.14% DSC improvement) while the highest improvement happened in Scheme No. 18 (i.e., 3.33% DSC improvement). Similar improvement was also seen after postprocessing: from 0.07% to 0.52% in DSC measurement. Two different architectures of CNNs (i.e., 5-layer network and 8-layer network) and different input of MRI sequences were deliberately tested in different

experiments to see whether the improvements could be observed in different cases. With the same intention, one-pathway (i.e., normal pathway) CNN was also evaluated (Scheme No. 7-12). All improvements listed in Table 3.5 (see label "diff") were tested using the paired two-sided Wilcoxon signed rank, and all of them were improving significantly with $p \leq 0.00015$.

General improvement of incorporating GSI into CNNs can be seen in Figure 3.6, which shows the curve of average DSC score produced by using different threshold values. Figure 3.6 shows that better performances in spatial agreement (DSC) can be produced by using different threshold values while the effect of GSI on CNN becomes smaller. It also shows that there is a limit of improvement that can be given by GSI to the CNN, especially in higher threshold values. However, it is worth mentioning that choosing the best threshold value for the best performance for all subjects is not practical (i.e., each patient has its own optimum threshold value). Furthermore, the best threshold value in a dataset might not work in different dataset due to different data acquisition protocols. In this study, the threshold value of $t \geq 0.5$ was chosen because it is the standard threshold value for two-class segmentation task (i.e., the probability of being one class is higher than the probability of being another class).

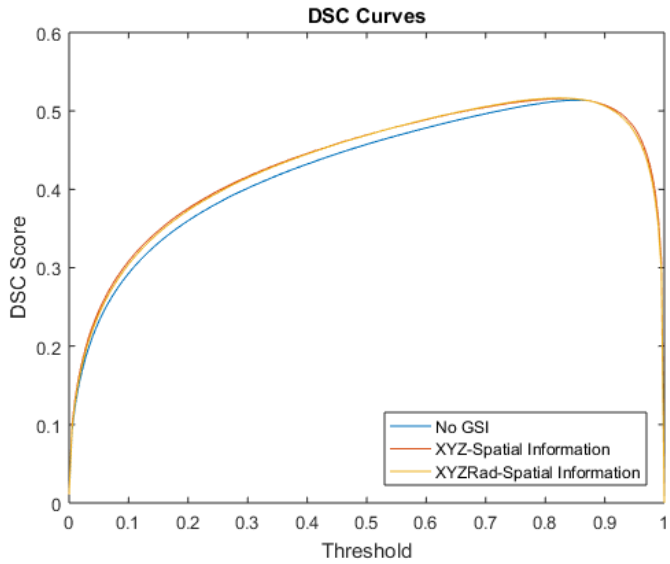


Figure 3.6: Average DSC score curve produced by using different threshold values where general improvement of incorporating GSI into CNN on WMH segmentation can be seen.

Interestingly, the impact of adding GSI into the CNNs was greater than adding an MRI sequence (i.e., T1-W) into the CNNs, especially in AUC-PR values. Adding

T1-W to Scheme No. 13 only improved AUC-PR from 0.5673 to 0.5676 (i.e., 0.03% improvement), whereas adding GSI to the same scheme improved AUC-PR up to 0.5767 (i.e., 0.94% improvement). Similarly happened adding T1-W to Scheme No. 16: AUC-PR only improved from 0.5716 to 0.5764 (i.e., 0.48% improvement). Whereas, adding GSI to the same scheme improved AUC-PR up to 0.5872 (i.e., 1.56% improvement).

Additional evaluation of Fazekas scores to the unlabelled MRI data in the second dataset was done using Spearman's correlation. The r-value indicates the strength in the correlation (i.e., variable $-1 \leq r \leq 1$ is used to describe *monotonic* relationship between paired data), and p-value indicates significance. As shown in Table 3.6, WMH volumes produced by CNNs with GSI correlated better with the corresponding total Fazekas score than the ones produced by CNNs without GSI. As a comparison, a preliminary experiment in the first dataset showed that Spearman's correlation between total Fazekas scores and the manual reference WMH segmentation was $r = 0.7385$ ($p < 0.0001$) and considered as the upper bound measurement of this experiment.

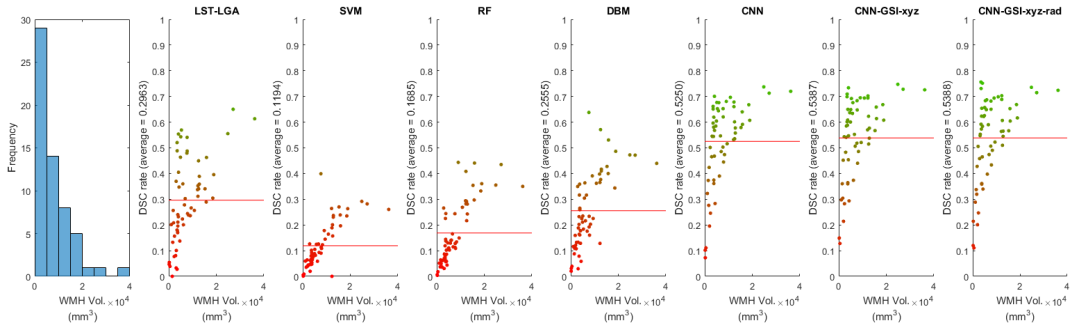


Figure 3.7: DSC values of automatic WMH segmentation in relation to the volume of WMH for each patient based on automated WMH segmentation done by using LST-LGA (Scheme No. 1), SVM (Scheme No. 3), RF (Scheme No. 5), DBM (Scheme No. 6), CNN without GSI (Scheme No. 22), and CNN-GSI (Schemes No. 23 and 24). Each dot represents one patient and its colour refers to its DSC value: red for low DSC, green for high DSC. The x axis indicates volume of WMH from the GT (given in mm^3) for each patient, whereas y indicates the correspondent DSC value. Red horizontal line indicates the mean of DSC values.

3.6.5 Influence of WMH burden

DSC and AUC-PR measurements in this study are low partly because almost half of MRI data have very small WMH burden (i.e., volume of WMH in one patient). The

Table 3.6: Spearman's correlation coefficient between WMH volume of MRI data automatically produced by CNNs and visual rating Fazekas score. Higher r-value and lower p-value are better.

		Corr. val.	
No.	**Scheme's Name**	***r*-value**	***p*-value**
1	CNN without GSI	0.4275	1.92E-72
2	CNN with X, Y, and Z GSI (CNN-GSI-xyz)	0.4341	7.00E-75
3	CNN with X, Y, Z, and radial GSI (CNN-GSI-xyz-rad)	0.4367	7.66E-76
4	CNN_one_FLAIR_T1W (5-layer)	0.3626	9.45E-10
5	CNN_one_FLAIR_T1W_GSI-xyz (5-layer)	0.3631	8.96E-10
6	CNN_one_FLAIR_T1W_GSI-xyz-rad (5-layer)	0.3779	1.60E-10
7	CNN_one_FLAIR_T1W	0.3816	1.02E-10
8	CNN_one_FLAIR_T1W_GSI-xyz	0.3894	3.92E-11
9	CNN_one_FLAIR_T1W_GSI-xyz-rad	0.3818	9.91E-11
10	CNN_two_FLAIR (5-layer)	0.4479	1.25E-14
11	CNN_two_FLAIR_GSI-xyz (5-layer)	0.4831	4.49E-17
12	CNN_two_FLAIR_GSI-xyz-rad (5-layer)	0.4981	3.26E-18
13	CNN_two_FLAIR	0.4864	2.54E-17
14	CNN_two_FLAIR_GSI-xyz	0.4865	2.51E-18
15	CNN_two_FLAIR_GSI-xyz-rad	**0.5104**	**3.51E-19**
16	CNN_two_FLAIR_T1W (5-layer)	0.4344	9.19E-14
17	CNN_two_FLAIR_T1W_GSI-xyz (5-layer)	0.4312	1.47E-13
18	CNN_two_FLAIR_T1W_GSI-xyz-rad (5-layer)	0.4369	6.39E-14
19	CNN_two_FLAIR_T1W	0.4691	4.55E-16
20	CNN_two_FLAIR_T1W_GSI-xyz	0.4702	4.48E-17
21	CNN_two_FLAIR_TW1_GSI-xyz-rad	0.4713	4.46E-17

volume of WMH can be calculated by multiplying the number of manual/predicted WMH voxels in one patient with the real-world voxel size (explained in Section 3.5.3). From Figure 3.7, it can be easily observed where all schemes evaluated performed better on brains with medium and high load of WMH, including the LST-LGA toolbox. Segmentation of small WMH was the most challenging. The DSC measurements for scans with small burden of WMH were low in most of machine learning algorithms except for deep learning algorithms, especially the CNNs, which performed much better than the others. Furthermore, it is also important to see in the right-side of the Figure 3.7 how incorporating GSI into CNN can push the dots to the top of the graphs, which means better performance of the CNN. Please note that CNN schemes depicted in Figure 3.7 are Schemes No. 22-24.

Table 3.7: Five groups of MRI data based on WMH volume.

No.	Group	Range of WMH vol. (mm^3)	Number of MRI Data
1	Very Small	[0, 1500]	5
2	Small	(1500, 4500]	22
3	Medium	(4500, 13000]	24
4	Large	(13000, 24000]	5
5	Very Large	$>$ 24000	3

Table 3.8: Average values of DSC, AUC-PR, and VD for grouped MRI data based on its WMH burden listed in Table 3.7. VS, S, M, L and VL stand for "Very Small", "Small", "Medium", "Large", and "Very Large" which are names of the groups. Average values listed below are directly corresponded to Figure 3.8. Bigger values of DSC and AUC-PR are better while VD value closer to zero is better. Values in bold are the best score whereas values in italic are the second-best score.

		DSC (mean)					AUC-PR (mean)					VD (mean)				
No.	**Scheme**	**VS**	**S**	**M**	**L**	**VL**	**VS**	**S**	**M**	**L**	**VL**	**VS**	**S**	**M**	**L**	**VL**
1	LST-LGA	0.0699	0.2867	0.3106	0.2992	0.6038	0.0140	0.1214	0.1153	0.1488	0.2076	4.1536	0.5921	0.2343	0.5448	-0.3404
2	SVM	0.0250	0.1091	0.1111	0.1753	0.2714	0.0186	0.1020	0.1311	0.1625	0.3017	124.2099	33.6717	11.9839	5.6529	2.8556
3	RF	0.0200	0.1452	0.1599	0.2735	0.3645	0.1703	0.3204	0.3961	0.4890	0.6448	121.31	32.9548	12.3818	4.3804	2.6595
4	DBM	0.0481	0.2423	0.2617	0.3892	0.4474	0.2061	0.3363	0.3616	0.4454	0.3251	47.3302	12.9548	4.8097	1.6414	0.3066
5	CNN	0.1599	0.4461	0.5262	0.5590	0.7187	**0.3187**	**0.5014**	0.6150	0.6358	0.7998	22.8059	6.0561	1.5364	0.9259	**-0.0155**
6	CNN-GSI-xyz	**0.1826**	*0.4596*	*0.5409*	*0.5837*	**0.7292**	*0.2959*	0.4922	*0.6239*	*0.6479*	*0.8154*	*15.7424*	*4.0804*	*1.4157*	**0.7298**	*0.0369*
7	CNN-GSI-xyz-rad	*0.1775*	**0.4623**	**0.5483**	**0.5849**	*0.7230*	0.2687	*0.5011*	**0.6302**	**0.6517**	**0.8161**	**14.7669**	**3.9256**	**1.3713**	*0.7697*	-0.0423

For clarity, in this analysis, all MRI data were divided into 5 different groups based on WMH volume (Table 3.7) and plotted the DSC and AUC-PR values in two separate boxplots (Figure 3.8). Note that the grouping of the dataset into "Very Small", "Small", "Medium", "Large", and "Very Large" groups is similar to (Brosch et al., 2016). Seven different schemes were plotted in Figure 3.7: LST-LGA (Scheme No. 1), SVM (Scheme No. 3), RF (Scheme No. 5), DBM (Scheme No. 6), CNN (Scheme No. 22), CNN-GSI-xyz (Scheme No. 23), and CNN-GSI-xyz-rad (Scheme No. 24). From Figure 3.8, it can be seen that GSI, both three axes and radial spatial information, helped to improve CNN's performance. This marks one of the purposes of this study: to improve WMH segmentation in brain MRI data from subjects with small WMH burden. Full report of average values from DSC, AUC-PR and VD measurements from grouped evaluation can be seen in Table 3.8: adding GSI improved CNN's performance up to 2.27% in the "Very Small" group and gives an overall similar rate of improvement in other groups.

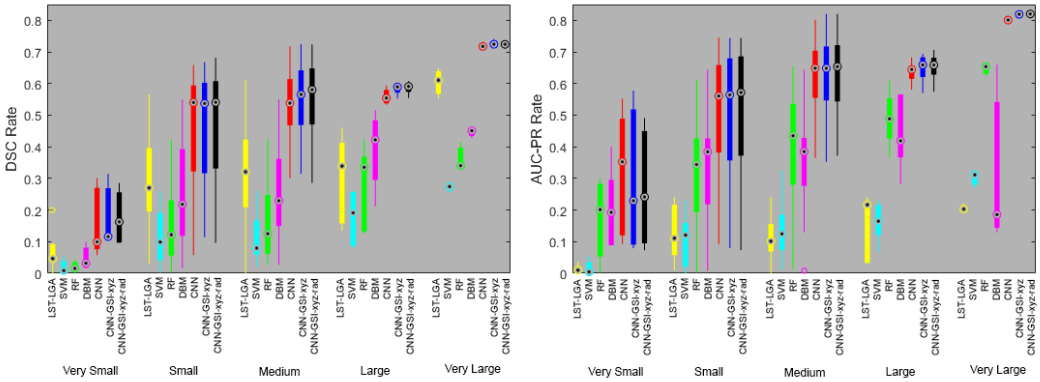


Figure 3.8: Comparison of WMH segmentation accuracy (i.e., in DSC and AUC-PR) using boxplot where all MRI data is grouped together based on its WMH burden for seven different schemes: LST-LGA (Scheme No. 1), SVM (Scheme No. 3), RF (Scheme No. 5), DBM (Scheme No. 6), CNN (Scheme No. 22), CNN-GSI-xyz (Scheme No. 23), and CNN-GSI-xyz-rad (Scheme No. 24). Criteria of each group are listed in Table 3.7, and the mean values for each scheme in each group are listed in Table 3.8.

3.6.6 Visual evaluation of the WMH segmentation results

Some visual examples of results from automatic WMH segmentation without postprocessing can be seen in Figure 3.9. In the figure, three axial slices of MRI data from three different subjects with different WMH volumes are presented. Raw segmentation results from Scheme No. 1 (LST-LGA), 3 (SVM), 5 (RF), 6 (DBM), 22 (CNN), and 23 (CNN-GSI-xyz) are presented to visually appreciate differences in performance. From the figure, it can be seen that the use of deep learning (i.e., DBM and CNN) made automatic segmentation results cleaner than SVM and RF, which have many false positives. How WMH volume affected the performance of each automatic WMH segmentation scheme can also be appreciated. In general, CNNs were more sensitive and precise than the other algorithms tested in this study.

To better appreciate the difference in performance between CNN and CNN-GSI (i.e., Scheme No. 22 and 23), the panels from Figure 3.9 were zoomed-in to Figure 3.10. GSI improved CNN's performance eliminating small false positives, which are pointed by yellow arrows, and correctly segmenting WMH in some cases, pointed by green arrows (as also seen in Table 3.5). Furthermore, also from Figure 3.10, it can be seen that the DSC of Subjects 1 and 3 improved considerably (i.e., 7.58% and 7.99% improvements). However, in the presence of extensive "dirty white matter", the introduction of GSI

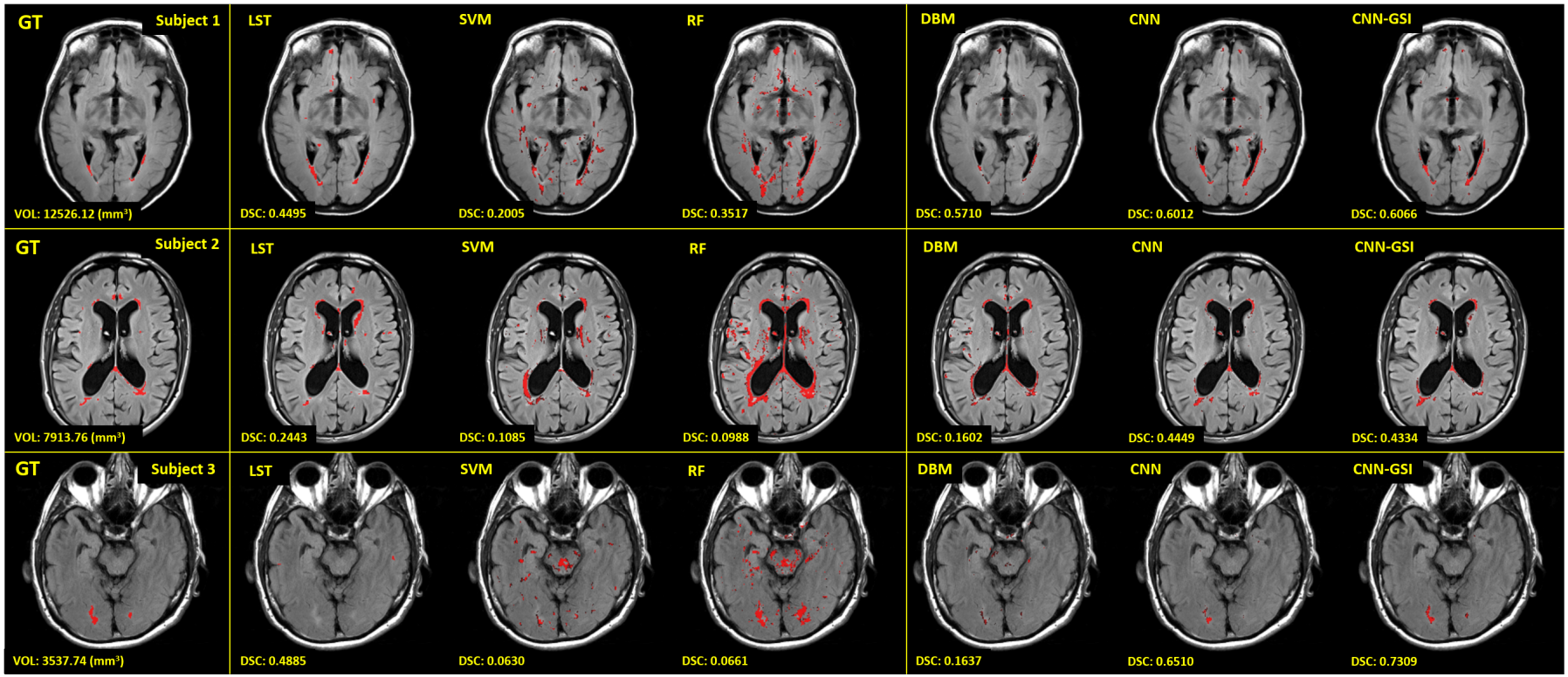


Figure 3.9: Visualisation of automatic WMH segmentation results (before postprocessing) from selected schemes of each algorithm (i.e., LST-LGA, SVM, RF, DBM, CNN, and CNN-GSI) and public toolbox LST-LGA. Red regions are WMH labelled by experts (GT) or machine/deep learning algorithms. Three different subjects with very different WMH burden are visualised to see how the WMH volume affects the performance of machine/deep learning algorithms. Volume of WMH and value of the DSC measurement for each algorithm are at the bottom left on each respective image.

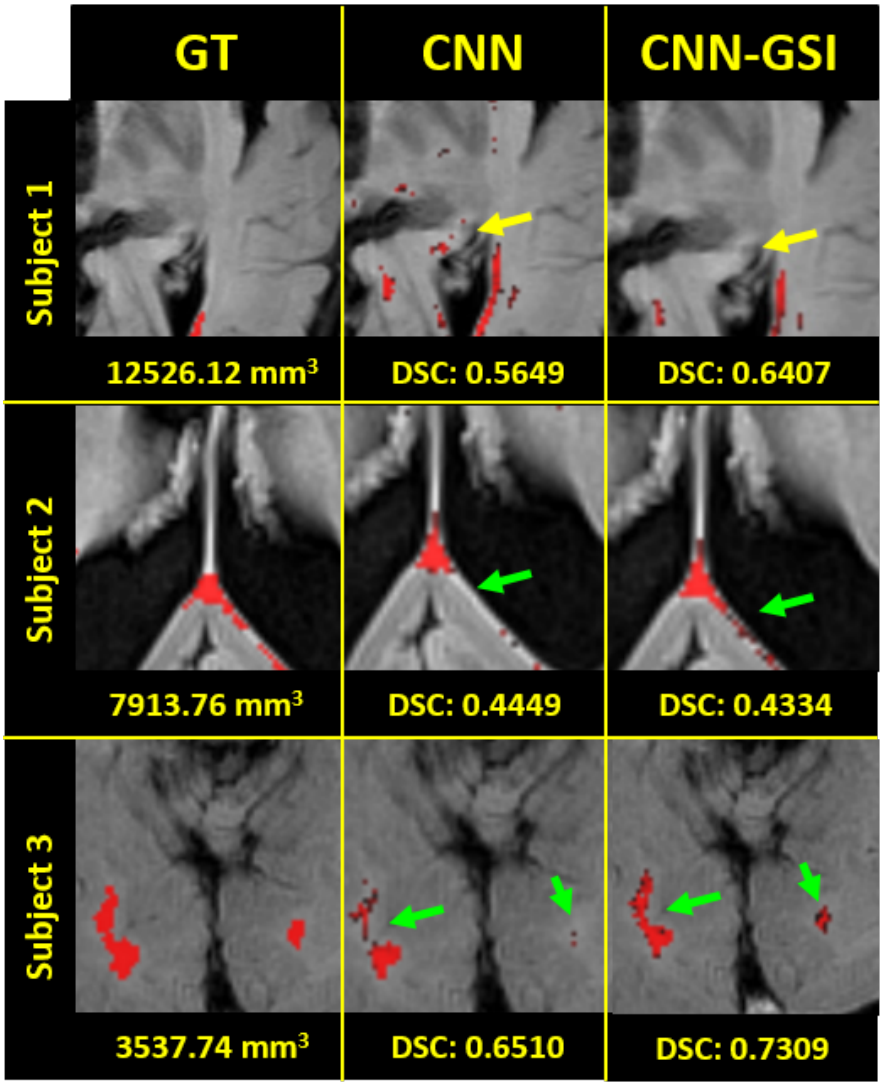


Figure 3.10: Close-up image of sections from selected cases showing WMH segmentation results from the CNN and CNN-GSI schemes. From left to right: GT, CNN (Scheme No. 22) and CNN-GSI (Scheme No. 23). The arrows indicate false positives which disappear (yellow) and true positives which appear (green) due to the use of GSI in CNN. Note that these are visualisations before postprocessing.

slightly decreased CNN's performance as shown in Subject 2, as many non-WMH regions appear very similar to WMH. This particular case can be observed more closely in Figure 3.9 by comparing CNN results with the GT.

Table 3.9: DSC scores for longitudinal test, VD for cross validation (CV) and longitudinal (Long.) experiments, and percentage of volumetric disagreement measurement[11](D) between automated scheme and multiple human observers (i.e., intra-/inter-observation) for LST-LGA, SVM, RF, DBM, CNN, CNN-GSI-xyz and CNN-GSI-xyz-rad (i.e., Scheme No. 1, 3, 5, 6, 22, 23 and 24 in Table 3.5 respectively). Captions of "[Intra]" and "[Inter]" refer to intra- and inter-observer evaluation. Higher DSC value is better, lower VD value is better, and value of D close to zero is better. Values in bold are the best score whereas values in italic are the second-best score.

No	Scheme	DSC Long.		VD (mean)		D of Observer #1 [Intra] (%)				D of both observers [Inter] (%)			
		mean	SD	CV	Long.	Label #1	SD	Label #2	SD	Obs. #1	SD	Obs. #2	SD
1	LST-LGA	-	-	0.6647	-	67.64	32.30	77.48	45.15	60.59	41.58	49.89	42.37
2	SVM	0.1478	0.1117	9.1551	4.0259	131.38	48.41	136.77	52.50	61.01	52.42	66.60	41.46
3	RF	0.1816	0.1517	15.857	11.260	140.13	43.28	147.76	41.62	123.72	49.07	112.70	50.12
4	DBM	0.3054	0.1513	1.5460	**0.1029**	78.05	50.26	94.58	60.08	75.63	38.19	65.11	48.66
5	CNN	0.5982	0.1410	*0.2541*	-0.1883	*38.92*	*32.79*	*63.87*	*60.57*	*33.18*	*38.48*	*35.01*	*36.62*
6	CNN-GSI-xyz	*0.6063*	*0.1411*	**0.2275**	-0.1997	**36.92**	**31.98**	**61.55**	**60.97**	**31.80**	**36.38**	**34.41**	**36.28**
7	CNN-GSI-xyz-rad	**0.6046**	**0.1512**	0.3304	*-0.1652*	41.62	34.47	64.55	60.88	36.03	36.50	42.56	40.38

3.6.7 Volumetric disagreement and intra-/inter-observer reliability analysis

VD (Equation (3.12)) evaluates WMH volume differences between manually segmented WMH (GT) and automatically segmented WMH. This analysis is clinically important if the WMH burden of one patient is to be expressed by the WMH volume. However, different observers can annotate WMH differently and one observer might give different opinion in the reassessment of the same data. Intra-/inter-observer reliability analysis can be done to evaluate the confidence level of the labels by using D measurement (Equation (3.13)). Intra-observer analysis evaluates agreement and reliability of multiple measurements generated by one human observer whereas inter-observer analysis evaluates agreement and reliability of measurements from multiple observers. The intra-observer D given by the percentage difference between measurements with respect to the average value of both for Observer #1 was 36.06% (SD 52.21%) whilst for Observer #2 it was 4.22% (SD 24.02%). The inter-observer D (i.e., between Observer #1 and #2) was 28.03% (SD 57.25%). These results mean that Observer #2 produced more consistent labels of WMH than the Observer #1 while the variations of Observer #1 is similar to the variations between Observer # 1 and Observer #2. The high level of intra-/inter-

[11]For clarity in the presentation of the agreement with the human observers, standard deviation (SD) values are given instead of 95% confidence intervals. Label #1 and Label #2 correspond to the two sets of measurements from Observer #1.

observer difference could be caused due to small and subtle WMH, which are abundant features in the first dataset. From preliminary observation and internal documentation, it has been observed that these types of WMH could be easily misidentified as artefacts (or vice versa), especially the ones located around insular cortex, midline parasagittal cortex, anterior temporal poles, inferior walls of the third ventricle, posterior thalamus, and periaqueductal regions. The inconsistencies in intra-/inter-observer experiment show the biggest shortcoming of supervised machine learning algorithms, especially when WMH labels from one expert are used exclusively in the training process. To reduce the inconsistencies, one study suggested to use high-field MR scan (i.e., 3T or over instead of 1.5T used in this study) that is more sensitive to small and early WMH and more MRI sequences (e.g., T2-W, DTI, positron density (PD), and magnetization transfer image (MTI)) for better characterisation of WMH (Kim et al., 2008).

VD and intra-/inter-observer analysis of seven learning algorithms (i.e., Scheme No. 1, 3, 5, 6, 22, 23, and 24 of Table 3.5) are shown in Table 3.9. VD rate of CNN-GSI (i.e., CNN-GSI-xyz) in the cross validation experiment is better than CNN without GSI (i.e., 0.2275 and 0.2541 respectively) and is the best performer in terms of VD. In the longitudinal test using the same measurement (VD), the performance of CNN-GSI (i.e., CNN-GSI-xyz-rad) is better than CNN without GSI. With regards to volumetric D against intra-/inter-observer reliability measurements, CNN-GSI (i.e., CNN-GSI-xyz) always performed better than SVM, RF, DBM, and CNN without GSI. This means that spatial XYZ information boosted the performance of CNNs according to VD, D, and DSC measurements.

3.6.8 Longitudinal evaluation

This evaluation aims to determine the schemes' performance in estimating the WMH regions in the two years following the baseline scan, providing that the baseline measurements are known. Hence, i.e., 1^{st} year samples are used for training and the rests of years are used for testing. Table 3.9 lists the DSC and VD measurements in longitudinal test for schemes No. 1, 3, 5, 6, 22, 23, and 24 (i.e., LST-LGA, SVM, RF, DBM, CNN, CNN-GSI-xyz, and CNN-GSI-xyz-rad respectively) listed in Table 3.5. From the table, the incorporation of four types of GSI improved the performance of CNN (i.e., 0.6046 compared to 0.5982 of CNN without spatial information). Furthermore, the incorporation of XYZ spatial information also improved CNN's performance, with DSC of 0.6063. In summary, these results (i.e., listed in Table 3.5, Table 3.8, and Table

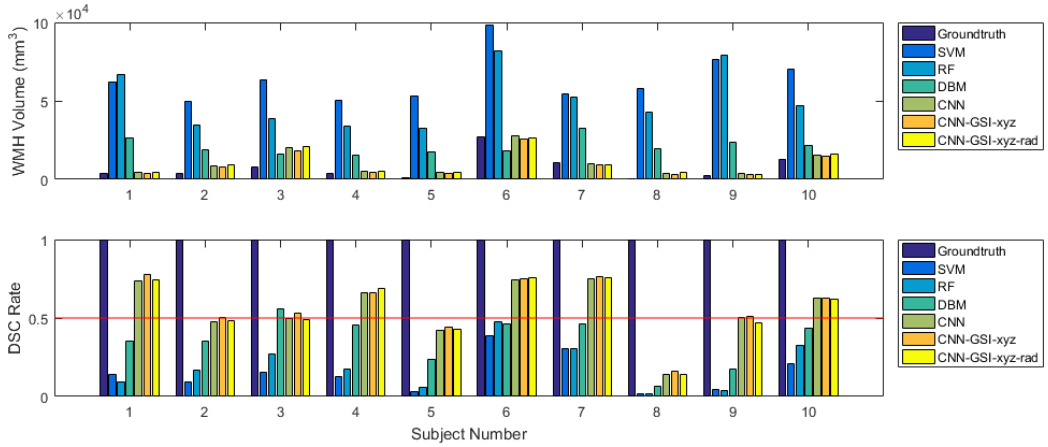


Figure 3.11: Results of the longitudinal evaluation for 10 random subjects where first year data is used as training data and second year data is used as testing data (shown in the charts). The upper chart presents the WMH volume (mm^3) of the GT and produced by the automatic WMH segmentation schemes, and the lower one presents the DSC values for the machine learning algorithms. See Figure 3.9 for reference of the schemes represented.

3.9) show that GSI successfully improved the performance of the CNNs. Figure 3.11 shows the WMH volumes and DSC rates obtained for 10 random subjects from several schemes, for schemes trained with data from the previous year. It can be seen that conventional machine learning algorithms (i.e., SVM and RF) produced low agreement of WMH volume and location while GSI improved CNN's performance in both WMH volume and location agreements.

3.6.9 Processing time

The processing time needed by each algorithm in training and testing processes was also evaluated. The results of this evaluation are shown in Table 3.10. Note that SVM, RF, and DBM used Central Processing Units (CPUs), and were run from a workstation in a Linux server with 32 Intel(R) Xeon(R) CPU E5-2665 @ 2.40GHz processors. Whereas, CNNs used General Processing Units (GPUs) and were run in a Linux Ubuntu desktop with Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz and EVGA NVIDIA GeForce GTX 1080 8GB GAMING ACX 3.0. Based on the evaluation, SVM was the fastest algorithm in the training process, but it was the slowest one in testing. On the other hand, CNNs were faster than DBM in training and the fastest in testing.

Table 3.10: Processing time of each algorithm in training phase and testing phases. Times are given in minutes and seconds respectively.

Algorithm	Training (minutes)	Testing one MRI data (seconds)
SVM	25.4589	82.4877
RF	36.4649	40.6431
DBM	1340.3209	16.9841
CNNs	317.8757	9.1879

3.6.10 Clinical plausibility of the results

Despite WMH have been found to be associated with hypertension, hypercholesterolaemia, and several vascular risk factors (Longstreth et al., 1996), their dynamic progression in short term has only been reported associated with their extent at certain time point considered the baseline measurement (Ramirez et al., 2016). In the ANCOVA models described in Section 3.5.3 that used the GT WMH volume, in agreement with clinical reports, the only predictor of the WMH volume one-year or two-year later was the WMH volume considered baseline on each model ($p < 0.0001$ in all cases). When these models were repeated but using the WMH volume obtained from all schemes evaluated, the results were not different.

Visual inspection of the results revealed that conventional machine learning methods do not distinguish T2-FLAIR hyperintense cortical sections well from subtle WMH as Figure 3.9 shows. Deep learning algorithms, on the other hand, correctly classify most of intense or obvious WMH, while misclassifying subtle white matter changes (i.e., pale WMH) in some cases. The fact that all schemes produced results clinically plausible (i.e., in agreement with published recent clinical reports) perhaps may be indicative that all T2-FLAIR hyperintensities, regardless of their location and relative intensity, may be part of a more generalised phenomena worth to be explored on a bigger sample.

3.6.11 Neuroradiological evaluation

Unlike image analysts (i.e., Observer #1 and #2) who measured WMH by delineating the boundary of WMH, in this study, the neuroradiologist evaluated the results from the six automated schemes that produced the best results on one scan (out of the three annual scans) per patient. As mentioned in Section 3.5.3, the neuroradiologist performed the evaluation by filling in a form which records the number of WMHs not identified,

missed partially, and misclassified in specific anatomical brain regions (completed forms can be seen in Appendix A). This evaluation was done to help regularising the location and cause of the misclassified/missed WMH partially or totally as well as to find out the effect of GSI on CNN from the point of view of a neuroradiologist. This evaluation is also useful to devise future improvement strategies. The automated schemes evaluated by the neuroradiologist were 5-layer dual-modality CNN with and without GSI incorporated (Schemes No. 19-21 in Table 4.1) and 8-layer dual-modality CNN with and without GSI incorporated (Schemes No. 22-24 in Table 4.1).

The neuroradiologist considered "missing" an average of 2 WMH clusters in the anterior white matter (i.e., white matter in the frontal and parieto-frontal lobes) on only 7/20 datasets (subjects). Of the WMH clusters correctly identified, the neuroradiologist did not consider relevant the differences in the extent of the clusters marked by any scheme. Therefore, no "WMH partially missed" were recorded. False positives were: artefacts in the pons, corpus striatum, deep white matter, and anterior cortex, on an average of 5 WMH clusters in total per patient. All schemes evaluated by the neuroradiologist were considered with "similar performance". These results indicate that GSI did not give negative impact to the CNN as per the neuroradiologist's visual observation, but at the same time GSI also did not give noticeable positive impact either. This is reasonable because, as per Table 3.8, GSI gives positive impact to the very small and small WMH which are easily missed by human observers. This also indicates that human observers easily overlook very small and small clusters of WMH in MRI.

3.7 Conclusion

Conventional machine learning algorithms evaluated in this study, SVM and RF, did not perform well on automatic WMH segmentation across the sample used in this study. The addition of the T2-W image to the T2-FLAIR and/or T1-W (i.e., the use of three structural MRI sequences instead of one or two) could increase the certainty of WMH delineation and reduce false positives. The experiments show that deep learning algorithms performed much better than the conventional ones for automatic WMH segmentation. Lastly, GSI set, which is incorporated into CNN's convolutional layer, successfully helps the performance of CNN in every CNN's schemes and tests done in this study especially in spatial agreement measurement (DSC) evaluations.

3.8 Future Work

The texture, shape, and prominence of WMH differ according to their anatomical location and are related to the overall "damage" of a particular brain, reflected on the presence of other indicators of small vessel disease (Wardlaw et al., 2013). Therefore, the best performing approach in this study, which is CNN-GSI, needs to be evaluated in brains with moderate to abundant vascular pathology (i.e., small vessel disease, strokes). Other types of GSI such as brain's landmark or tissue priors probability maps can be investigated. Different approaches of incorporating GSI into the CNN like in (Ghafoorian et al., 2017a), where GSI is incorporated in the segmentation layer, can also be evaluated. Different deep neural network architectures, like the autoencoder could be promising. Further study to increase the performance of automatic WMH segmentation schemes on brains with heterogeneous WMH load and appearance, and with images acquired with different acquisition protocols is needed.

Chapter 4

Quantitative Assessment of WMH using Irregularity Map

In this chapter, a novel unsupervised method called Limited One-time Sampling Irregularity Map (LOTS-IM) is described, tested, and evaluated for WMH segmentation. This chapter is based on the following publications:

1. Rachmadi, M. F., Valdés-Hernández, M. D. C., Li, H., Guerrero, R., Meijboom, R., Wiseman, S., Waldman, A., Zhang, J., Rueckert, D., Wardlaw, J., and Komura, T. (2020). Limited One-time Sampling Irregularity Map for automatic unsupervised assessment of white matter hyperintensities and multiple sclerosis lesions in structural brain magnetic resonance images. *Computerized Medical Imaging and Graphics*, 79:101685.

2. Rachmadi, M. F., Valdés-Hernández, M. D. C.,M., and Komura, T. (2018a). Transfer learning for task adaptation of brain lesion assessment and prediction of brain abnormalities progression/regression using irregularity age map in brain MRI. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 85–93, Cham. Springer International Publishing.

4.1 Motivation

Since the widespread use of deep neural network algorithms (i.e. hereinafter referred to as "deep learning") in computer vision, these methods have become the state-of-the-art for detection and segmentation problems in brain MRI. For example, it has been shown in Chapter 3 that deep learning algorithms based on DeepMedic (Kamnitsas

et al., 2017) outperformed the conventional machine learning algorithms (i.e., SVM and RF) on automatic segmentation of WMH. However, as supervised methods, they are highly dependent on manual labels produced by experts (i.e., physicians) for training process. This dependency to expert's opinion limits their applicability due to the expensiveness of manual WMH labels and their limited availability. Furthermore, the quality of manual label itself depends on and varies according to the expert's skill. Section 3.2.3 exhibits this problem clearly where inconsistency of Observer #1 is high in intra-observer reliability test and inconsistency between Observer #1 and #2 is also high in inter-observer reliability test.

Conventional unsupervised segmentation methods, such as LST-LGA (Schmidt et al., 2012a) and Lesion-TOADS (Shiee et al., 2010), do not have the aforementioned dependencies to segment WMH in brain MRI. Hence, these methods have been tested in many studies and become the standard references to the other segmentation methods. Unfortunately, their performance is usually worse than that of supervised machine learning and deep learning methods. On the other hand, the more recent unsupervised deep learning methods based on GAN (Goodfellow et al., 2014), such as Anomaly GAN (AnoGAN) (Schlegl et al., 2017) and Adversarial Auto-Encoder (AAE) (Chen and Konukoglu, 2018)), need large number of both healthy and unhealthy data for adversarial training processes, usually not easily accessible.

Recently, a new unsupervised segmentation method named Irregularity Age Map (IAM) (Rachmadi et al., 2017c) and its faster version One-time Sampling Irregularity Age Map (OTS-IAM) (Rachmadi et al., 2018c) have been proposed and reported to work better than the state-of-the-art unsupervised WMH segmentation method LST-LGA, the conventional supervised machine learning methods (i.e., SVM and RF), and some deep learning methods of DBM (Salakhutdinov and Larochelle, 2010) and Convolutional Encoder Network (CEN) (Brosch et al., 2016). IAM and OTS-IAM uniquely produce an irregularity map (IM) that has several advantages over deep learning's probability map (PM). Unlike PM, IM captures regular and irregular regions by retaining changes of the original T2-FLAIR intensities. This cannot be achieved with deep neural network algorithms, which are trained to reproduce manually generated binary masks. For example, the gradual changes of hyperintensities along the border of WMH, usually referred to as "penumbra" (Maillard et al., 2011), can be well represented in IM. The penumbra of WMH has been subject of many studies in recent years, which debate criteria to correctly identify the WMH borders (Firbank et al., 2003; Jeerakathil et al., 2004; Valdés Hernández et al., 2010). Further discussion of WMH penumbra is

described in Section 2.1.

While IAM and OTS-IAM have been tested in previous studies and produced very good results in the segmentation of WMH in MRI scans from individuals with minor vascular pathology, they had one main limitation: their lengthy computing time. The most recent OTS-IAM takes 13 minutes (on GPU) to 174 minutes (on CPU) for processing a single MRI scan data of $256 \times 256 \times 35$ voxels in average. The aforementioned computation times are not ideal especially if thousands of MRI are to be processed.

In this study, a new version of IAM method called Limited OTS-IM (LOTS-IM) is proposed. LOTS-IM greatly improves the processing time compared to IAM and OTS-IAM without any perceivable quality degradation. This study also documents in more detail the generation of the IM, the method's performance (i.e. including limits of validity), describes and evaluates the internal parameters involved in the computation of the IM, and demonstrates the use of IM for simulating the evolution of abnormalities inside the brain.

4.2 Irregularity Age Map Method

The IAM for WMH assessment on brain MRI was originally proposed in (Rachmadi et al., 2017c) which is based on a computer graphics method developed to synthesise time-varying weathered texture images (Bellini et al., 2016). In the original study, Bellini et al. (2016) proposed a method to calculate the degree (i.e., age) of weathering (e.g., mold and stains on the exterior walls caused by prolong exposure to weather) at different regions of input texture by analysing the prevalence of texture patches. The term "age value" and "age map" were originally used by Bellini et al. (2016) for the 2D array of values between 0 and 1 denoting the weathered regions considered texture irregularities in texture images. In this study, the terms of age value and IAM are changed to "irregularity value" and "irregularity map" (IM) as the concept of detecting "aged/weathered" textural regions no longer applies. In the IM, the closer the value to 1, the more probable a pixel/voxel belongs to a neighbourhood that has different texture from that considered "normal".

After segmenting the regions of interest where the algorithm will work (e.g. brain tissue) using well established fully automatic computational methods (Step 1, depicted in Figure 4.1(A)), IM is calculated from each structural MRI slice (i.e. preferably in axial or coronal orientation) by applying the following steps: patch generation (Step

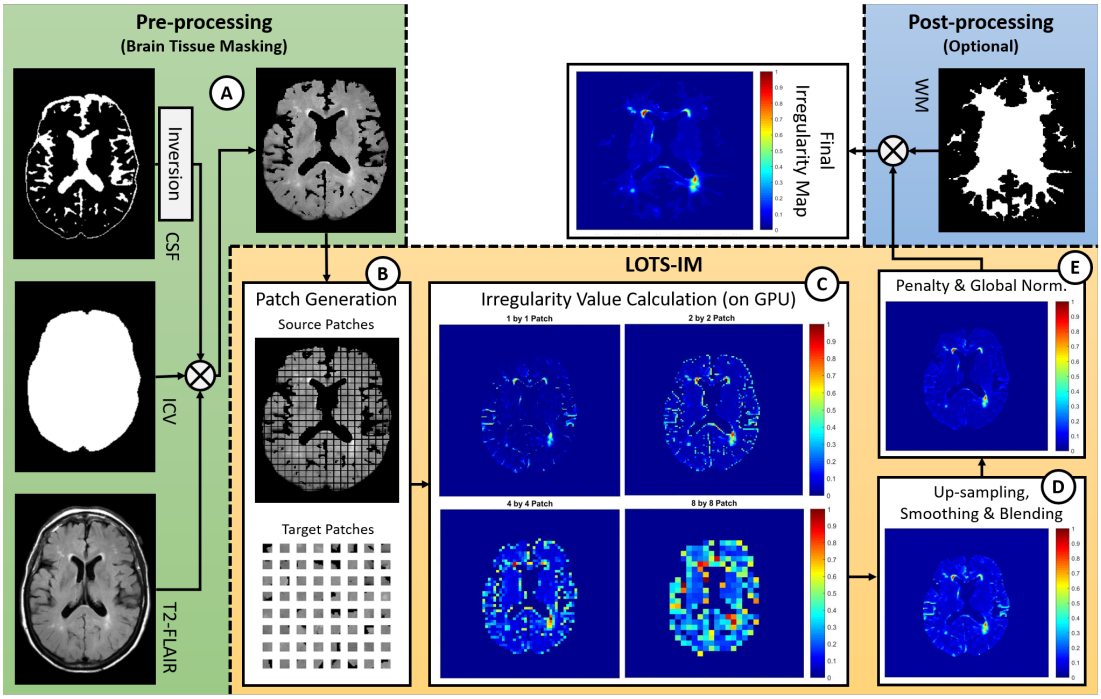


Figure 4.1: Flow of the proposed LOTS-IM. **1) Pre-processing**: brain tissue-only T2-FLAIR MRI 2D slices are generated from the original T2-FLAIR MRI and its corresponding brain masks (i.e., intracranial volume (ICV) and cerebrospinal fluid combined with pial regions (CSF)). **2) LOTS-IM**: the brain tissue-only T2-FLAIR MRI slice is processed through the LOTS-IM algorithm on GPU. **3) Post-processing**: final age map of the corresponding input MRI slice is produced after a post-processing step (optional).

2, depicted in Figure 4.1(B)), irregularity value calculation (Step 3, depicted in Figure 4.1(C)), and final IM generation (Step 4, depicted in Figures 4.1(D) and 4.1(E)). These four steps are described in the rest of this section. Note that steps 2 to 4 are executed slice by slice (i.e., in 2D).

4.2.1 Brain tissue masking

For brain MRI scans, the brain tissue mask is necessary to exclude non-brain tissues which can represent "irregularities" *per se* (e.g., skull, cerebrospinal fluid, veins, and meninges). In other words, brain tissue patches are compared with themselves, not with patches from the skull, other extracranial tissues, or fluid-filled cavities. For this purpose, two binary masks, ICV and CSF masks, are used where the latter containing also blood vessels and pial elements like venous sinuses and meninges. In this study, the ICV mask was generated using optiBET (Lutkenhoff et al., 2014). However,

several tools that produce accurate output exist and can be used for this purpose (e.g., bricBET[1], freesurfer[2]). The CSF mask was generated by using a multispectral algorithm (Valdés Hernández et al., 2015a). The brain tissue masking is schematically represented in Figure 4.1(A).

The pre-processing step before computing LOTS-IM only involves the generation of these two masks as per in the original IAM and OTS-IAM (Rachmadi et al., 2017c, 2018c). Their subtraction generates the brain tissue mask, which is, then, multiplied by the T2-FLAIR volume. This study also uses the NAWM mask in a post-processing step to exclude brain areas in the cortex that could be identified as false positives. The NAWM masks were generated using FSL-FAST (Zhang et al., 2001), but these can also be generated using other tools (e.g., freesurfer).

4.2.2 Patch generation

Similar to IAM (Rachmadi et al., 2017c), LOTS-IM requires the generation of two sets of patches: non-overlapping grid patches called *source patches* and randomly-sampled patches called *target patches*, which can geometrically overlap each other (Figure 4.1(B)). In the IM computation, a source patch is used as reference to the underlying pixel (or patch) while a target patch is used to represent a sample of all possible image textures. A set of target patches is randomly sampled from the same image. Note that the distribution of randomly sampled target patches closely follows the underlying distribution of all target patches, i.e., brain tissues' textures.

Source and target patches are used to calculate the irregularity value, where each of the source patches is compared with several randomly sampled target patches using a distance function (Bellini et al., 2016). This will be discussed in the next subsection. Hierarchical subsets of 2D image array are used where four different sizes of source and target patches, which are 1×1, 2×2, 4×4 and 8×8 pixels. The patch generation process is schematically depicted in Figure 4.1(B).

[1] https://sourceforge.net/projects/bric1936/files/MATLAB_R2015a_to_R2017b/BRIClib/

[2] https://surfer.nmr.mgh.harvard.edu/

4.2.3 Irregularity value calculation

The *irregularity value* calculation is the core of the IM generation process. Let **s** be a source patch and **t** be a target patch, then the distance (d) between **s** and **t** is defined as:

$$d = \text{average}(|\max(\mathbf{s}-\mathbf{t})|,\ |\text{mean}(\mathbf{s}-\mathbf{t})|). \tag{4.1}$$

Based on Equation (4.1) above, the distance between source patch (**s**) and target patch (**t**) can be calculated by averaging the maximum difference and the mean difference between **s** and **t**. The difference between **s** and **t** is calculated by subtracting their intensities pixel wise. The averaging of maximum and mean differences is applied to make the distance value robust against outliers. To capture the distribution of textures in the image (i.e., slice MRI), each source patch is compared against a set of target patches (e.g., 2,048 target patches in (Rachmadi et al., 2018c)) for which the same number of distance values are produced.

The *irregularity value* for a source patch can be calculated by sorting all distance values and averaging the second half of the third quartile (Q_3) of the samples, i.e., outliers. The rationale is simple: the mean of outliers' distance values produced by an irregular source patch is still comparably higher than the one produced by a normal source patch. Also, the mean is chosen as irregularities are compared to the normal-appearing white matter, and normal tissue intensities are known to be normally distributed, although other descriptive statistics (e.g., percentiles) have been identified to discern degree of pathology (Dickie et al., 2015, 2014).

All irregularity values from all source patches are then mapped and normalised to real values between 0 and 1 to create the *IM for one MRI slice* (see Figure 4.1(C)). Lastly, the IM is up-sampled to fit the original size of MRI slice and smoothed using a Gaussian filter as per (Bellini et al., 2016).

4.2.4 Final IM generation

The generation of the final IM consists of three sub-steps: a) blending of the four IMs produced in the irregularity value calculation step, b) penalty, and c) global normalisation.

Blending of four IMs is performed by the following formulation:

$$\text{IM}_{blended} = \alpha_{IM} \times \text{IM}_1 + \beta_{IM} \times \text{IM}_2 + \gamma_{IM} \times \text{IM}_4 + \delta_{IM} \times \text{IM}_8 \tag{4.2}$$

where $\alpha_{IM} + \beta_{IM} + \gamma_{IM} + \delta_{IM}$ is equal to 1 and IM_1, IM_2, IM_4, and IM_8 are IMs from 1×1, 2×2, 4×4, and 8×8 pixels of source/target patches. Note that combining all

information from patches of different sizes is performed to capture different levels of details, where smaller patches capture a more detailed information of the MRI's intensity while bigger patches capture a bigger contextual information of the brain (Rachmadi et al., 2017c). The blended IM is depicted in Figure 4.1(D).

The blended IM is then *penalised* using:

$$pen = blend \times ori \tag{4.3}$$

where $blend$ is the voxel value from the $IM_{blended}$, ori is the voxel value from the original T2-FLAIR MRI, and pen is the penalised voxel value. Penalisation is performed to eliminate artefacts often caused by low quality ICV/CSF mask (Rachmadi et al., 2017c). Artefacts might be produced in previous step (Equation (4.1)) when non-brain tissues represented as hypo-intensities in T2-FLAIR MRI are unsuccessfully excluded by ICV/CSF mask. Note that Equation (4.1) cannot differentiate between hyper-intensities (bright voxels) and hypo-intensities (dark voxels).

Lastly, all IMs from different MRI slices are normalised together to produce values between 0 to 1 for each voxel to estimate "irregularity" with respect to the normal brain tissue across all slices. This normalisation procedure is called *global normalisation*. The resulted IM, penalised, and globally normalised, is depicted in Figure 4.1(E).

Some important notes on IM computation are: 1) source and target patches need to have the same size within the hierarchical framework, 2) the centre of source/target patches needs to be inside the brain and outside the CSF masks at the same time to be included in the irregularity value calculation, and 3) slices which do not provide any source patch (i.e where no brain tissue is observed) are skipped.

4.3 Limited One-time Sampling Irregularity Map

As previously mentioned, while the original IAM has been reported to work well for WMH segmentation, its computation takes considerable time because it performs one target patch sampling for each source patch, selecting different target patches per source patch. For clarity, this scheme is named Multiple-time Sampling (MTS) scheme. The MTS scheme is performed in the original IAM so that no target patch is too close to the source patch (location condition) (Bellini et al., 2016). Extra time in MTS to sample target patches for each source patch is, therefore, unavoidable under these premises.

To accelerate the computation, a new scheme called One-time Sampling (OTS) was proposed and evaluated, where target patches are randomly sampled only once for each

MRI slice, hence abandoning MTS's location based condition (Rachmadi et al., 2018c). In other words, distance values of all source patches from one slice were computed against one (i.e. the same) set of target patches. This new method is named "One-time Sampling IAM" (OTS-IAM).

In this study, limited number of the OTS's target patches is proposed to accelerate the computation even more. This new method is named Limited OTS-IM (LOTS-IM). Note that the original IAM, which runs on CPUs, uses an undefined large number of target patches which could range from 10% to 75% of all possible target patches, depending on the size of the brain tissue in an MRI slice.

Six numbers of target patches are sampled and evaluated for the computation of LOTS-IM; 2048, 1024, 512, 256, 128, and 64. A more systematic way to calculate the irregularity value is also proposed where the $1/8$ largest distance values are used instead of a fixed number of 100. The ratio of the $1/8$ largest distance values is used as it represents the second half of the third quartile (Q_3) of the samples, i.e., outliers. Thus, the 256, 128, 64, 32, 16, and 8 largest distance values, deemed as outliers, are used to calculate irregularity values for 2048, 1024, 512, 256, 128, and 64 target patches respectively. Smaller number of target patches in the LOTS-IM enables us to implement it on GPU to accelerate the computation. The limited number of samples in power-of-two is carefully chosen to ease GPU memory allocation.

4.4 IM for Simulation of Brain Abnormalities

Brain lesions evolution over a period of time is very important in medical image analysis because it not only helps estimating the pathology's level of severity but also selecting the "best" treatment for each patient (Rekik et al., 2014). However, predicting brain lesions evolution is challenging because it is influenced by various hidden parameters unique to each individual. Hence, brain lesions can appear and disappear at any point in time (Rekik et al., 2014) while the reasons behind it are still not fully well known. Previous studies that have modelled brain lesion progression/regression, use longitudinal (i.e., time-series) data to formulate lesions metamorphosis by estimating direction and speed of the lesions evolution over time (Hong et al., 2012; Rekik et al., 2014). Hence, multiple scans are necessary to simulate the evolution of the lesion.

The use of IM is proposed for simulating brain lesion evolution (i.e., progression and regression) by using one MRI scan at one time point. This is possible thanks to the nature of IM which retains original T2-FLAIR MRI's complex textures while indicating

irregular textures of WMH. Compared to manually produced WMH binary mask by experts or automatically produced PM by machine learning algorithms, information contained/retained in IM is much richer (see Figure 4.2). Note that this is based on the original study of (Bellini et al., 2016) in computer graphics where simulation of weathering effect on textured natural image was proposed.

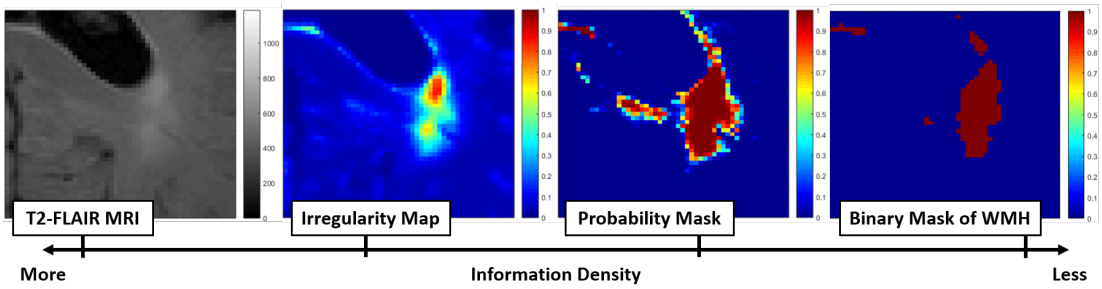


Figure 4.2: Information density retained in each domain of the original T2-FLAIR, IM, PM, and binary mask of WMH.

Algorithm 1: Brain lesions regression (shrinkage) simulation algorithm

input : Original T2-FLAIR MRI

output : Irregularity map and sequential time points of "healthier" T2-FLAIR

$t = 1$;
$\eta = 0.05$;
$Fl(1) =$ T2-FLAIR ;
$I_{Original} =$ LOTS-IM$(Fl(1))$;
$Fl(0) =$ load/make pseudo-healthy of T2-FLAIR (see Algorithm 2);
while $t > 0$ **do**
    $t = t - \eta$;
    $Irr(t) = I_{Original} - (1 - t)$;
    $Fl(t) = Fl(0) + (Fl(1) \times Irr(t))$ (Equation 4.4);
    save $Irr(t)$ and $Fl(t)$;
end

4.4.1 Brain lesions regression (shrinkage) simulation algorithm

The regression pattern of brain lesions is simulated by lowering the irregularity values of the IM gradually. This is possible as each voxel of IM contains different irregularity

Algorithm 2: Pseudo-healthy MRI generation algorithm

input : Original T2-FLAIR MRI

output : Pseudo-healthy T2-FLAIR MRI

$I_{Original} =$ LOTS-IM(T2-FLAIR);

for each patch that has $I_{Original} > 0.178$ **do**

  sample all normal patches (i.e., $I_{Original} \leq 0.178$) close to the original patch (i.e., distance between the original and normal patches is ≤ 12 pixels) from original T2-FLAIR MRI;

  calculate distance values between the original patch and all possible sampled normal patches from T2-FLAIR MRI using Equation (4.1);

  randomly pick a normal patch with small distance value (e.g., from 64 smallest normal patches) and average it with the original patch;

end

value associated with the original T2-FLAIR. It can be observed in Figure 4.2 where irregularity values of brain lesion decrease gradually from the centre to the border of each brain lesion. This is not possible using PM produced by most machine learning algorithms or binary masks of WMH produced manually by expert where most lesion voxels have flat value of 1.

The algorithm for simulating brain lesions regression is described in Algorithm 1 where the irregularity values of IM are gradually decreased by a fixed step (η) in each loop. In this study, irregularity values of IM in time t (i.e., $Irr(t)$) are decreased by $\eta = 0.05$ to get the irregularity values in time $t - \eta$ (i.e., $Irr(t - \eta)$). Note that t in this study is a real number, and $t = 1$ and $t = 0$ are reserved for the original input and pseudo-healthy respectively. After $Irr(t - \eta)$ is generated using Algorithm 1, the corresponding T2-FLAIR with regressed abnormalities (i.e., $Fl(t - \eta)$) can be generated by using Equation (4.4) below:

$$Fl(st) = Fl(0) + (Fl(1) \times Irr(st)) \tag{4.4}$$

where $st = t - \eta$ stands for "simulated time" and represents the regression of the abnormalities from time-t with fixed step η, $Fl(1)$ is the original T2-FLAIR MRI, and $Fl(0)$ is the pseudo-healthy of T2-FLAIR MRI.

For simulating the brain lesions regression in T2-FLAIR, a pseudo-healthy of T2-FLAIR MRI is needed and generated first. Pseudo-healthy is a generated (fake) subject-specific "healthy" image from a pathological one (Xia et al., 2019). In this

study, this can be done by replacing the original "abnormal" brain tissue patches with the nearest neighbour of "normal" brain tissue patches with the help of distance value calculated by Equation (4.1) and Algorithm 2. Let $\mathbf{s}$ be the original (not-normal) patch and $\mathbf{t}$ be the candidate of nearest neighbour (normal) patch, the distance d between the two patches is calculated by using Equation (4.1). The patch's size used in this study to produce pseudo-healthy T2-FLAIR MRI is 3×3.

4.4.2 Brain lesions progression (growth) simulation algorithm

Compared to the previous algorithm for simulating regression, the algorithm for simulating brain lesions progression is more complex as it involves nearest neighbour searching and patch replacement processes. The idea is simple; similar IM patches (i.e., nearest neighbours) with slightly higher irregularity values ($\eta = 0.05$) than the original IM patch are needed for each original IM patch. Once the nearest IM patch is found, the original IM patch is then replaced. Once all patches are replaced by their nearest IM patches, a new T2-FLAIR MRI showing brain lesion progression can be produced by blending the new IM with the pseudo-healthy T2-FLAIR MRI by using Equation (4.4) where $st = t + \eta$ stands for "simulated time" and represents the progression of the abnormalities from time-t with fixed step η. The algorithm for simulating brain lesion progression is detailed in Algorithm 3.

4.5 Experimental Setup

In this section, subjects, MRI data, other MWH segmentation methods, and evaluation measurements used in this study are described.

4.5.1 Subjects and MRI data

In this study, T2-FLAIR MRI from the ADNI (Mueller et al., 2005) database[3] is used. The dataset contains brain MRI data from 20 subjects with MCI and early AD. Note that this is the first dataset described in Chapter 3 (see Section 3.2.1). All T2-FLAIR MRI sequences have the same dimension of $256 \times 256 \times 35$ pixels where each voxel is 3.69 mm^3. Full data acquisition information are described in Table 3.2. GT WMH label were produced by following the description in Section 3.2.2. For more details on this

[3] http://adni.loni.usc.edu/

Algorithm 3: Brain lesions progression (growth) simulation algorithm

input : Original T2-FLAIR MRI

output : Irregularity map and sequential time points of "more severe" T2-FLAIR

$\eta = 0.05$;
$Fl(1) =$ T2-FLAIR ;
$I_{Original} =$ LOTS-IM$(Fl(1))$;
$Fl(0) =$ load/make pseudo-healthy of T2-FLAIR (see Algorithm 2);
$\varepsilon = 0.05$; `/* maximum increase of irregularity value */`
for $t = 1.05 : \eta : 2.00$ **do** `/* progression by` η `at one step */`
    $[patches] = find(I_{Original} \geq 0.16)$; `/* patch's size is` 3×3 `*/`
    for *patch* ***in*** *[patches]* **do**
        $[patches_{temp}] = find(I_{Original} > patch + \eta$ **and** $I_{Original} \leq patch + \eta + \varepsilon))$;
        select 128 random *patches* from $[patches_{temp}]$ **as** $[candidates]$;
        for *candidate* ***in*** *[candidates]* **do**
            rotate *candidate* by 90° four times `/* data augmentation */`
        end
        calculate distance values between *patch* and $[candidates]$ using distance function (Equation (4.1));
        select a nearest neighbour *patch*;
        if *irregularity value* ***in*** *nearest neighbour* $>$ *irregularity value* ***in*** *patch* **then**
            replace *irregularity value*;
        end
    end
    $Irr(t)$ is produced here ;
    $Fl(t) = Fl(0) + (Fl(1) \times Irr(t))$ (Equation 4.4);
    save $Irr(t)$ and $Fl(t)$;
end

dataset, please see data-share page URL [4]. The investigators within ADNI[5] contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this study.

4.5.2 Other WMH segmentation methods

As LOTS-IM is an unsupervised method, LOTS-IM's performance is mainly compared with that from other unsupervised segmentation method: the LST-LGA (Schmidt et al., 2012a) described in Section 3.3. Similar to Chapter 3, LST-LGA's kappa value of $\kappa = 0.05$ was used for WMH segmentation.

The performance of LOTS-IM is also evaluated and compared with that from several supervised machine learning algorithms described and evaluated in Chapter 3, which are SVM, RF, DBM, CEN, patch-based 2D CNN with global spatial information (DeepMedic-GSI-2D), patch-based 2D UNet (Patch2D-UNet), and patch-based 2D UResNet (Patch2D-UResNet). Note that UNet and UResNet are used in this study as they have been applied for WMH segmentation in recent years (Guerrero et al., 2018). This comparison aims to give broader insight of LOTS-IM's performance compared to other machine learning WMH segmentation methods.

Similar to Chapter 3, all supervised segmentation methods used in this study were trained and tested using 5-fold cross validation and evaluated on all 60 WMH labelled ADNI MRI scans (full explanation can be read in Section 3.5.1). Class balancing (i.e., WMH and non-WMH) for UNet and UResNet is performed similarly to DeepMedic-GSI-2D (Kamnitsas et al., 2017). Configurations for SVM/RF, DBM, and DeepMedic-GSI-2D algorithms are described in detail in Sections 3.3, 3.4.1, and 3.4.2 respectively. Whereas, configurations for CEN, Patch2D-UResNet, and Patch2D-UNet can be found in (Brosch et al., 2016), (Guerrero et al., 2018), and (Li et al., 2018) respectively.

4.5.3 Evaluation measurements

DSC (Dice, 1945), which measures similarity between GT and automatic segmentation results, is used in this study as the primary measurement of evaluation. Higher DSC score means better performance, and the DSC score itself can be computed using Equation (3.11). Additional measurements positive PPV (Equation (3.9)), TPR (Equation

[4] http://hdl.handle.net/10283/2214

[5] http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

(3.10)), and Specificity (SPC) (i.e., True Negative Rate (TNR)) (Equation (4.5) where True Negative (TN)) are also calculated.

$$Specificity = TNR = \frac{TN}{FP + TN} \tag{4.5}$$

Non-parametric Spearman's correlation coefficient (Myers et al., 2010) is used to compute monotonic correlation between WMH volume produced by each segmentation method and visual ratings of WMH. In this study, Fazekas (Fazekas et al., 1987) and Longstreth visual rating scales (Longstreth et al., 1996) are used for evaluation of each automatic method. The grades of Fazekas and Longstreth visual rating scales are described in Section 2.2.1.

Furthermore, the paired two-sided Wilcoxon signed rank test is performed to see whether the difference between the performance of two algorithms is significant or not, described by p-value and h-value. The latter shows the result of testing the null hypothesis that there is no significant difference of performance between the two algorithms (i.e., if $h = 1$ then the null hypothesis is rejected, and if $h = 0$ then the null hypothesis is not rejected). In this study, if $p < 0.05$ then the null hypothesis is rejected.

4.6 Results and Discussions

In this section, LOTS-IM is evaluated for WMH segmentation, longitudinal WMH assessment, and comparison with other methods including the original IAM and OTS-IAM. In addition, LOTS-IM's performance for scans with different WMH burden is evaluated. Its speed, blending weights, and random sampling are also evaluated and analysed.

4.6.1 LOTS-IM for WMH segmentation

Table 4.1 shows the performance of all methods evaluated for WMH segmentation. Note that the original IAM is listed as IAM-CPU and different optimum thresholds (i.e. TRSH in Table 4.1) are used to produce the best WMH segmentation for each methods. The best values of DSC, PPV, SPC, and TPR evaluation measurements are underlined.

From Table 4.1, it can be seen that the binary WMH segmentations produced by all IM method configurations (i.e., IAM, OTS-IAM, and LOTS-IM methods) outperformed LST-LGA in mean DSC, PPV, SPC, and TPR measurements. Especially for LOTS-IM-512, the best performer of all LOTS-IM methods, the performance differed up

Table 4.1: Experiment results of WMH segmentation based on DSC, PPV, SPC, and TPR. Best values for each measurements are underlined. Column "$\pm$ (%)" shows relative performance difference (mean of DSC) to the LOTS-IM-512. The paired two-sided Wilcoxon signed rank test (with 5% significance level) is performed between LOTS-IM-512 and other methods to see whether the performance difference is significant or not. "Speed increase" is relative to IAM-CPU. **Abbreviations**: "DL" for deep learning method, "#TP" for number of target patches, "TRSH" for optimum threshold, and "Train/Test" for training/testing time in minute (min).

Method		DL	#TP	TRSH	DSC		Wilcoxon		PPV	SPC	TPR	Train	Test	Speed
					mean (SD)	± (%)	h	p	(mean)	(mean)	(mean)	(min)	(min)	increase
LST-LGA	UNSUPERVISED	✗	-	0.134	0.3037 (0.166)	-16.92	1	0.000	0.3158	0.9946	0.3625	-	0.67	-
IAM (CPU)		✗	75%	0.179	0.3930 (0.121)	-7.99	1	0.000	0.7001	0.9993	0.3757	-	217.18	-
OTS-IAM-CPU		✗	75%	0.164	0.4297 (0.173)	-4.32	1	0.000	0.6994	0.9992	0.3827	-	173.50	1.26
LOTS-IM-2048		✗	2,048	0.178	0.4710 (0.182)	-0.19	0	0.051	0.6111	0.9984	0.4564	-	12.43	17.52
LOTS-IM-1024		✗	1,024	0.178	0.4721 (0.183)	-0.08	0	0.054	0.6082	0.9983	0.4607	-	3.82	56.85
LOTS-IM-512		✗	512	0.178	0.4729 (0.185)	-	-	-	0.5918	0.9980	0.4710	-	1.87	116.14
LOTS-IM-256		✗	256	0.178	0.4711 (0.188)	-0.18	0	0.225	0.5722	0.9977	0.4865	-	0.77	282.05
LOTS-IM-128		✗	128	0.178	0.4660 (0.192)	-0.69	0	0.556	0.5357	0.9970	0.5158	-	0.45	482.62
LOTS-IM-64		✗	64	0.178	0.4539 (0.204)	-1.90	0	0.752	0.4769	0.9952	0.5589	-	0.42	517.10
SVM	SUPERVISED	✗	-	0.925	0.2630 (0.150)	-20.99	1	0.000	0.0474	0.9869	0.1259	26	1.38	-
RF		✗	-	0.995	0.3633 (0.184)	-10.96	1	0.002	0.0482	0.9860	0.1320	37	0.68	-
DBM		✓	-	0.687	0.3235 (0.135)	-14.94	1	0.000	0.0642	0.9955	0.0542	1,341	0.28	-
CEN		✓	-	0.284	0.4308 (0.158)	-4.21	1	0.009	0.5255	0.9975	0.4815	152	0.08	-
Patch2D-UResNet		✓	-	0.200	0.5277 (0.173)	+5.48	1	0.000	0.5899	0.9970	0.5968	215	0.08	-
Patch2D-UNet		✓	-	0.200	0.5030 (0.149)	+3.01	1	0.047	0.6480	0.9985	0.4886	211	0.08	-
DeepMedic-GSI-2D		✓	-	0.801	0.5225 (0.169)	+4.96	1	0.000	0.5950	0.9985	0.5276	392	0.45	-

to 16.92% compared to LST-LGA. Furthermore, IAM/OTS-IAM/LOTS-IM not only outperformed LST-LGA but also conventional supervised machine learning algorithms (i.e., SVM and RF), and some of them outperformed supervised deep learning methods of DBM and CEN in DSC measurement. Based on the paired two-sided Wilcoxon signed rank test, the performance of all LOTS-IM configurations were significantly different to LST-LGA, SVM, RF, and DBM with $p < 0.05$.

It is worth mentioning that the best performer of LOTS-IM method, LOTS-IM-512, did not outperform the supervised deep learning methods of Patch2D-UResNet, Patch2D-UResNet, and DeepMedic-GSI-2D. However, LOTS-IM produced output modality (i.e., IM) that is richer and has more granularity than the output of supervised deep learning methods. Figure 4.3 (top) shows that the IM produced by LOTS-IM retains the texture information of both non-WMH and WMH regions, including penumbra of WMH. On the other hand, the PMs produced by DeepMedic-GSI-2D and UNet/UResNet lack the ability to represent non-WMH regions and the penumbra of WMH. Furthermore, IM also can be used for WMH segmentation by thresholding its values, as shown in

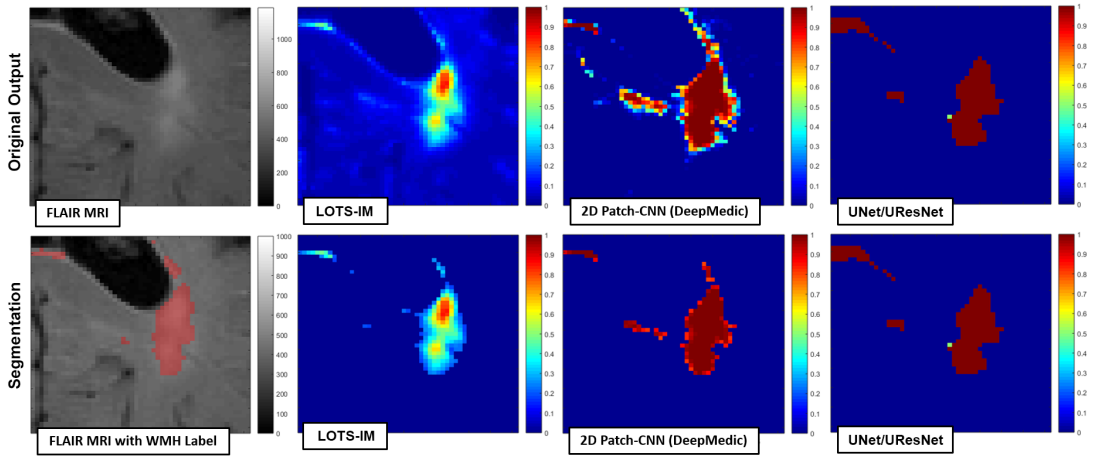


Figure 4.3: **Top:** Visualisation of original outputs produced by LOTS-IM (i.e., IM) and other machine learning methods such as CNN, UNet, and UResNet (i.e., PMs). **Bottom:** Visualisation of WMH segmentation by cutting off the original values of IM and PM. This figure shows that IM not only well represents the penumbra of WMH by retaining the original textures but also is able to segment WMH by cutting off its values.

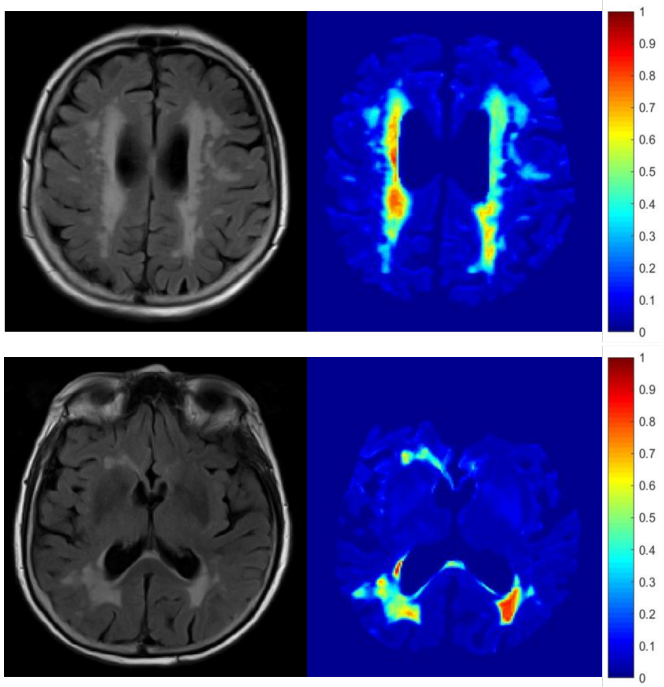


Figure 4.4: Large WMH visualised using IM produced by the proposed LOTS-IM method. Note how both non-WMH and WMH regions, including the penumbra of WMH, are well represented by irregularity values.

Figure 4.3 (bottom). Visualisation of the IM on a scan with large WMH load can be seen in Figure 4.4.

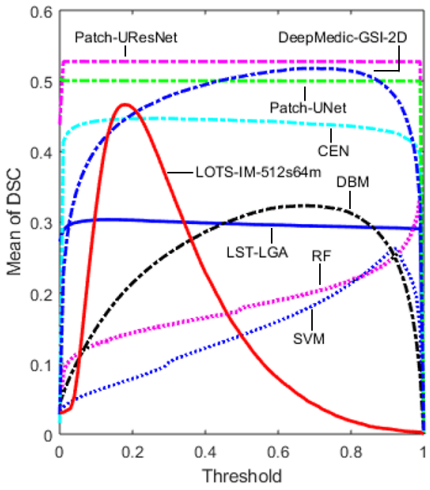


Figure 4.5: Mean of DSC score from all subjects for LST-LGA, SVM, RF, DBM, CEN, Patch2D-UResNet, Patch2D-UNet, DeepMedic-GSI-2D, and LOTS-IM-512 against possible threshold values.

The unique ability of IM to retain the texture information of both non-WMH and WMH regions means that it has different characteristic on performing WMH segmentation. Figure 4.5 shows the DSC performance curves of LOTS-IM and other WMH segmentation methods by cutting off the irregularity or probability values on different threshold values. LOTS-IM uses lower threshold values than the other methods to produce better WMH segmentation as the IM gives finer brain tissues details than the other methods. It is also worth mentioning that the peak of LOTS-IM's performance is located close to the performance of supervised deep learning methods.

4.6.2 LOTS-IM vs. IAM and OTS-IAM

Table 4.1 shows that Limited One-time Sampling (LOTS) scheme not only accelerated the computational time but also improved the overall performance due to the use of limited number of target patches. Implementation of LOTS-IM on GPU increased the processing speed by 17 to 435 times with respect to the original IAM (implemented on CPU). Furthermore, it is worth stressing that this increase in processing speed was not only due to the use of GPU instead of CPU, but also due to the limited number of target patch samples used in LOTS-IM. Furthermore, one of the implementations (i.e., LOTS-IM-64) ran faster than LST-LGA. The increase in speed shows the effectiveness of the proposed method of LOTS-IM GPU in terms of computational time and overall performance. Note that the testing time in Table 4.1 excludes registrations and

generation of brain masks in pre-/post-processing step.

4.6.3 Speed vs. quality of LOTS-IM

The biggest contribution of this work is the increase in processing speed without compromising the quality of the results. The first iteration of IAM can only be run on CPU because it uses MTS. OTS-IAM samples patches only once, but still uses a high number of target patches to compute the IM. Through this study, it can be seen that using a limited number of target patches leads not only to faster computation but also to achieve small to none quality degradation.

The relation between speed and quality of the output (mean DSC) produced by IAM, OTS-IAM, and all configurations of LOTS-IM is illustrated and described in Figure 4.6 and Table 4.1 respectively. Also, it is worth mentioning that the use of more target patches in LOTS-IM produced better PPV and SPC evaluation measurements. The TPR measurement, on the contrary, is better when less target patches are used.

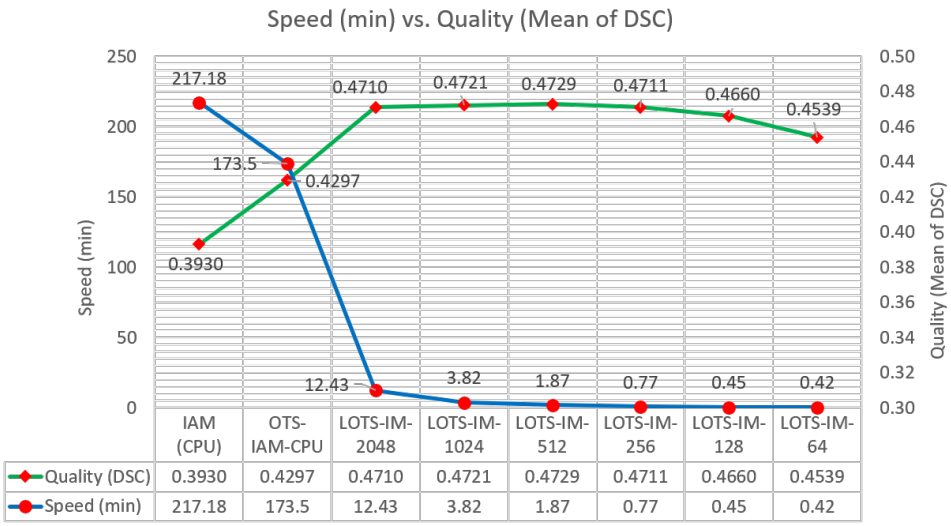


Figure 4.6: Speed (min) versus quality (mean of DSC) of different settings of LOTS-IM (extracted from Table 4.1). By implementing LOTS-IM on GPU and limiting the number of target patch samples, computational time and result's quality are successfully improved and retained.

The paired two-sided Wilcoxon signed rank test shows that there was no significant difference between LOTS-IM methods (i.e. $p \geq 0.05$). Thus, LOTS-IM is more flexible than other methods in terms of speed as its computation speed can be adjusted as needed without compromising the output's quality.

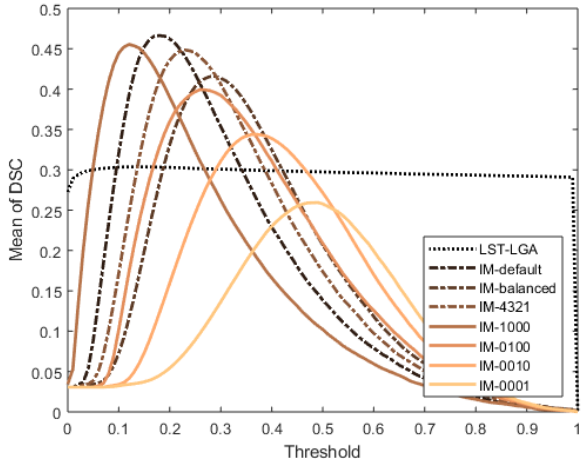


Figure 4.7: Curves of mean DSC produced by using different settings of blending weights. LOTS-IM used in this experiment is LOTS-IM-512, and all weights are listed in Table 4.2.

Table 4.2: Mean and SD of DSC produced by using different settings of blending weights. Plots corresponding to settings listed in this table can be seen in Figure 4.7. The LOTS-IM tested in this experiment is LOTS-IM-512.

Name	Blending Weights				TRSH	DSC	
	α_{IM} (1x1)	β_{IM} 2x2	γ_{IM} 4x4	δ_{IM} 8x8		mean	SD
LST-LGA	-	-	-	-	0.134	0.2936	0.1658
IM-1000	1	0	0	0	0.128	0.4555	0.1774
IM-0100	0	1	0	0	0.267	0.3995	0.1646
IM-0010	0	0	1	0	0.376	0.3439	0.1627
IM-0001	0	0	0	1	0.495	0.2594	0.1289
IM-balanced	0.25	0.25	0.25	0.25	0.287	0.4158	0.1754
IM-4321	0.40	0.30	0.20	0.10	0.228	0.4486	0.1776
IM-default	0.75	0.19	0.05	0.01	0.179	0.4692	0.1820

4.6.4 Analysis on LOTS-IM's blending weights

LOTS-IM has four internal parameters used to blend four IMs, hierarchically produced by four different sizes of source/target patches, to generate the final IM (see Equation (4.2) in Section 4.2.4). In this experiment, different sets of blending weights in LOTS-IM's computation were evaluated. 7 different sets of blending weights were tested and listed in Table 4.2. The effect of different sets of blending weights is illustrated in Figure 4.7.

From Figure 4.7 and Table 4.2, it can be seen that blending irregularity values from different IMs produced better WMH segmentation results. The IM produced by 1×1 pixels of source/target patches influences the WMH segmentation more than the others

Table 4.3: Three groups of MRI data based on WMH volume.

No.	Group	WMH Vol. (mm^3)	# MRI Data
1	Small	WMH $\leq$ 4500	27
2	Medium	4500 $<$ WMH $\leq$ 13000	25
3	Large	WMH $>$ 13000	8

(i.e. those of dimensions 2×2, 4×4, and 8×8 pixels). Furthermore, the skewed blending weights of 0.75, 0.19, 0.05, and 0.01 produced the best DSC score. The skewed blending weights come from the ceiling operation of normalising the power of two (i.e., $2^6/85$, $2^4/85$, $2^2/85$, and $2^0/85$ where $85 = 2^6 + 2^4 + 2^2 + 2^0$). Based on the paired two-sided Wilcoxon signed rank test, the performances of the skewed blending weights to the IM produced by 1×1 pixels of source/target patches were significantly different ($p < 0.05$). As the skewed blending weights of 0.75, 0.19, 0.05, and 0.01 produced the best DSC score in this experiment, it is chosen to become the default blending set for the LOTS-IM. Also, note that this default blending set was used for all other experiments in Section 4.6.

Through this experiment, it can be seen that it is necessary to consider not only the intensity of the individual pixels but also those from the group of pixels (textons) which convey the textural information. Furthermore, combining IMs produced by different sizes of non-overlapping sources is also similar to calculating IM using overlapping source patches. It is also useful to reduce pixellation or discretisation of IM by averaging (i.e., generalising) irregularity values from different sizes of patch instead of using one irregularity value from pixel-wise computation. Nevertheless, individual pixel intensities constitute the strongest feature for irregularity detection.

4.6.5 WMH burden scalability test

In this experiment, all methods were evaluated to see their performances on segmenting WMH in MRI scans with different burden of WMH. The DSC measurement is still used, but the dataset is categorised into three different groups according to each patient's WMH burden (Table 4.3). The results can be seen in Figure 4.8 and Table 4.4. Note that LOTS-IM is represented by LOTS-IM-512, the best performer amongst the LOTS-IM methods in Table 4.1.

From Figure 4.8, it can be appreciated that LOTS-IM-512 performed better than LST-LGA in all groups. LOTS-IM-512 also performed better than the conventional

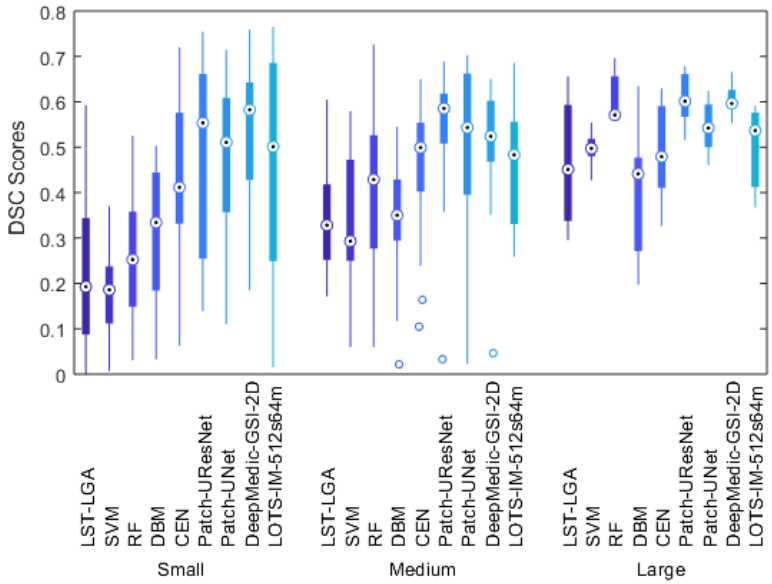


Figure 4.8: Distributions of DSC scores for all methods tested in this study in respect to WMH burden of each patient (see Table 4.3).

supervised machine learning algorithms (i.e. SVM and RF) in "Small" and "Medium" WMH burden groups. Whereas, LOTS-IM-512's performance was at the level, if not better, than the supervised deep learning algorithms DBM and CEN. However, LOTS-IM-512 still could not beat the state-of-the-art supervised deep learning algorithms in any group. It also can be observed that the SD of LOTS-IM-512's performances in "Small" WMH burden is still relatively high compared to one from the other methods evaluated. However, LOTS-IM-512's performance is more stable in "Medium" and "Large" WMH burdens (i.e., lower SD). Furthermore, it is worth mentioning that all tested methods have high SD in "Small" burden of WMH which exhibits the challenge of performing small (i.e., early) WMH even for deep learning algorithms. Note that small WMH has marginal effect to the overall performance of machine learning algorithms on WMH segmentation, so supervised machine learning algorithms usually "sacrifice" the performance of small WMH segmentation in the training process most of the time.

4.6.6 Analysis on LOTS-IM's random sampling

To automatically detect T2-FLAIR's irregular textures (i.e., WMH) without any expert supervision, LOTS-IM works on the assumption that normal brain tissue is predominant compared with the extent of abnormalities. Based on this assumption, random sampling is used in the computation of LOTS-IM to choose the target patches. However, it raises

Table 4.4: Mean and SD values of DSC score's distribution for all methods tested in this study in respect to WMH burden of each patient (see Table 4.3). Note that LOTS-IM-512 is listed as LIM-512 in this table.

Method	TRSH	DSC - Small		DSC - Medium		DSC - Large	
		mean	SD	mean	SD	mean	SD
LST-LGA	0.138	0.2335	0.1785	0.3524	0.1208	0.4645	0.1399
LIM-512	0.179	0.4682	0.2278	0.4660	0.1331	0.4940	0.0932
SVM	0.925	0.1792	0.0958	0.3360	0.1284	0.4966	0.0377
RF	0.995	0.2512	0.1298	0.4150	0.1662	0.6055	0.0559
DBM	0.687	0.3127	0.1432	0.3442	0.1350	0.4014	0.1474
CEN	0.284	0.4359	0.1802	0.4474	0.1485	0.4896	0.1122
Patch2D-UResNet	0.200	0.5007	0.2064	0.5403	0.1432	0.6064	0.0579
Patch2D-UNet	0.200	0.4872	0.1596	0.5079	0.1697	0.5447	0.0574
2D Patch-CNN	0.801	0.5230	0.1722	0.5118	0.1340	0.6053	0.0341

Table 4.5: Distribution measurements (mean and SD) based on DSC for each LOTS-IM's settings. Each LOTS-IM setting is tested on a random MRI data 10 times.

No	Method	TRSH	DSC	
			mean	SD
1	LOTS-IM-2048	0.178	0.5681	0.0041
2	LOTS-IM-1024	0.178	0.5901	0.0018
3	LOTS-IM-512	0.178	0.5922	0.0033
4	LOTS-IM-256	0.178	0.5925	0.0075
5	LOTS-IM-128	0.178	0.5848	0.0092
6	LOTS-IM-64	0.178	0.5852	0.0141

an important question on the stability of LOTS-IM's performance to produce the same level of results for one exact MRI data, especially using different number of target patches.

In the first experiment, a random MRI data was chosen out of the available 60 MRI data and LOTS-IM was performed for 10 times using different number of target patches. Each result was then compared to the GT and listed in Table 4.5. From this experiment, it can be seen that each setting produced low SD values which indicates that the results are closely distributed around the corresponding mean values. However, there is an indication that higher deviations are produced when using fewer number of target patches.

In the second experiment, three random MRI data were chosen from each group

Table 4.6: Distribution measurements (mean and SD) based on DSC for subject with different WMH burden. Each subject is tested 10 times using LOTS-IM-512.

WMH Burden	Subject	DSC	
		mean	SD
"Small"	S1	0.2481	0.0148
	S2	0.1998	0.0038
	S3	0.5516	0.0067
"Medium"	S4	0.6301	0.0058
	S5	0.3044	0.0013
	S6	0.2907	0.0039
"Large"	S7	0.5659	0.0037
	S8	0.3623	0.0045
	S9	0.5671	0.0051

Table 4.7: Mean and SD values produced in longitudinal test (see Figure 4.9). LOTS-IM-GPU-512 is listed as LIM-512 in this table. The best values are written in bold while the second-best values are underlined. In this longitudinal test, LIM-512 outperformed LST-LGA while competed with the supervised deep learning methods.

Method	DSC					
	Grow		Stay		Shrink	
	Mean	SD	Mean	SD	Mean	SD
LST-LGA	0.1301	0.0350	0.2343	0.0199	0.2706	0.0058
LIM-512	<u>0.2260</u>	0.0084	<u>0.4585</u>	0.0104	<u>0.3715</u>	0.0018
Patch2D-UNet	0.2242	0.0125	0.4207	0.0125	0.3675	0.0242
Patch2D-UResNet	**0.2523**	0.0199	**0.4664**	0.0211	**0.3912**	0.0044
2D Patch2DCNN	0.1440	0.0228	0.4066	0.0298	0.3660	0.0129

of WMH burden (i.e., "Small", "Medium", and "Large" listed in Table 4.3). Then, LOTS-IM-512 was performed 10 times on the selected MRI data. Lastly, the results were compared with the GT. The results are listed in Table 4.6. Similar to the first experiment, low SD values were produced for each subject, regardless of the WMH burden.

The results indicate that LOTS-IM produces stable results of WMH segmentation in multiple test instances regardless of WMH burden while employing a simple random sampling scheme. However, of course, more sophisticated sampling method could be applied to make sure patches of normal brain tissue are more likely to be sampled.

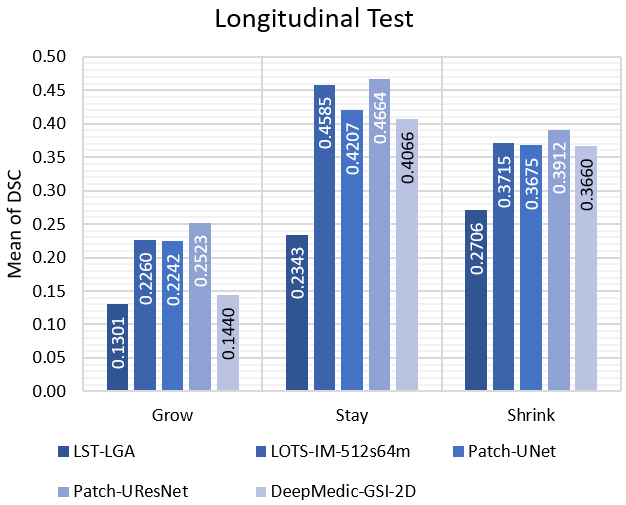


Figure 4.9: Quality of spatial agreement (mean of DSC) of the produced results in longitudinal test. Longitudinal test is done to see the performance of tested methods in longitudinal dataset of MRI (see Section 4.6.7 and Table 4.7 for full report).

4.6.7 Longitudinal test on MCI/AD patients

In this experiment, spatial agreement analysis between the produced results in three consecutive years was evaluated. For each subject, Y2 and Y3 MRI were aligned to the Y1 using niftyReg through TractoR (Clayden et al., 2011), performed subtraction between the aligned WMH labels of baseline/previous year and follow-up year(s) (i.e., Y2-Y1, Y3-Y2, and Y3-Y1), and then labelled each voxel as either “Grow”, “Shrink”, or “Stay”. The voxel is labelled “Grow” or “Shrink” if it has value above zero or below zero after subtraction respectively. Whereas, it is labelled as “Stay” if it has value of zero after subtraction and one before subtraction. This way, it can be seen whether the method captures the progression of WMH across time or not.

Figure 4.9 depicts the results of longitudinal test listed in Table 4.7 for all methods, where LOTS-IM is represented by LOTS-IM-512. In this longitudinal test, LOTS-IM-512 is the second-best performer (underlined) on “Grow”, “Shrink”, and “Stay” regions segmentation task evaluated using DSC measurement after Patch2D-UResNet (written in bold). This, again, confirms that the LOTS-IM shows comparable performance with the state-of-the-art supervised deep learning methods (i.e., Patch2D-UNet, Patch2D-UResNet, and DeepMedic-GSI-2D).

Table 4.8: Non-parametric correlation using Spearman's correlation coefficient between manual/automatic WMH volume and Fazekas and Longstreth visual ratings.

Visual Rating	Fazekas (Total)		Longstreth	
Method	Spearman's Corr.		Spearman's Corr.	
	ρ	p	ρ	p
Manual label	0.7562	1.04×1^{-12}	0.7752	1.45×10^{-12}
LST-LGA	0.5718	3.38×6^{-12}	0.4813	1.50×10^{-4}
LIM-2048	0.4727	2.05×10^{-4}	0.4579	3.42×10^{-4}
LIM-1024	0.4892	1.13×10^{-4}	0.4849	1.32×10^{-4}
LIM-512	0.5010	7.19×10^{-5}	0.5065	5.82×10^{-4}
LIM-256	0.5009	7.22×10^{-5}	0.5085	5.37×10^{-4}
LIM-128	0.4505	4.38×10^{-4}	0.4946	9.22×10^{-4}
LIM-64	0.4393	6.30×10^{-4}	0.4858	1.28×10^{-4}
SVM	0.4062	1.70×10^{-2}	0.3602	5.90×10^{-3}
RF	0.2447	6.66×10^{-2}	0.2128	1.12×10^{-1}
DBM	0.2436	6.79×10^{-2}	0.1659	2.17×10^{-1}
CEN	0.2359	7.74×10^{-2}	0.3618	5.70×10^{-3}
Patch2D-UResNet	0.3602	5.90×10^{-3}	0.5171	3.80×10^{-5}
Patch2D-UNet	0.4618	2.99×10^{-4}	0.5140	4.33×10^{-5}
DeepMedic-GSI-2D	0.7054	9.01×10^{-10}	0.8664	3.19×10^{-18}

4.6.8 Correlation with visual scores

This experiment was performed to see how close LOTS-IM's results correlate with visual rating scores of WMH, specifically Fazekas and Longstreth visual scores. Table 4.8 shows the results of Spearman's correlation coefficient between 1) the total Fazekas score (i.e., the sum of PVWMH and DWMH) and manual/automatic WMH volumes and 2) Longstreth total score and manual/automatic WMH volumes. The grades of Fazekas and Longstreth visual rating scales are described in Section 2.2.1.

Table 4.8 shows that, although not much better, all LOTS-IM methods are highly correlated with visual rating clinical scores. It is worth mentioning that LST-LGA produced WMH segmentation results that are highly correlated with visual ratings but produced the lowest DSC measurement of all (see Table 4.1). On the other hand, LOTS-IM produced high values of DSC measurement and high correlation with visual scores at the same time. Visual inspection of the LOTS-IM results revealed systematic false positive detection in the cerebellum, aqueduct, Sylvian fissure, and some cortical regions. These errors are consistent with those reported by other WMH segmentation

methods (Valdés Hernández et al., 2010).

4.6.9 Simulation of Brain Abnormalities

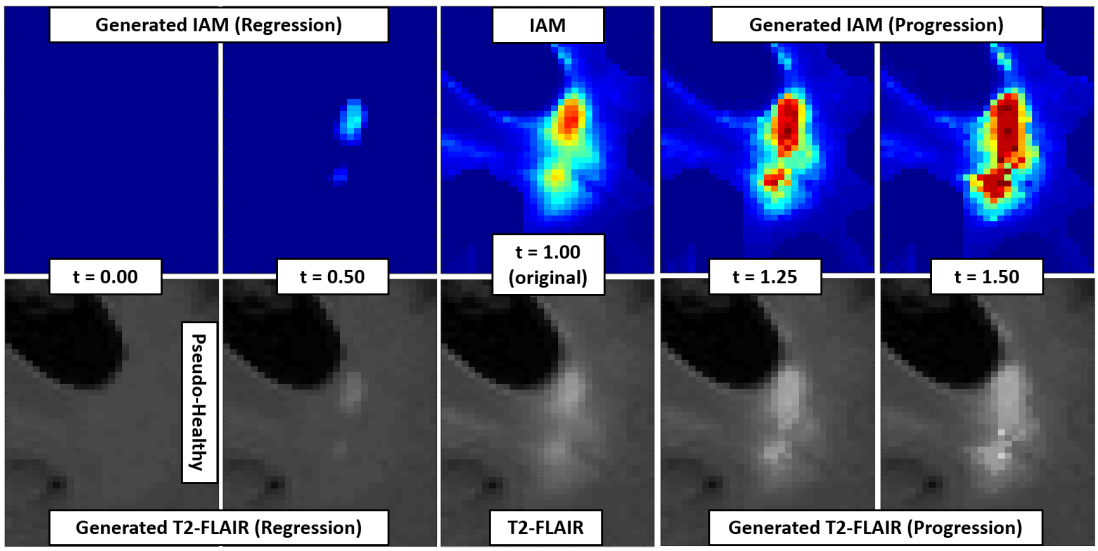


Figure 4.10: Visualisation of brain lesions progression and regression simulation by manipulating irregularity values of IM[2].

Figure 4.10 shows an example of simulated IM and T2-FLAIR from the original IM and T2-FLAIR (centre with $t = 1.00$). The regression step of IM and T2-FLAIR (2^{nd} column with $t = 0.50$) was generated by using Algorithm 1. Whereas, the progression steps of IM and T2-FLAIR (4^{th} and 5^{th} column with $t = 1.25$ and $t = 1.50$) were generated by using Algorithm 3. On the other hand, the pseudo-healthy T2-FLAIR (1^{st} column with $t = 0.00$) was generated using Algorithm 2.

As Figure 4.10 shows, simulation of brain lesions regression works really well for WMH, but simulation of brain lesions progression shows a small unmatched tessellation problem, which is a common problem in computer graphics field. Once this shortcoming is tackled, the simulation results can be used for other purposes such as sources of data augmentation for supervised deep learning methods. However, more investigations are needed to ensure that simulation results follow clinical risk factors of WMH (e.g., blood pressure) and other brain pathologies that usually appear alongside WMH (e.g., stroke and brain atrophy). Note that large WMH is usually followed by deformation of the brain (e.g., large volume of ventricle and brain atrophy). Nevertheless, this experiment shows the suitability of IM for simulating brain lesions progression/regression.

[2]Full simulation can be seen at `https://github.com/febrianrachmadi/iam-tl-progression`.

4.7 Conclusion and Future Work

In this study, the development and use of LOTS-IM for WMH segmentation and simulation of brain abnormalities are described and evaluated. It has been shown that the optimisation of the proposed IM method (LOTS-IM) accelerates processing time by large margin without excessive quality degradation compared with the previous iterations (IAM and OTS-IAM). LOTS-IM speeds up the overall computational time, attributable not only to implementation on GPU but also to the use of a limited number of target patch samples. In addition, different scenarios and settings of LOTS-IM are tested, evaluated, and reported.

Unlike other WMH segmentation methods, LOTS-IM successfully identifies and represents both non-WMH and WMH regions using IM, including the "penumbra" of WMH. Despite not being a WMH segmentation method *per se*, LOTS-IM can be applied for this purpose by thresholding the value of the IM. Being unsupervised confers an additional value to this fully automatic method as it does not depend on expert-labelled data, and therefore is independent from any subjectivity and inconsistency from human experts, which typically influence supervised machine learning algorithms. The results show that LOTS-IM outperforms LST-LGA (i.e., the current state-of-the-art unsupervised method for WMH segmentation), conventional supervised machine learning algorithms (i.e., SVM and RF), and some supervised deep learning algorithms (i.e., DBM and CEN). Furthermore, the results also show that LOTS-IM has comparable performance with the state-of-the-art supervised deep learning algorithms (DeepMedic, UResNet, and UNet).

IM also has shown to be very useful for the simulation of brain lesions progression and regression. There are still some problems in the simulation of progression such as unmatched tessellation, T2-FLAIR contrast changes, and slightly higher computation time compared to simulating regression. The accuracy of simulated image against the original data (i.e., MR image and other clinical data) have to be investigated as well. However, it does not change the fact that the use of IM facilitates the simulation of brain lesions progression and regression.

One limitation of LOTS-IM is the influence that the quality of brain masks (i.e., CSF and NAWM) has in its performance. It has been shown that the tested random sampling has a small impact to the final result on WMH segmentation, but more effective sampling method could be used as well. Some improvements also could be done by adding other brain tissues masks, such as cortical and cerebrum masks.

In the future, the IM could provide unsupervised information for pre-training supervised deep learning, such as UResNet and UNet. In (Rachmadi et al., 2018a), UNet successfully learned the IM produced by the LOTS-IM. The simulation results of brain abnormalities regression and progression could potentially be used for data augmentation of training data in supervised deep learning WMH segmentation methods. Due to its principle, it could be applicable to segment brain lesions in CT scans or different brain pathologies, but further evaluation would be necessary. Further works could also explore its implementation on a multispectral approach that combines different MRI sequences. The implementation of LOTS-IM on GPU is publicly available.[6]

[6] https://github.com/febrianrachmadi/lots-iam-gpu.

Chapter 5

Disease Evolution Predictor Deep Neural Networks

In this chapter, deep learning models for predicting and estimating the evolution of WMH are described. This chapter is based on the following publications:

1. Rachmadi, M. F., del C. Valdés-Hernández, M., Makin, S., Wardlaw, J. M., and Komura, T. (2019a). Predicting the evolution of white matter hyperintensities in brain MRI using generative adversarial networks and irregularity map. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 146–154, Cham. Springer International Publishing.

2. Rachmadi, M. F., Valdés-Hernández, M. D. C., Makin, S., Wardlaw, J. M., & Komura, T. (2019b). Automatic spatial estimation of white matter hyperintensities evolution in brain MRI using disease evolution predictor deep neural networks. *bioRxiv*, 738641. Submitted to *Medical Image Analysis* (in revision).

5.1 Motivation

In Section 2.1.1, it has been described that WMH are commonly associated with the progression of stroke, dementia, and congnitive decline (Wardlaw et al., 2013; Prins and Scheltens, 2015). Furthermore, in Section 2.1.2, it also has been described that WMH have dynamic changes over time where WMH in a patient may simultaneously shrink (regress), stay unchanged (stable), and grow (progress) as indicated by recent previous studies (Ramirez et al., 2016; Chappell et al., 2017; Wardlaw et al., 2017). This chapter describes the development of deep learning methods for predicting WMH changes over

time. For simplicity, the aforementioned WMH changes are referred as "evolution of WMH" in this study.

In this study, an end-to-end training model for automatically predicting and spatially estimating the dynamic evolution of WMH from *baseline* to the *following time point* using deep neural networks is proposed. The proposed model is called Disease Evolution Predictor (DEP) model (discussed in Section 5.3). The DEP model produces a map named Disease Evolution Map (DEM) which characterises each voxel of WMH or brain tissues as progressing, regressing, or stable WMH (discussed in Section 5.2). Deep neural networks are chosen for this study due to their exceptional performance on WMH segmentation (Rachmadi et al., 2017a; Li et al., 2018; Kuijf et al., 2019), reportedly have produced better results than the conventional machine learning algorithms. Specifically, GAN (Goodfellow et al., 2014) and UResNet (Guerrero et al., 2018) are chosen as base architectures for the DEP model. These architectures represent the state-of-the-art unsupervised and supervised deep neural network models, respectively.

This study differs from previous studies on predictive modelling in the fact that predicting the evolution of specific neuroradiological MRI features (i.e., WMH in T2-FLAIR) is the main interest and objective of this study, not the progression of a disease as a whole and/or its effect. For example, previous studies have proposed methods for predicting the progression from MCI to AD (Spasov et al., 2019) and progression of cognitive decline in AD patients (Choi and Jin, 2018). Instead, the proposed DEP model generates three outcomes: 1) prediction of WMH volumetric changes (i.e., either progressing or regressing), 2) estimation of WMH spatial changes, and 3) spatial distribution of white matter evolution at the voxel-level precision. Thus, using the DEP model, clinicians can estimate the size, extent, and location of WMH in time to study their progression/regression in relation to clinical health and disease indicators, for ultimately design more effective therapeutic interventions. Results and evaluations can be seen in Section 5.8.

The main contributions of this study are as follows.

1. A standard training scheme to predict the evolution pattern of WMH between two time points of assessment is proposed. The proposed scheme consists of two parts: 1) generation of the spatial representation of WMH evolution named DEM and 2) generation of the DEM using deep neural networks.

2. Three different modalities to produce the DEM, which are 1) irregularity map, 2) probability map, and 3) binary WMH label, are proposed and evaluated.

3. Three different DEP learning approaches, which are 1) unsupervised learning, 2) indirectly supervised learning, and 3) supervised learning, are proposed. Unsupervised and indirectly supervised learning approaches are based on GAN (i.e., DEP based on Generative Adversarial Network (DEP-GAN)) whereas the supervised learning approach is based on UResNet (i.e., DEP based on U-Residual Network (DEP-UResNet)). DEP-GAN and DEP-UResNet are discussed in Sections 5.3.1 and 5.3.2 respectively.

4. An ablation study of using different kinds of GAN for DEP-GAN model, namely 1) WGAN-GP, 2) Visual Attribution GAN (VA-GAN), 3) DEP-GAN with 1 critic (DEP-GAN-1C), and 4) DEP-GAN with 2 critics (DEP-GAN-2C), is performed and analysed.

5. An ablation study of four different auxiliary inputs to the DEP model: 1) no auxiliary input, 2) baseline WMH load, 3) baseline WMH and SL loads, and 4) Gaussian noise, is performed and analysed. Further explanation can be read in Section 5.4 while the results can be seen in Section 5.8.2.

6. An analysis of plausibility of the WMH volumetric changes predicted by the DEP models and risk factors of WMH evolution using ANCOVA is performed and analysed. The results can be seen in Section 5.8.2.4.

5.2 Disease Evolution Map

In this study, a standard representation of WMH evolution named DEM is proposed. DEM is produced by using a simple subtraction operation between two images from two time points (i.e., baseline assessment and follow-up assessment). In this study, three different modalities for the subtraction operation are proposed: irregularity map, probability map, and binary WMH label.

As previously described in Chapter 4, irregularity map (IM) is a map/image that describes the "irregularity" level of each voxel with respect to the normal brain tissue using real values between 0 and 1. The IM is unique as it retains some of the original MRI textures (e.g., from the T2-FLAIR image intensities), including gradients of WMH. Furthermore, IM is also independent from a human rater or training data, as it is produced using an unsupervised method (i.e., LOTS-IM). DEM resulted from the subtraction of two IMs has values ranging from -1 to 1 (first row of Figure 5.1). Note

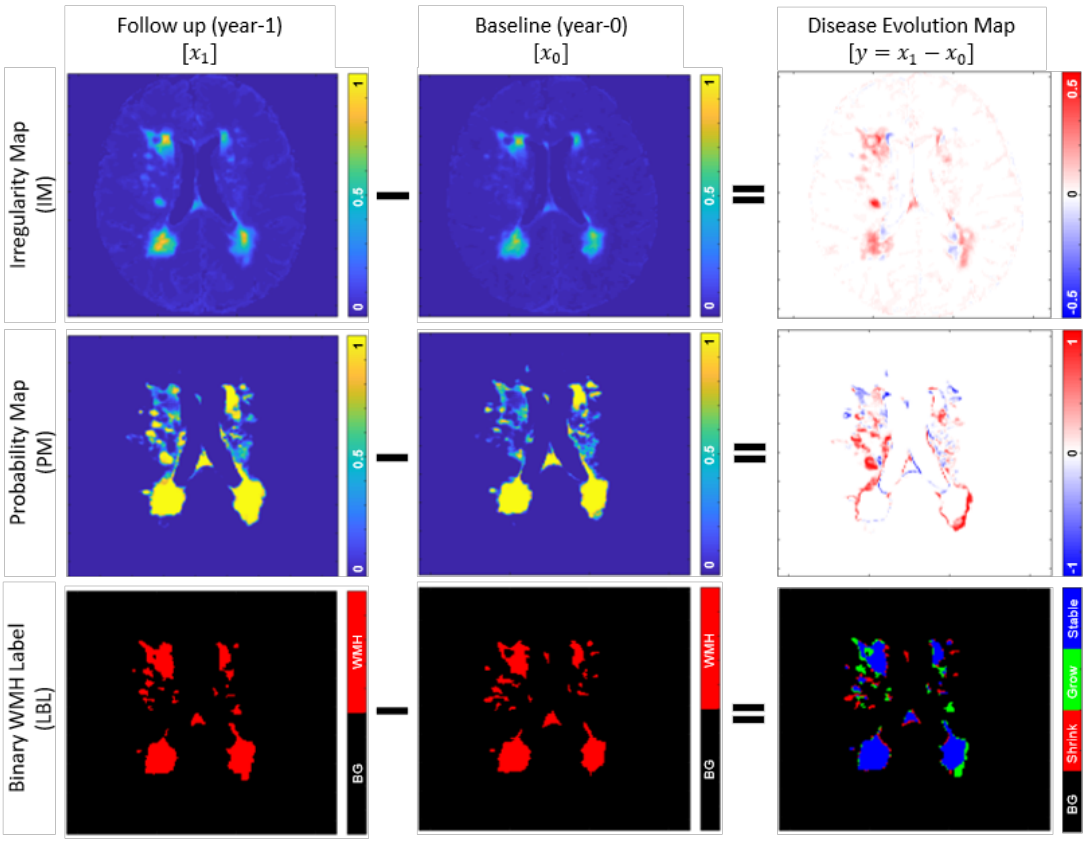


Figure 5.1: DEM (**right**) is produced by subtracting baseline images (**middle**) from follow-up image (**left**). In DEM produced by IM (**first row**) and PM (**second row**), bright yellow pixels represent positive values (i.e., progression) while dark blue pixels represent negative values (i.e., regression). On the other hand, DEM produced by LBL (**third row**) has three foreground labels which represent progression or "Grow" (green), regression or "Shrink" (red), and "Stable" (blue). This special DEM is named LBL-DEM.

how both regression and progression (i.e. dark blue from negative values and bright yellow pixels from positive values in Figure 5.1) are well represented at the voxel level precision on the DEM obtained from IMs.

Probability map (PM) in the present study refers to the WMH segmentation output from a supervised machine learning method. Similar to IM, PM has real values between 0 and 1 which describe the probability for each voxel of being WMH. However, PM differs from IM in the fact that PM only has WMH gradients on the borders of WMH (note that the centre of relatively big WMH clusters have probability of 1). Thus, the DEM produced from the subtraction of two PMs also has values ranging from -1 to 1 representing regression and progression respectively, but these are usually located on the WMH clusters' borders and/or representing small WMH. On the other hand,

the rest of DEM's regions (i.e., the centers of big WMH and non-WMH regions) have probability value of 0 (see the second row of Figure 5.1). Another caveat is that the quality (i.e., accuracy and meaning) of DEM from PM depends on the performance of the automatic WMH segmentation method that generated the PM.

Lastly, binary WMH label (LBL) refers to the WMH label produced by an expert's manual segmentation, which is often considered as gold standard (Valdés Hernández et al., 2015a). DEM from LBL can be produced by subtracting the baseline LBL from the follow-up LBL, and each voxel of the resulted image is then labelled as either "Shrink" if it has value below zero, "Grow" if it has value above zero, or "Stable" if it has value of zero. In this study, this DEM is called three-class DEM label (LBL-DEM), and its depiction can be seen in the bottom-right of Figure 5.1.

5.3 DEP Model using Deep Neural Networks

In this study, two learning approaches of DEP model are proposed and evaluated: 1) non-supervised DEP model based on GANs (DEP-GAN) and 2) supervised DEP model based on UResNet (DEP-UResNet). Each DEP model's workflow consists on two parts: 1) construction of the WMH spatial representation and 2) generation of the predicted DEM. The general flow of DEP model is depicted in Figure 5.2.

DEP-GAN uses either IM or PM to represent the WMH while DEP-UResNet uses T2-FLAIR and LBL-DEM. DEP-GAN using IM is categorised as unsupervised learning because the input modality (IM) is produced by an unsupervised method: LOTS-IM. DEP-GAN using PM is categorised as indirectly supervised learning because the PM is produced by a supervised deep learning algorithm, which is UResNet in this case (see Section 5.6). Finally, DEP-UResNet is categorised as supervised learning as it simply learns DEM labels from LBL-DEM.

5.3.1 DEP Generative Adversarial Network

DEP-GAN is based on GAN, a well established unsupervised deep neural network model commonly used to generate fake natural images (Goodfellow et al., 2014). Thus, in this study, DEP-GAN is proposed to predict the evolution of WMH when there are no longitudinal WMH labels available. DEP-GAN is based on a VA-GAN, originally proposed to detect atrophy in T2-W MRI of AD (Baumgartner et al., 2018). DEP-GAN consists of a generator based on a UResNet (Guerrero et al., 2018) and two separate

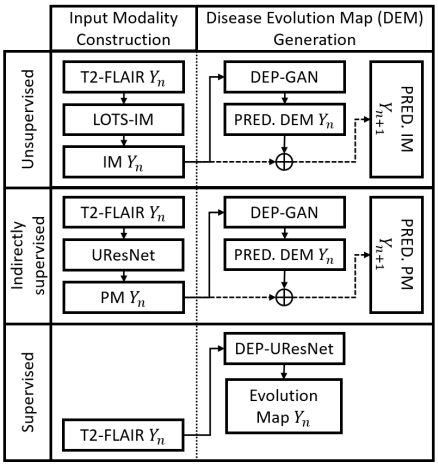


Figure 5.2: Flow diagram of DEP models grouped by learning approach. Each DEP model's workflow is divided into two, which are input modality construction and DEM generation. See Section 5.3 for explanation of DEP models.

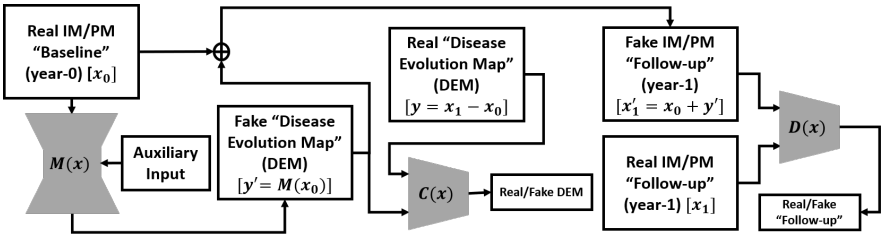


Figure 5.3: Schematic of the proposed DEP-GAN with 2 discriminators (critics). DEP-GAN can take either IM or PM as input. DEP-GAN also has an auxiliary input to deal with the non-deterministic factors in WMH evolution (see Section 5.4 for full explanation).

convolutional networks based on FCN and used as discriminators (hereinafter will be referred as critics). The schematic of DEP-GAN can be seen in Figure 5.3.

Let $\mathbf{x_0}$ be the baseline (year-0) image and $\mathbf{x_1}$ be the follow-up (year-1) image. Then, the "real" DEM ($\mathbf{y}$) can be produced by a simple subtraction ($\mathbf{y} = \mathbf{x_1} - \mathbf{x_0}$). To generate the "fake" DEM ($\mathbf{y}'$), i.e. without $\mathbf{x_1}$, a generator function ($M(x)$) is used: $\mathbf{y}' = M(\mathbf{x_0})$. Thus, a "fake" follow-up image ($\mathbf{x}'_\mathbf{1}$) can be produced by $\mathbf{x}'_\mathbf{1} = \mathbf{x_0} + \mathbf{y}'$. Once $M(x)$ is fully trained, the "fake" follow-up ($\mathbf{x}'_\mathbf{1}$) and the "real" follow-up ($\mathbf{x_1}$) should be indistinguishable by a critic function $D(x)$, while "fake" DEM ($\mathbf{y}'$) and "real" DEM ($\mathbf{y}$) should be also indistinguishable by another critic function $C(x)$. Full schematic of DEP-GAN's architecture (i.e., its generator and critics) can be seen in Figure 5.4.

The DEP-GAN's UResNet-based generator ($M(x)$) has two parts, an encoder which encodes the input image information to a latent representation and a decoder which decodes back image information from the latent representation. The baseline IM/PM

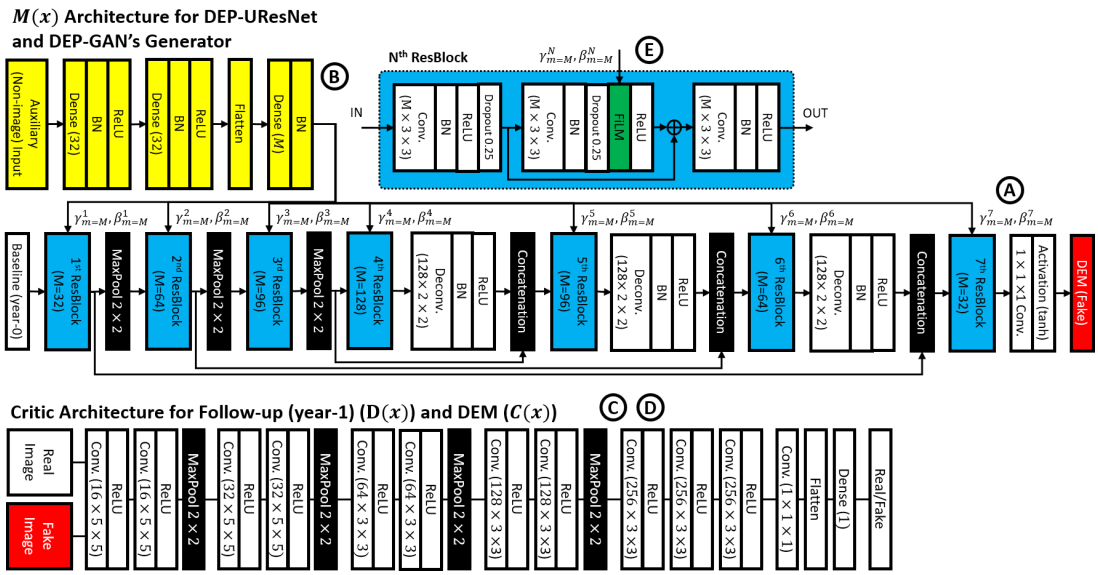


Figure 5.4: Architecture of DEP-GAN, which consists of one generator (**upper side**, "A") and two critics (**lower side**, "C" and "D"). Note how the proposed auxiliary input is feed-forwarded to convolutional layers (yellow, "B") and then modulated to the generator using FiLM layer (green) inside Residual Block (ResBlock) (light blue, "E"). Please see Section 5.4 for full explanation about auxiliary input. On the other hand, DEP-UResNet (**upper right side**, "F") is based on DEP-GAN's generator, including its auxiliary input, with modification of the last non-linear activation function (i.e., from $tanh$ to $softmax$).

($\mathbf{x_0}$) is feed-forwarded to this generator to generate a "fake" DEM ($\mathbf{y}'$). There is also an auxiliary input modulated into the generator using a Feature-wise Linear Modulation (FiLM) layer (Perez et al., 2018) inside the ResBlock to deal with non-deterministic factors of WMH evolution. This auxiliary input and its modulation will be fully discussed in Section 5.4. The architecture of the DEP-GAN's generator is depicted in the upper side of Figure 5.4 (with "A", "B", and "E" annotations for UResNet-based generator of $M(x)$, auxiliary input, and ResBlock respectively).

Unlike VA-GAN that uses only one critic (i.e., only $D(x)$) (Baumgartner et al., 2018), DEP-GAN uses two critics (i.e., $D(x)$ and $C(x)$) to enforce anatomically realistic modifications to the follow-up images (Baumgartner et al., 2018) and encode realistic plausibility in the modifier (i.e., DEM). Anatomically realistic modifications to the follow-up images can be achieved by optimising the critic $D(x)$ and the anatomically realistic plausibility of the modifier can be achieved by optimising the critic $C(x)$. In other words, an anatomically realistic DEM is also essential to produce anatomically realistic (fake) follow-up images. The architecture of the DEP-GAN's critics and their

connection to the generator are depicted in the lower side of Figure 5.4 (with "C" and "D" annotations for critic $C(x)$ and $D(x)$ respectively).

The DEP-GAN's optimisation process is the same as the optimisation of VA-GAN, where the optimisation processes of WGAN-GP using a gradient penalty factor of 10 is used (Gulrajani et al., 2017). The optimisation of $M(x)$ is given by the following function

$$M^* = \arg\min_{M} \max_{D \in \mathcal{D}} \mathcal{L}_{critic}(M, D) + \arg\min_{M} \max_{C \in \mathcal{C}} \mathcal{L}_{critic}(M, C) + \mathcal{L}_{reg}(M) \tag{5.1}$$

where

$$\mathcal{L}_{critic}(M, D) = \mathbb{E}_{\mathbf{x_1} \sim \mathbb{P}_1}[D(\mathbf{x_1})] - \mathbb{E}_{\mathbf{x_0} \sim \mathbb{P}_0}[D(\mathbf{x_0} + M(\mathbf{x_0}))], \tag{5.2}$$

$$\mathcal{L}_{critic}(M, C) = \mathbb{E}_{\mathbf{x_0}, \mathbf{x_1} \sim \mathbb{P}_0, \mathbb{P}_1}[C(\mathbf{x_1} - \mathbf{x_0})] - \mathbb{E}_{\mathbf{x_0} \sim \mathbb{P}_0}[C(M(\mathbf{x_0}))], \tag{5.3}$$

$$\mathcal{L}_{reg}(M) = \lambda_1 \text{MAE}(\mathbf{x'_1}, \mathbf{x_1}) + \lambda_2 (1 - \text{DSC}(\mathbf{x'_1}, \mathbf{x_1})) + \lambda_3 \text{MAE}(vol(\mathbf{x'_1}), vol(\mathbf{x_1})), \tag{5.4}$$

$\mathbf{x_0}$ is the baseline image that has an underlying distribution $\mathbb{P}_0$, $\mathbf{x_1}$ is the follow-up image that has an underlying distribution $\mathbb{P}_1$, $M(\mathbf{x_0})$ represents the "fake" DEM, $\mathbf{x'_1}$ is the "fake" follow-up image, *vol* is a function which computes volumetric measurement by multiplying the number of voxels in the segmentation with the real-world voxel size (i.e., $0.9375 \times 0.9375 \times 4$ mm^3), $\mathcal{D}$ and $\mathcal{C}$ are the critics (i.e. a set of 1-Lipschitz functions (Baumgartner et al., 2018; Gulrajani et al., 2017)), and MAE and MSE are mean absolute error and mean square error (i.e., L1 and L2 losses) respectively. The optimisation is performed by updating the parameters of the generator and critics alternately, where (each) critic is updated 5 times per generator update. Also, in the first 25 iterations and every 100 iterations, the critics are updated 100 times per generator update (Baumgartner et al., 2018; Gulrajani et al., 2017).

In summary, Equation (5.1), which optimises both critics (i.e., $D(x)$ and $C(x)$ using Equations (5.2) and (5.3) respectively) based on WGAN-GP's optimisation process and is regularised using Equation (5.4), needs to be solved to optimise the generator $M(x)$. Each term in the regularisation function (Equation (5.4)) simply says:

1. Intensities of "fake" follow-up images ($\mathbf{x'_1}$) have to be similar to the "real" follow-up images ($\mathbf{x_1}$) based on mean absolute error (MAE) (i.e., L1 loss).

2. The WMH segmentation estimated from $\mathbf{x'_1}$ has to be spatially similar to the WMH segmentation estimated from $\mathbf{x_1}$ based on the DSC (Equation (3.11)).

3. The WMH volume (in ml) estimated from $\mathbf{x_1'}$ has to be similar to the WMH volume estimated from $\mathbf{x_1}$ based on mean square error (MSE) (i.e., L2 loss).

The WMH segmentation of $\mathbf{x_1'}$ and $\mathbf{x_1}$ is estimated by either thresholding IM values (i.e., irregularity values) to be above 0.178 (see Section 4.6.1) or PM values (i.e., probability values) to be above 0.5. Furthermore, each term in Equation (5.4) is weighted by λ_1, λ_2, and λ_3 which equals to 100 (Baumgartner et al., 2018), 1 and 100 respectively. The importance of each regularisation term is discussed in Section 5.8.3.

5.3.2 DEP U-Residual Network

In case LBL for both time points (i.e., baseline and follow-up in longitudinal dataset) are available, a simple supervised deep neural network method can be used to automatically estimate WMH evolution. As previously described in Section 5.2, DEM produced from LBL (i.e., LBL-DEM) consists of 3 foreground labels (i.e., "Grow" (green), "Shrink" (red), and "Stable" (blue)) and 1 background label (black). An example of LBL-DEM can be seen in the bottom-right figure of Figure 5.1.

In this case, the DEP-GAN's generator is detached from the critics and modified into DEP-UResNet by changing the last non-linear activation layer of *tanh* (i.e., for regression) to $softmax$ (i.e., for multi-label segmentation). Thus, the DEP-UResNet's schematic is similar to the DEP-GAN's generator, which can be seen in Figure 5.4 (with "A", "B", and "E" annotations). DEP-UResNet uses T2-FLAIR as input and LBL-DEM as target output. Note that this configuration makes all DEP models have similar generator networks based on UResNet (Guerrero et al., 2018). Furthermore, the auxiliary input proposed in this study can be also applied to the DEP-UResNet. See Section 5.4 for the full explanation about auxiliary input in DEP model.

5.4 Auxiliary Input in DEP Model

The biggest challenge in modelling the evolution of WMH is mainly the amount of factors involved in WMH evolution. In this study, an auxiliary input module which modulates non-image features involved in WMH evolution is proposed. To modulate the auxiliary input to every layer of the DEP-GAN's generator, FiLM layer (Perez et al., 2018) is used. The FiLM layer is depicted as the green block inside the ResBlock in Figure 5.4 (annotated as "E"). In the FiLM layer, γ_m and β_m modulate feature maps F_m,

where subscript m refers to m^{th} feature map, via the following affine transformation

$$FiLM(F_m|\gamma_m,\beta_m) = \gamma_m F_m + \beta_m. \tag{5.5}$$

where γ_m and β_m for each ResBlock in each layer are automatically determined by convolutional layers (depicted as yellow blocks in Figure 5.4 with "B" annotation). Please note that the proposed auxiliary input module can be easily applied to any deep neural network model. Thus, the proposed auxiliary input module is applied to the two DEP models proposed in the present study: DEP-GAN and DEP-UResNet.

In this study, an ablation study of auxiliary input modalities for DEP model was performed. Auxiliary input modalities tested in this study are 1) no auxiliary input (No Auxiliary), 2) baseline WMH volume (+WMH), 3) both baseline WMH and SL volumes (+WMH+Stroke), and 4) Gaussian noise (+Gaussian). Firstly, DEP models without any auxiliary input were tested. Secondly, some risk factors that have been commonly associated with WMH evolution were used. Note that while all factors which influence WMH evolution are not fully well known, baseline WMH load (i.e., cited as the most common and strongest predictor) (Schmidt et al., 2003; Sachdev et al., 2007; Van Dijk et al., 2008; Wardlaw et al., 2017; Chappell et al., 2017) and baseline SL load (Gouw et al., 2008a; Wardlaw et al., 2017) have been found strongly associated with WMH evolution over time. The WMH and SL volumes were obtained from WMH and SL labels/masks. Please see Section 5.5 for full explanation on how WMH and SL masks were produced. Lastly, an array of 32 random noises which follow Gaussian distribution of $z \sim \mathcal{N}(0,1)$ was used as auxiliary input. Hereinafter, this array is referred as Gaussian noise. It is worth to mention that changing the auxiliary input modality from WMH and SL loads to Gaussian noise changes the nature of the DEP model from deterministic to non-deterministic.

5.5 Subjects and Data

An MRI dataset from stroke patients ($n = 152$) enrolled in a study of stroke mechanisms, from which full recruitment and assessments have been published (Wardlaw et al., 2017), was used. Written informed consent was obtained from all patients on protocols approved by the Lothian Ethics of Medical Research Committee (REC 09/81101/54) and NHS Lothian R+D Office (2009/W/NEU/14), on the 29th of October 2009. In the clinical study that provided the data, patients were imaged at three time points (i.e., first time (*baseline*) 1-4 weeks after presenting to the clinic with stroke symptoms,

Table 5.1: Demographics and clinical characteristics of the samples used in this study ($n = 152$).

Vascular risk factors	Diabetes (n, (%))	18 (12)
	Hypertension (n, (%))	114 (75)
	Hypercholesterolaemia (n, (%))	86 (57)
	Recent or present smoker (n, (%))	96 (64)
Relevant SVD imaging markers	Presence of at least 1 microbleed (n, (%))	26 (17)
	Presence of a previous lacune (n, (%))	37 (24)
	SVD score (median [interquartile range (IQR)])	1 [0 2]
	PV WMH Fazekas score (median [IQR])	1 [1 2]
	Deep WMH Fazekas score (median [IQR])	1 [1 2]

at approximately 3 months, and a year after (*follow-up*)). All images were acquired at a GE 1.5T MRI scanner following the same imaging protocol (Valdés Hernández et al., 2015a). Ground truth segmentations were performed using a multi-spectral semi-automatic method (Valdés Hernández et al., 2015a) only from baseline and 1-year follow-up scan visits in the image space of the T1-W scan of the second visit, in $n = 152$ (out of 264) patients. T2-W, T2-FLAIR, gradient echo, and T1-W structural images at baseline and 1-year scan visits were rigidly and linearly aligned using FSL-FLIRT (Jenkinson et al., 2002). The resulted resolution of the images is $256 \times 256 \times 42$ with voxel size of $0.9375 \times 0.9375 \times 4$ mm^3. Note that this voxel size is used to calculate the volume of manual/automated segmentation of WMH. All patients who had the three scan visits and ground truth generated as mentioned above were used in this study. Hence, the total MRI scans are 304 ($n \times 2$) which consist of baseline and 1-year follow-up data. Out of all patients, there are 70 of them that have stroke subtype lacunar (46%) with median SVD score of 1. Other demographics and clinical characteristics of the patients that provided data for this study can be seen in Table 5.1.

The primary study that provided the data used a semi-automatic multi-spectral method to produce several brain masks including ICV, CSF, SL, and WMH, all which were visually checked and manually edited by an expert (Valdés Hernández et al., 2015a). The image processing protocol followed to generate these masks is fully explained in (Valdés Hernández et al., 2015a). Extracranial tissues, SL, and skull were removed from the baseline and follow-up T2-FLAIR images using the SL and ICV binary masks from previous analyses (Chappell et al., 2017; Wardlaw et al., 2017). Furthermore, binary WMH labels produced for the primary study that provided the data (Valdés Hernández et al., 2015a) were used as the gold standard (i.e. ground truth)

for evaluating the DEP models. As per these labels, 98 and 54 out of the 152 subjects have increasing (i.e., progression) and decreasing (i.e., regression) volume of WMH respectively.

As previously explained, IM and PM are needed for DEP-GAN (i.e., the non-supervised learning approach of DEP model). LOTS-IM with 128 target patches was used to generate IM from each MRI data. To generate PM, a 2D UResNet (Guerrero et al., 2018) with gold standard WMH and SL masks was trained and used for WMH and SL segmentation. For this training, all subjects in the dataset were used in a 4-fold cross validation training scheme. Thus, out of 304 MRI data (152 subjects $\times$ 2 scans), 228 MRI data (114 subjects $\times$ 2 scans) were used for training and 76 MRI data (38 subjects $\times$ 2 scans) were used for testing in each fold. Note that this UResNet is different from the DEP-UResNet, which is newly proposed in this study. Notice that "DEP" key-word is affixed to any model's name used for prediction and delineation of WMH evolution. Whereas, the UResNet was previously proposed for WMH and SL segmentation by (Guerrero et al., 2018).

5.6 Experiment Setup

For the present study, 2D architectures were chosen for all networks rather than 3D ones because the number of data available in this study is limited (i.e. only 152 subjects). VA-GAN (i.e., the GAN scheme used as basis for DEP-GAN) used roughly 4,000 subjects for training its 3D network architecture, yet there was still an evidence of over-fitting (Baumgartner et al., 2018). The 2D version of VA-GAN has been previously tested on synthetic data (Baumgartner et al., 2018).

To train DEP models (i.e., DEP-GAN and DEP-UResNet), 4-fold cross validation was performed. In each fold, out of 304 MRI data (152 subjects $\times$ 2 scans), 228 MRI data (114 subjects $\times$ 2 scans) were used for training and 76 MRI data (38 subjects $\times$ 2 scans) were used for testing. Note that DEP models are subject-specific models, so pairwise MRI scans (i.e., baseline and follow-up) are needed and necessary for both training and testing. Out of all slices from the training set in each fold (i.e., 114 pairwise MRI scans), 20% of them were randomly selected for validation. Furthermore, slices without any brain tissues were omitted. Thus, around 4,000 slices were used in the training process in each fold. Values of IM and PM did not need to be normalised as these are between 0 and 1. Finally, each DEP model was trained for 200 epochs (i.e., 200 generator updates for DEP-GAN).

In this study, an ablation study using different GAN architectures for DEP model was performed first. GAN architectures testsed in this study are WGAN-GP, VA-GAN, DEP-GAN-1C, and DEP-GAN-2C. This ablation study is intended to see the impact of the number of critics, the location of the critic(s) and the additional losses proposed in this study. WGAN-GP only generates DEM and has one critic for DEM (i.e., $C(x)$). On the other hand, VA-GAN and DEP-GAN-1C generate both: DEM and the follow-up image, but only have one critic for generating the follow-up image (i.e., $D(x)$). The difference between VA-GAN and DEP-GAN-1C is that DEP-GAN-1C has additional losses for optimisation in the training (see Section 5.3.1). Lastly, DEP-GAN-2C, which generates both: DEM and follow-up image, has two critics for both of them (i.e., $C(x)$ and $D(x)$), and has additional losses for the training.

Furthermore, an ablation study using different types of auxiliary input was also performed and analysed to see the effect of auxiliary input to the DEP models (i.e., DEP-UResNet, DEP-GAN using IM, and DEP-GAN using PM). Note that DEP-GAN used in this ablation study is the DEP-GAN-2C used in the previous ablation study. The procedure of using auxiliary input depends on the input modality and training/testing process. If SL and WMH volumes were used as auxiliary input, these (i.e., not the volumes per slice, but the volume per subject) were feed-forwarded together with one MRI slice. Thus, all slices from one subject used the same number of WMH and SL volumes. Note that WMH and SL loads for the whole dataset (i.e., all subjects) were first normalised to zero mean unit variance before their use in training/testing.

If Gaussian noise were used as auxiliary input, an array of Gaussian noise was feed-forwarded together with an MRI slice in the training process as follows: 10 different sets of Gaussian noise were first generated and only the "best" set (i.e., the set that yielded the lowest M^* loss (Equation (5.1))) was used to update the DEP model's parameters. Note that this approach is similar to and inspired by Min-of-N loss in 3D object reconstruction (Fan et al., 2017) and variety loss in Social GAN (Gupta et al., 2018). In the testing process, 10 different sets of Gaussian noise were generated and the average performance was calculated. Furthermore, in the evaluation, the "best" prediction of WMH evolution based on DSC was also reported.

5.7 Evaluation Measurements

In this study, the following tests and evaluation measurements were performed to assess the performance of DEP models:

1. Prediction error of WMH volumetric change (i.e., whether WMH volume in a subject will increase or decrease).

2. Volumetric agreement between ground truth and predicted WMH volumes of the follow-up assessment using Bland-Altman plot (Bland and Altman, 1986).

3. Volumetric correlation between ground truth and predicted WMH volumes of the follow-up assessment.

4. Spatial agreement of the automatic map of WMH evolution in a patient (i.e. after binarisation) using DSC (Dice, 1945).

5. Clinical plausibility test between the outcome of DEP models in relation with baseline WMH load and clinical risk factors of WMH evolution suggested in clinical studies.

Prediction error is a simple measurement to assess how good a DEP model can predict the WMH evolution in the future follow-up assessment (i.e., increasing or decreasing). On the other hand, volumetric agreement using Bland-Altman plot presents the mean volumetric difference and upper/lower limit of agreement (LoA) (i.e., mean $\pm$ 1.96 $\times$ SD) between ground truth and predicted WMH volumes of the follow-up assessment. Volumetric correlation between ground truth and follow-up predicted WMH volumes was also calculated, complementary to the Bland-Altman plot. Whereas, for evaluating the spatial agreement between ground truth and automatic delineation results, DSC was used. Higher DSC means better performance. The DSC itself can be computed by using Equation (3.11).

In addition, clinical plausibility test, which evaluate the outcome of DEP models in relation with the baseline WMH load and clinical risk factors of WMH change and evolution suggested in clinical studies, was also performed . For this, ANCOVA were performed as follows:

1. The WMH volume at follow-up, predicted from each of the schemes evaluated was used as outcome variable.

2. The baseline WMH volume was the dependent variable or predictor.

3. After running Belsley collinearity diagnostic tests, the covariates in the models were: 1) type of stroke (i.e. lacunar or cortical), 2) basal ganglia perivascular spaces (BG PVS) score, 3) presence/absence of diabetes, 4) presence/absence of

hypertension, 5) recent or current smoker status (yes/no), 6) volume of the index SL (abbreviated as "index SL"), and 7) volume of old SL (abbreviated as "Old SL").

The outcome from an ANCOVA model using the baseline and follow-up WMH volumes of the gold-standard expert-delineated binary masks was used as reference to compare the outcome of the ANCOVA models that used the volumes generated by thresholding the input and output of the DEP models. All volumetric measurements involved in the ANCOVA models were previously adjusted by patient's head size. Therefore, all ANCOVA models used the percentage of these volumetric measurements in ICV rather than the raw volumes.

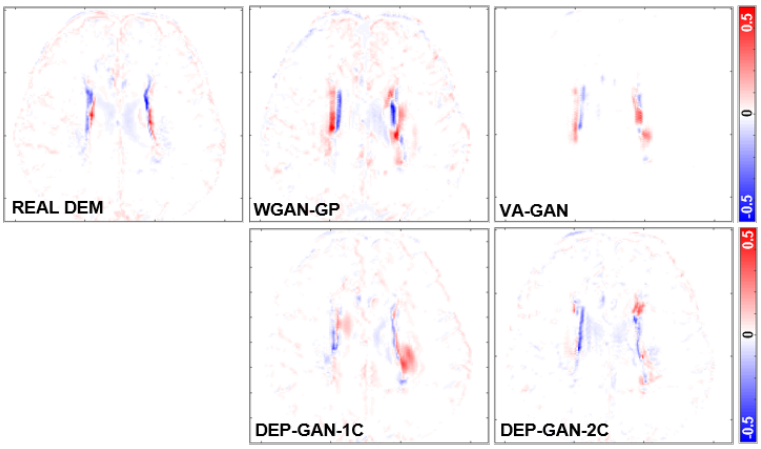
5.8 Results and Discussion

5.8.1 Ablation study of different GAN architectures for DEP model

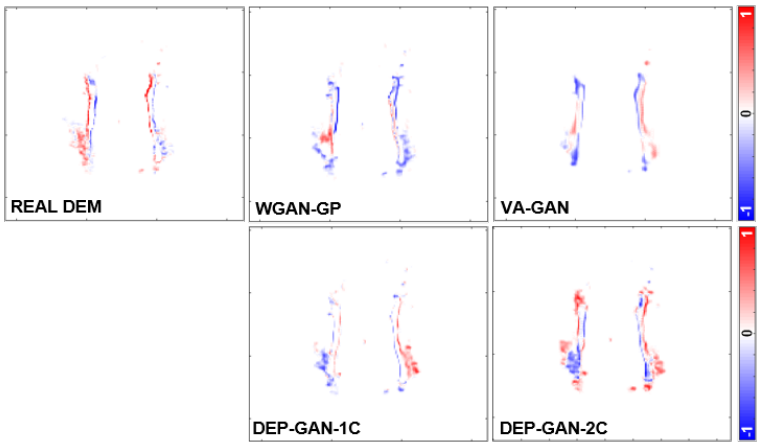
In this ablation study, different GAN architectures were used and evaluated for DEP model to see the impact of number of critics, location of critic(s), and additional losses. As previously described in Section 5.6, WGAN-GP has one critic for DEM (i.e., $C(x)$), VA-GAN has one critic for the follow-up image (i.e., $D(x)$), DEP-GAN-1C has one

Table 5.2: Results from ablation study of different GAN architectures for DEP models. Prediction error of WMH change, volumetric agreement of WMH volume, and spatial agreement of WMH evolution were calculated and compared to the gold standard expert-delineated WMH masks (i.e., LBL-DEM). "Vol." stands for volumetric and "G" and "S" stand for percentage of subjects correctly predicted as having growing and shrinking WMH by DEP models. The best value for each learning approaches and evaluation measurements is written in bold.

Unsupervised (IM)	**Grow (G)** [%]	**Shrink (S)** [%]	**Avg.** [%] ((G+S)/2)	**Vol. Bias** [ml] *mean(SD)*	**Lower LoA** [ml]	**Upper LoA** [ml]	**Entire WMH**	**Change (C)**	**Stable (St)**	**Shrink (Sr)**	**Grow (Gr)**	**Avg.** ((St+Sr+Gr)/3)
WGAN-GP	**85.71**	40.74	63.23	-11.70(24.12)	-59.11	35.70	0.3179	0.0809	0.3294	**0.0595**	0.0325	0.1405
VA-GAN	65.31	62.96	64.13	**2.52(16.43)**	-29.69	**34.72**	**0.3361**	0.0789	0.3506	0.0356	0.0361	0.1408
DEP-GAN-1C	65.31	68.52	**66.91**	3.88(15.93)	-27.33	35.10	0.3343	0.0583	**0.3711**	0.0388	0.0265	0.1454
DEP-GAN-2C	61.22	**72.22**	66.72	5.54(15.98)	**-25.79**	36.87	0.3204	**0.0946**	0.3684	0.0238	**0.0445**	**0.1456**
Indirectly Supervised (PM)	**Grow (G)** [%]	**Shrink (S)** [%]	**Avg.** [%] ((G+S)/2)	**Vol. Bias** [ml] *mean(SD)*	**Lower LoA** [ml]	**Upper LoA** [ml]	**Entire WMH**	**Change (C)**	**Stable (St)**	**Shrink (Sr)**	**Grow (Gr)**	**Avg.**((St+Sr+Gr)/3)
WGAN-GP	55.10	79.63	67.37	4.19(8.28)	-12.05	20.42	**0.6139**	0.2082	0.5906	0.1494	0.0899	0.2766
VA-GAN	42.86	**94.44**	68.65	5.78(8.13)	**-10.15**	21.70	0.6070	0.1946	0.5952	**0.1584**	0.0641	0.2726
DEP-GAN-1C	59.18	85.19	72.18	3.66(7.64)	-11.32	**18.63**	0.6116	0.1711	**0.6012**	0.1186	0.0800	0.2666
DEP-GAN-2C	**69.30**	75.93	**72.66**	**2.48(8.47)**	-14.13	19.08	0.6083	**0.2246**	0.5812	0.1515	**0.1105**	**0.2811**



(a) DEMs using IM.



(b) DEMs using PM.

Figure 5.5: Examples of real DEM and generated DEMs produced by different GAN architectures for DEP model. From left to right: real DEM and generated DEMs produced by WGAN-GP, VA-GAN, DEP-GAN-1C, and DEP-GAN-2C respectively.

critic for the follow-up image (i.e., $D(x)$) and additional losses for optimisation in the training (see Section 5.3.1), and DEP-GAN-2C has two critics for both of DEM and follow-up image (i.e., $C(x)$ and $D(x)$) and additional losses. Furthermore, all methods evaluated used IM and PM as main input modality and did not use any auxiliary input.

5.8.1.1 Spatial agreement (DSC) and qualitative (visual) analyses

Based on Table 5.2 (columns 8-13), it can be seen that DEP-GAN-2C produced better spatial agreement (i.e., higher DSC score) than WGAN-GP, VA-GAN, and DEP-GAN-1C, especially for changing and growing WMH. Qualitative (visual) assessment of

Table 5.3: Volumetric correlation analysis in ablation study of GAN architectures for DEP model. The best value for each correlation measurement is written in bold.

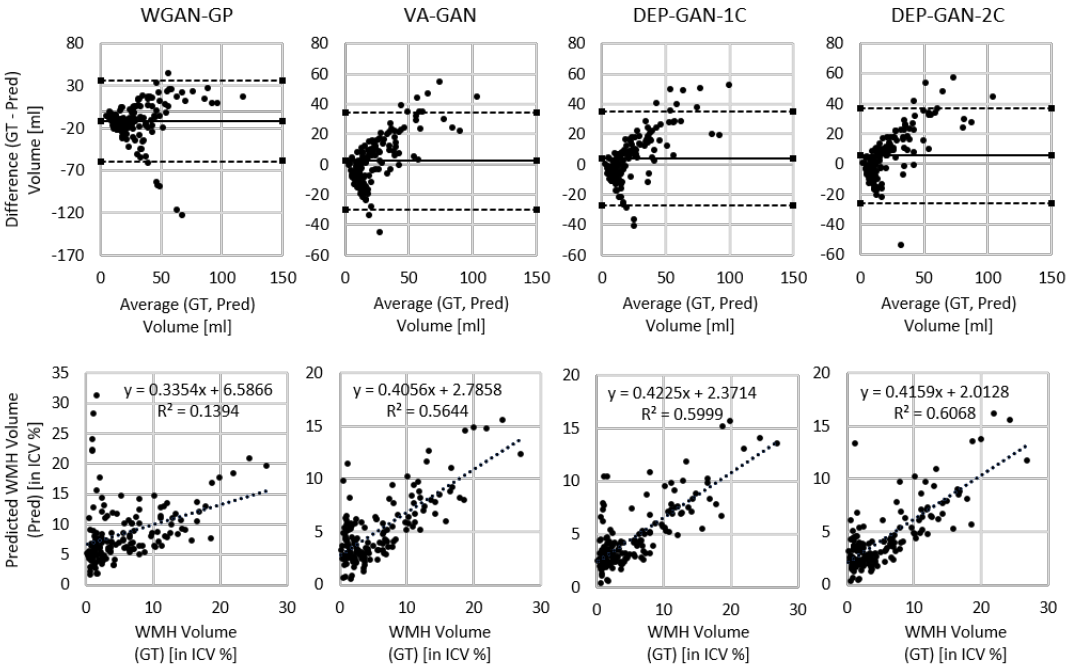
Unsupervised (IM)	WGAN-GP	VA-GAN	DEP-GAN-1C	DEP-GAN-2C
$\mathbf{R^2}$	0.1394	0.5644	0.5999	**0.6068**
Trend	$y = 0.3354x + 6.5866$	$y = 0.4056x + 2.7858$	$\mathbf{y = 0.4225x + 2.3714}$	$y = 0.4159x + 2.0128$
Indirectly Supervised (PM)	WGAN-GP	VA-GAN	DEP-GAN-1C	DEP-GAN-2C
$\mathbf{R^2}$	0.8735	0.8813	**0.8916**	0.8659
Trend	$y = 0.8525x - 0.1265$	$y = 0.8289x - 0.3792$	$y = 0.8799x - 0.1667$	$\mathbf{y = 0.898x + 0.0258}$

generated DEM depicted in Figure 5.5 also shows that DEP-GAN-2C produced more detailed DEM than the other methods, especially when compared to VA-GAN. These results show that DEP-GAN-1C and DEP-GAN-2C are more responsive to the changes of WMH and better in predicting the changes of WMH than VA-GAN. Furthermore, it can also be seen from both Table 5.2 and Figure 5.5 that the use of PM produced better spatial agreement than IM, regardless of the GAN architecture.

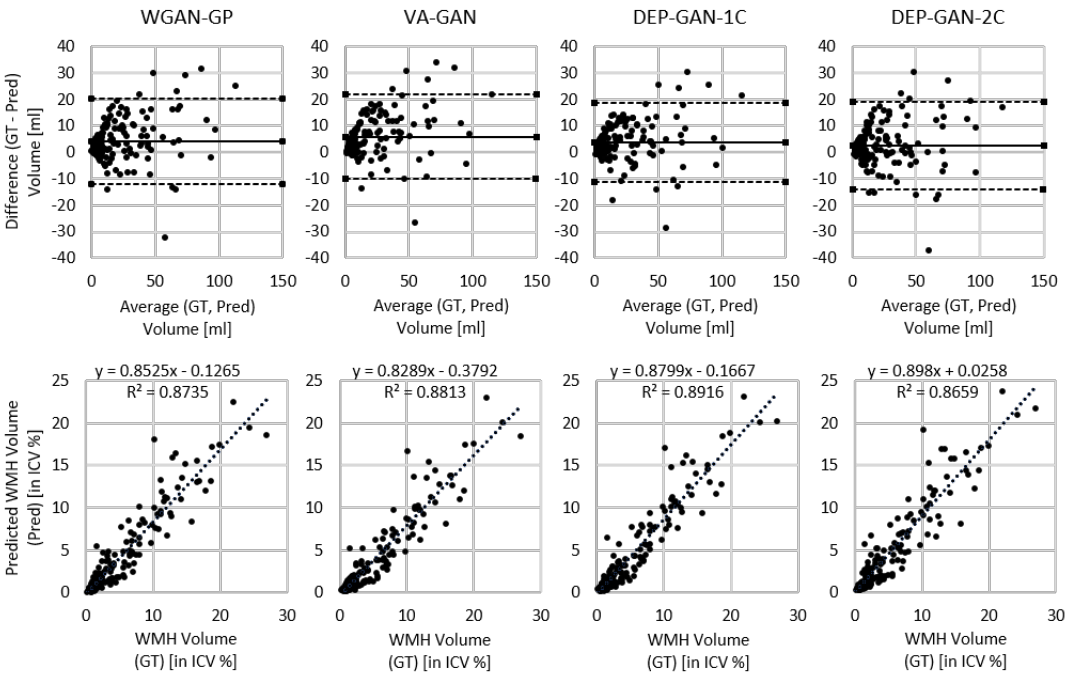
5.8.1.2 Volumetric agreement (Bland-Altman) and correlation analyses

From Table 5.3, it can be seen that the volume of WMH predicted by DEP-GAN-1C and DEP-GAN-2C correlated better with the volume of the ground truth than the volume of WMH predicted uusing WGAN-GP and VA-GAN. However, as per the volumetric agreement analysis (Bland-Altman plot), the performance of DEP-GAN-1C and DEP-GAN-2C depended on the working domain, IM or PM (see columns 5-7 of Table 5.2). If PM was used, DEP-GAN-1C and DEP-GAN-2C performed better than the other methods. On the other hand, VA-GAN achieved the best volumetric agreement when IM was used. However, it is worth to mention that VA-GAN's good performance in the volumetric agreement analysis did not translate to good spatial agreement as previously described in Section 5.8.1.1.

Based on the Bland-Altman and correlation plots depicted in Figure 5.6, it can be seen that PM is better than IM for representing the volumetric change of WMH. From the correlation plots, it can be seen that the correlation between ground truth and predicted WMH volumes when PM was used is higher than when IM was used, regardless of the GAN architecture. Furthermore, Bland-Altman plots show evidence of increasing discrepancy and variability between ground truth and predicted volumes with increasing volume of WMH when IM was used. These discrepancy and variability are less prominent when PM was used.



(a) GAN architectures for DEP model using IM.



(b) GAN architectures for DEP model using PM.

Figure 5.6: Volumetric agreement (in ml) and correlation (in ICV %) analyses between BG PVS and predicted volume of WMH (Pred) produced by WGAN-GP, VA-GAN, DEP-GAN-1C, and DEP-GAN-2C using (a) IM and (b) PM using Bland-Altman and correlation plots.

5.8.1.3 Prediction error analysis and discussion

From Table 5.2 (columns 2-4), it can be seen that most the GAN architectures for DEP models could correctly predict the progression/regression of WMH volume, as they performed better than a random guess system ($\geq 50\%$). Furthermore, based on this ablation study, it can be concluded that DEP-GAN-2C performed generally better for predicting the evolution of WMH due to additional losses and two critics in the architecture. Note that DEP-GAN is used to refer the DEP-GAN-2C in other experiments. Furthermore, there is evidence that PM is better for representing the evolution of WMH than IM when GAN architectures are used for DEP model.

Table 5.4: Results from ablation study of auxiliary input in DEP models. Prediction error of WMH change, volumetric agreement of WMH volume, and spatial agreement of WMH evolution were calculated to the gold standard expert-delineated WMH masks (i.e., LBL-DEM). "Vol." stands for volumetric and "G" and "S" stand for percentage of subjects correctly predicted as having growing and shrinking WMH by DEP models. The best value for each machine learning approaches and evaluation measurements is written in bold. Furthermore, the best value of all learning approaches for each evaluation measurements is underlined and written in bold.

Supervised (DEP-UResNet)	**Grow (G)**[%]	**Shrink (S)** [%]	**Avg.** [%] ((G+S)/2)	**Vol. Bias** [ml] *mean(SD)*	**Lower LoA** [ml]	**Upper LoA** [ml]	**Entire WMH**	**Change (C)**	**Stable (St)**	**Shrink (Sr)**	**Grow (Gr)**	**Avg.** ((Sr+ Gr+St)/3)
No Auxiliary	70.41	72.22	71.32	1.16(7.31)	**-13.17**	15.48	0.6091	0.2234	**0.6332**	0.1551	0.1128	0.3004
+WMH	73.47	**77.78**	75.62	1.59(7.85)	-13.80	16.97	0.6005	0.2532	0.6188	0.1688	0.1409	0.3095
+WMH+Stroke	79.59	75.93	**77.76**	0.81(8.14)	-15.14	16.76	0.6080	0.2565	0.6311	0.1688	0.1415	0.3138
+Gaussian (mean)	**81.63**	59.26	70.45	**-0.58(7.99)**	-16.24	15.09	**0.6135**	**0.2629**	0.6230	**0.1717**	**0.1477**	**0.3141**
+Gaussian (best)	81.63	57.41	69.52	-0.79(7.96)	-16.40	14.81	0.6162	0.2686	0.6280	0.1787	0.1409	0.3159
Unsupervised (DEP-GAN & IM)	**Grow (G)** [%]	**Shrink (S)** [%]	**Avg.** [%] ((G+S)/2)	**Vol. Bias** [ml] *mean(SD)*	**Lower LoA** [ml]	**Upper LoA** [ml]	**Entire WMH**	**Change (C)**	**Stable (St)**	**Shrink (Sr)**	**Grow (Gr)**	**Avg.** ((Sr+ Gr+St)/3)
No Auxiliary	61.22	**72.22**	66.72	5.58(15.98)	-25.79	36.87	0.3204	0.0946	0.3684	0.0238	0.0445	0.1456
+WMH	**75.51**	53.70	64.61	-1.18(19.71)	-39.80	37.45	0.3249	0.0901	0.3551	0.0580	0.0458	0.1530
+WMH+Stroke	71.43	64.81	**68.12**	**0.92(19.91)**	-38.11	39.95	0.3291	0.0922	0.3476	**0.0590**	**0.0468**	0.1511
+Gaussian (mean)	61.22	70.37	65.80	4.59(14.99)	**-24.79**	**33.98**	**0.3359**	**0.2252**	**0.3768**	0.0485	0.0361	**0.1538**
+Gaussian (best)	72.45	64.81	68.83	0.44(15.37)	-29.67	30.56	0.3429	0.1053	0.3795	0.0619	0.0633	0.1682
Indirectly Spv. (DEP-GAN & PM)	**Grow (G)** [%]	**Shrink (S)** [%]	**Avg.** [%] ((G+S)/2)	**Vol. Bias** [ml] *mean(SD)*	**Lower LoA** [ml]	**Upper LoA** [ml]	**Entire WMH**	**Change (C)**	**Stable (St)**	**Shrink (Sr)**	**Grow (Gr)**	**Avg.** ((Sr+ Gr+St)/3)
No Auxiliary	**69.39**	75.93	72.66	2.48(8.47)	-14.13	19.08	0.6083	0.2246	0.5812	0.1515	0.1105	0.2811
+WMH	68.37	70.37	69.37	**1.70(8.24)**	-14.45	**17.84**	**0.6125**	**0.2295**	0.6006	0.1467	**0.1267**	**0.2913**
+WMH+Stroke	66.33	75.93	71.13	2.69(9.14)	-15.22	20.60	0.6098	0.2229	0.5943	**0.1581**	0.1091	0.2872
+Gaussian (mean)	58.16	**79.63**	**68.90**	2.91(8.81)	**-14.36**	20.18	0.6107	0.1801	**0.6245**	0.1216	0.0868	0.2776
+Gaussian (best)	65.31	88.89	77.10	3.63(7.85)	-11.75	19.02	0.6155	0.2415	0.6044	0.1834	0.1265	0.3048

5.8.2 Ablation study of auxiliary input in DEP models

In this ablation study, different types (modalities) of auxiliary input were used and evaluated to see how they affect the performance of DEP models for predicting the evolution of WMH. 4 modalities of auxiliary input were tested, namely 1) no auxiliary input (No Auxiliary), 2) baseline WMH volume (+WMH), 3) both baseline WMH and SL volumes (+WMH+Stroke), and 4) Gaussian noise (+Gaussian). Specific to the Gaussian noise, both of the mean and "best" performances are evaluated and reported. All quantitative results can be seen in Tables 5.4 and 5.5. Whereas, qualitative results (image examples) can be seen in Figures 5.10, 5.11, and 5.12.

5.8.2.1 Volumetric agreement (Bland-Altman) and correlation analyses

From Table 5.4 (columns 5-7), it can be seen that DEP-UResNet using Gaussian noise (+Gaussian (mean)) produced the best estimation of WMH volumetric changes with -0.58 ± 7.99 ml mean difference with respect to the gold standard in volumetric agreement analysis. Furthermore, it can also be seen that almost all DEP-UResNet

Table 5.5: Volumetric correlation analysis of DEP models with different types/modalities of auxiliary input in ablation study of auxiliary input.

Supervised (DEP-UResNet)	$\mathbf{R^2}$	**Trend**
No Auxiliary	**0.9031**	$y = 0.9781x - 0.1397$
+WMH	0.8893	$\mathbf{y = 1.0113x - 0.2435}$
+WMH+Stroke	0.8939	$y = 0.984x - 0.2768$
+Gaussian (mean)	0.8855	$y = 0.9772x + 0.2841$
+Gaussian (best)	0.8869	$y = 0.9821x + 0.3073$
Unsupervised (DEP-GAN & IM)	$\mathbf{R^2}$	**Trend**
No Auxiliary	0.6068	$y = 0.4159x + 2.0128$
+WMH	0.3293	$y = 0.3539x + 3.9732$
+WMH+Stroke	0.3129	$y = 0.3817x + 3.275$
+Gaussian (mean)	**0.6461**	$\mathbf{y = 0.4684x + 1.9418}$
+Gaussian (best)	0.6037	$y = 0.4724x + 2.9103$
Indirectly Spv. (DEP-GAN & PM)	$\mathbf{R^2}$	**Trend**
No Auxiliary	0.8659	$y = 0.898x + 0.0258$
+WMH	0.8755	$\mathbf{y = 0.9541x - 0.1169}$
+WMH+Stroke	**0.8916**	$y = 0.9102x - 0.0987$
+Gaussian (mean)	0.8541	$y = 0.9228x - 0.23$
+Gaussian (best)	0.8836	$y = 0.8972x - 0.2629$

models with auxiliary input performed better in volumetric agreement analysis than ones without auxiliary input (No Auxiliary). Only DEP-UResNet with WMH performed slightly lower than DEP-UResNet without auxiliary input. This shows the importance of auxiliary input for predicting the evolution of WMH using deep neural networks.

On the other hand, from all DEP models, DEP-GAN models using IM produced the worst SD and (lower and upper) limits of agreement (LoA) in the volumetric agreement analysis, regardless of the modalities of auxiliary input. This is another indication that IM is not adequate for predicting the evolution of WMH. Interestingly, DEP-GAN using PM, which seemingly had better (lower and upper) LoA than the DEP-GAN using IM, had some of the worst mean of volumetric bias. This indicates that there is a bias towards regression (i.e., shrinking of WMH) when DEP-GAN using PM was used for predicting the evolution of WMH.

From Bland-Altman plots depicted in Figure 5.7, the volumetric agreement of DEP-GAN using PM is similar to the volumetric agreement of DEP-UResNet. In contrast, Bland-Altman plots produced by DEP-GAN using IM show increasing discrepancy and variability between ground truth and predicted volumes with increasing volume of WMH, similar to the results from previous experiment in Section 5.8.1.2. Furthermore, the correlations between ground truth and predicted volumes of WMH for DEP-UResNet and DEP-GAN using PM were much higher than the ones produced by DEP-GAN using IM, especially when auxiliary input is incorporated (see Table 5.5 and Figure 5.8).

5.8.2.2 Spatial agreement (DSC) analysis

On the automatic delineation of WMH change's boundaries in the follow-up year, DEP-UResNet using Gaussian noise produced the best performances with mean DSC of 0.6135 and average of stable, shrinking, and growing WMH clusters with mean DSC of 0.3141 (see "DEP-UResNet+Gaussian (mean)" in Table 5.4 columns 8-13). Furthermore, it also outperformed the rest of the models on changing, shrinking, and growing WMH clusters. These results clearly show the advantage of performing fully supervised learning and modulating Gaussian noise as auxiliary input for predicting the evolution of WMH. It is also worth to mention that its performance could be improved if the "best" Gaussian noise is used and evaluated (see "DEP-UResNet+Gaussian (best)" in Table 5.4 columns 8-13)

Based on Table 5.4 results, the (indirectly supervised) DEP-GAN using PM had close performance to the (supervised) DEP-UResNet in all performed analyses, espe-

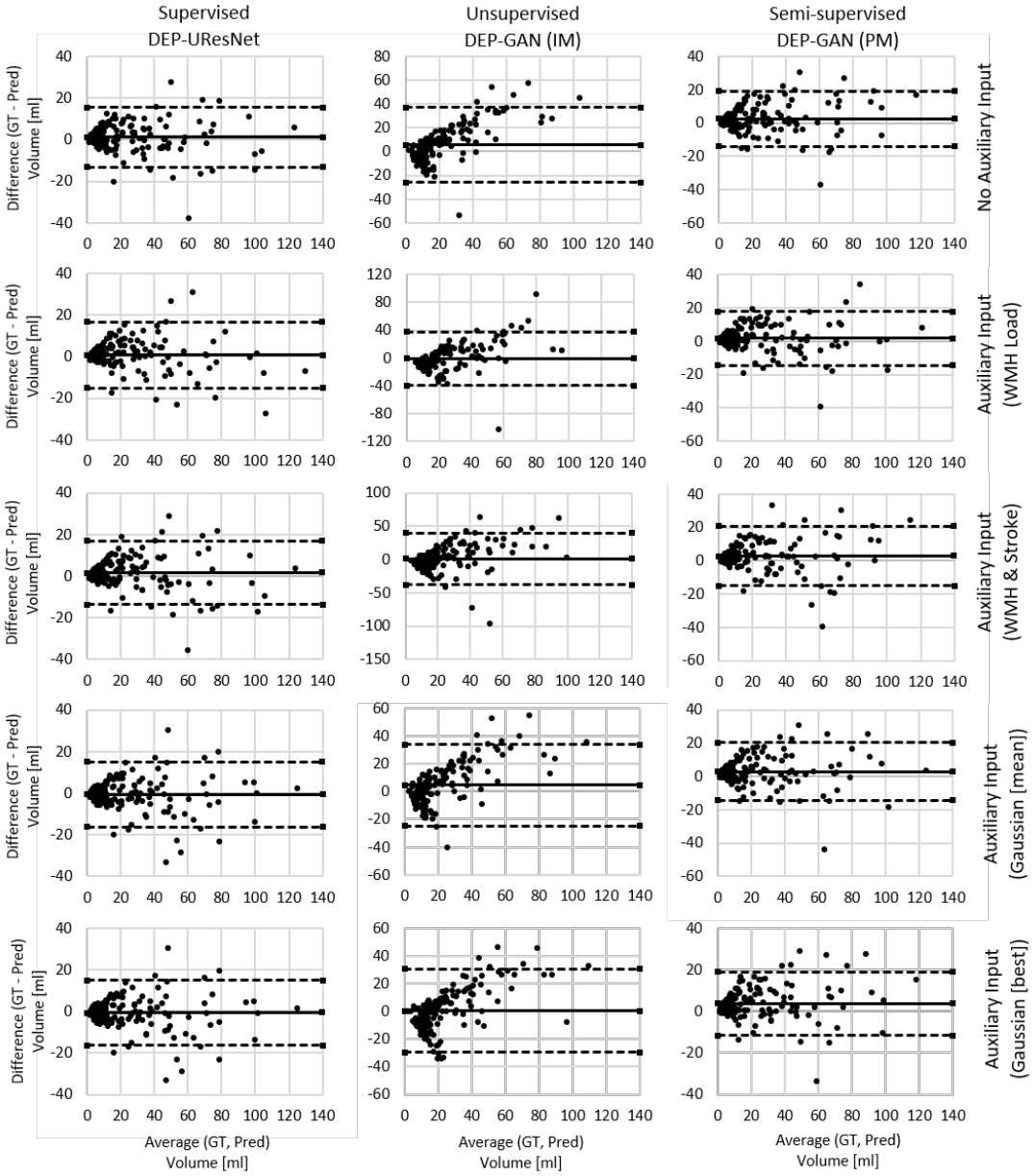


Figure 5.7: Volumetric agreement analysis (in ml) between GT and predicted volume of WMH with different types/modalities of auxiliary input (Pred) using Bland-Altman plot which correspond to data presented in Table 5.4. Solid lines correspond to “Vol. Bias” while dashed lines correspond to either “Lower LoA” or “Upper LoA” of the same table.

cially in the spatial agreement analysis (columns 8-13). To give a better visualisation of the spread of the performances, the distributions of DSC scores for all WMH categories (i.e., entire WMH, changing WMH, shrinking WMH, growing WMH, and stable WMH) produced by all DEP models and different types of auxiliary input were plotted by using box-plot in Figure 5.9. Furthermore, paired two-sided Wilcoxon signed rank

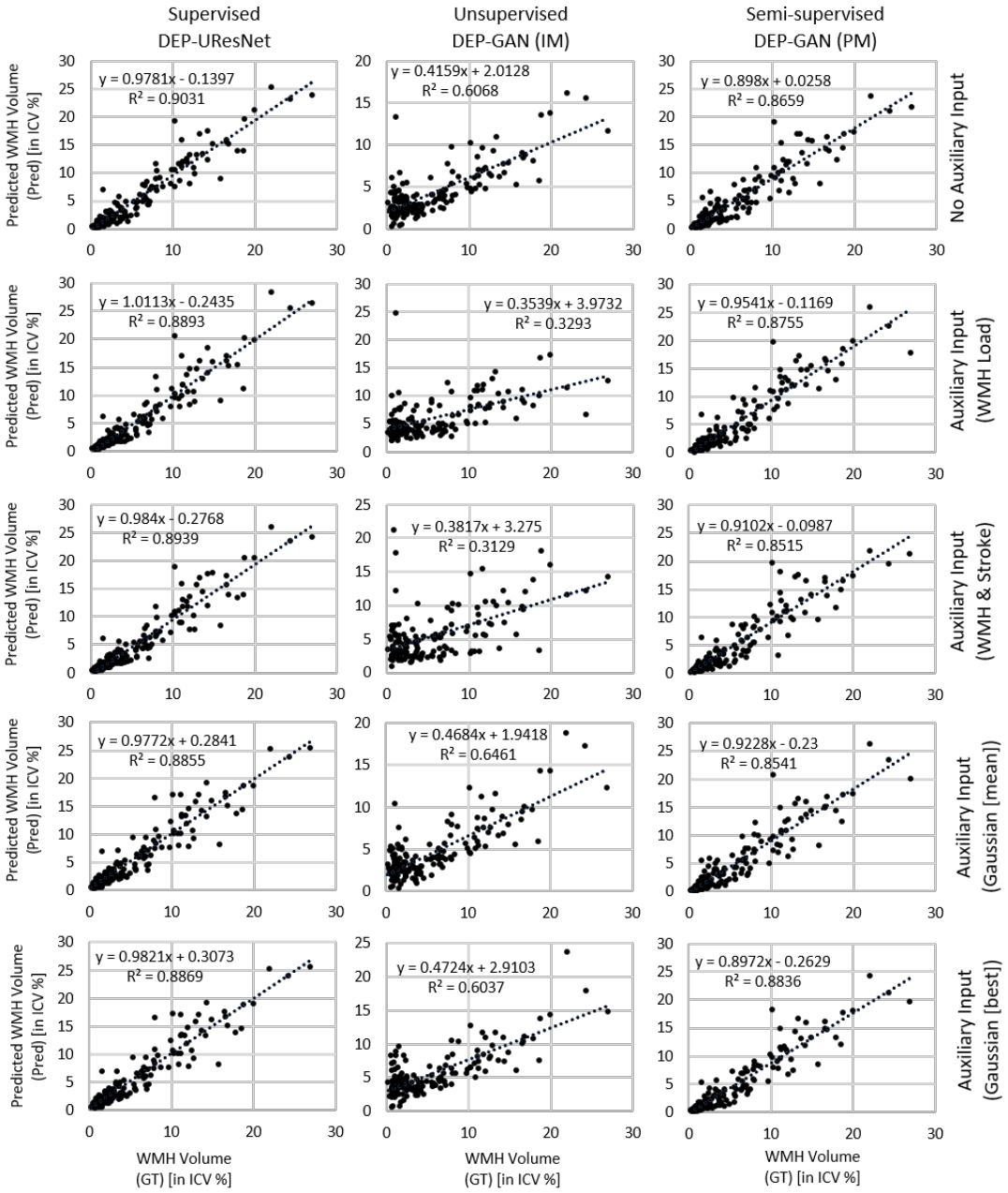
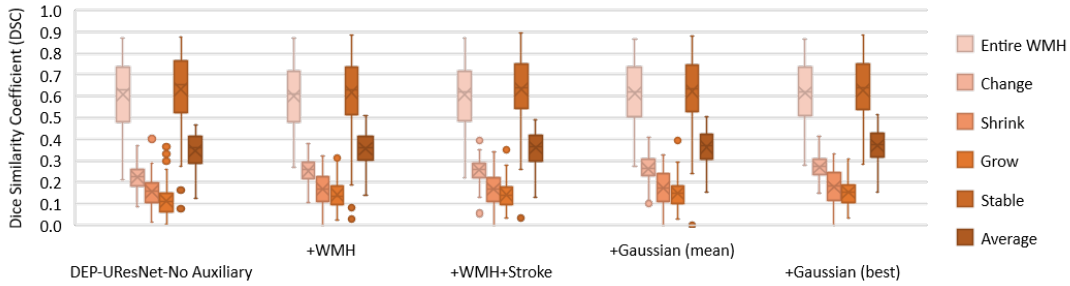


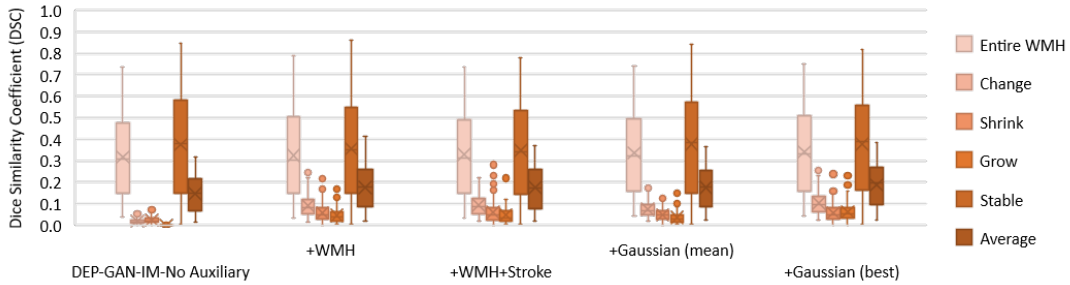
Figure 5.8: Correlation plots between manual WMH volume produced by the expert (GT) and predicted WMH volume by various DEP models with different types/modalities of auxiliary input (Pred). WMH volume is in the percentage of ICV to remove any potential bias associated with head size.

tests were also performed to evaluate whether the medians and distributions of DSC scores produced by the non-supervised DEP-GAN using IM and PM were significantly different to those produced by the supervised DEP-UResNet.

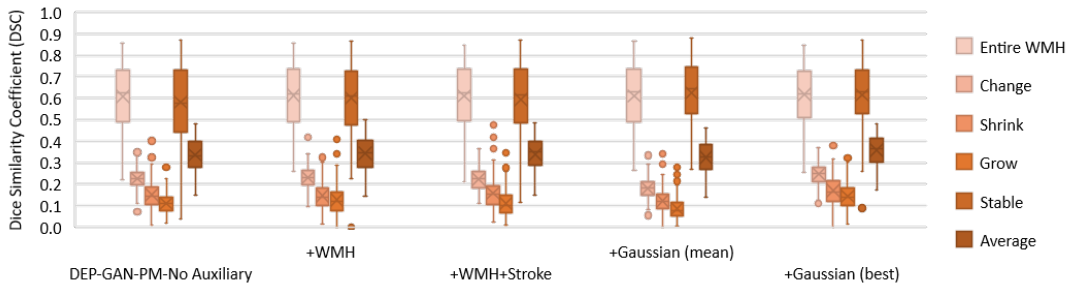
From Figure 5.9, it can be seen that performances of DEP-GAN using PM and



(a) Distributions of DSC scores from the (supervised) DEP-UResNet models.



(b) Distributions of DSC scores from the (unsupervised) DEP-GAN using IM models.



(c) Distributions of DSC scores from the (indirectly supervised) DEP-GAN using PM models.

Figure 5.9: Distributions of DSC scores from all evaluated DEP models in auxiliary input ablation study. These distributions correspond to the Table 5.4, columns 8-13.

DEP-UResNet on delineating different WMH clusters did not differ from each other in term of the distribution of DSC scores. Based on the result from the Wilcoxon tests, there is no significant difference between the performances of DEP-GAN using PM and DEP-UResNet in all WMH clusters, especially when the same auxiliary input was used, with p-value > 0.17. In contrast, the distribution of DSC scores produced by DEP-GAN using IM and DEP-UResNet are significantly different to each other with p-value < 0.0012.

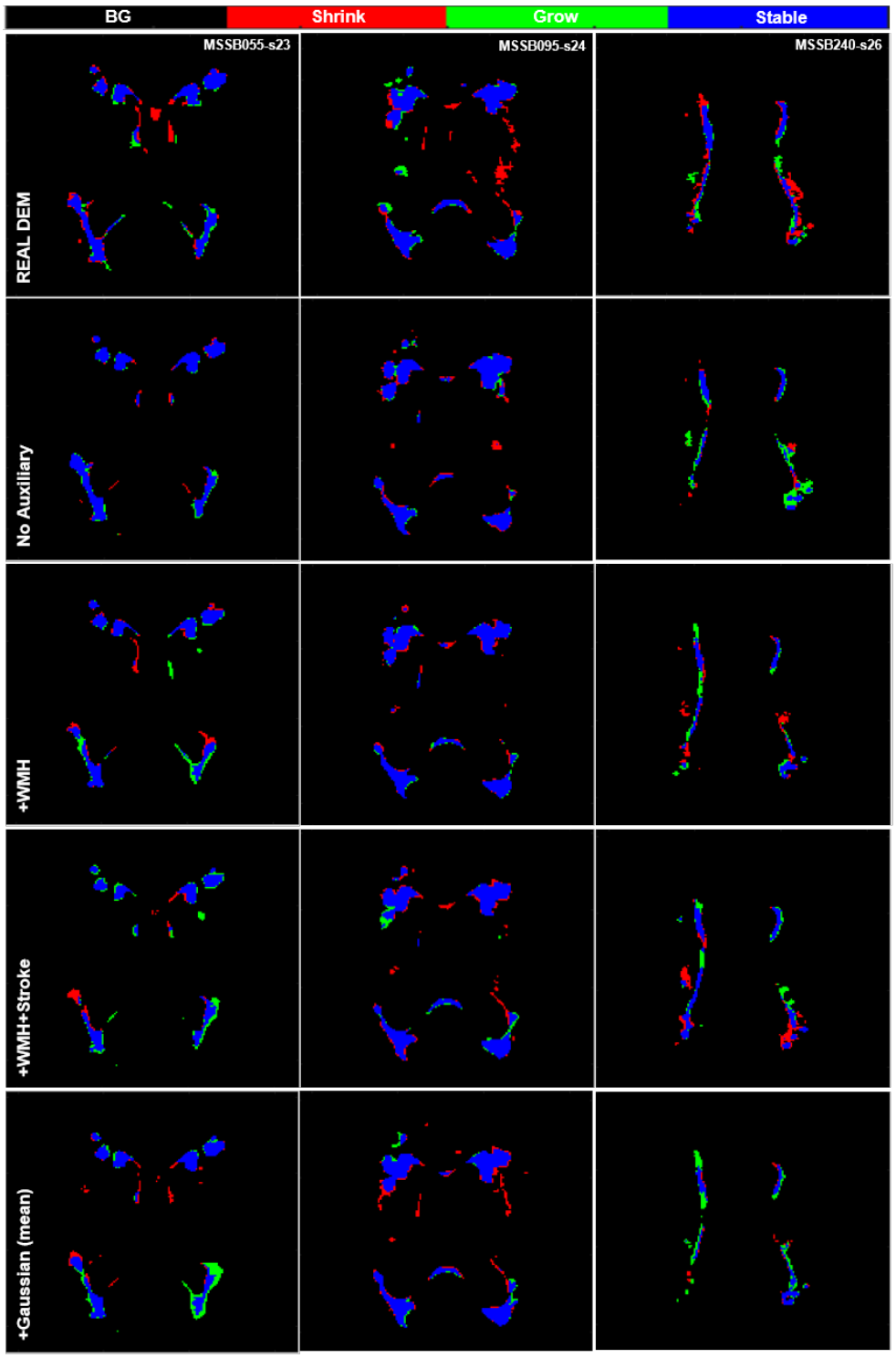


Figure 5.10: Qualitative (visual) assessment of DEM label produced by the supervised DEP model, DEP-UResNet, with different types/modalities of auxiliary input. The corresponding T2-FLAIR (input data) can be seen in Figure 5.12.

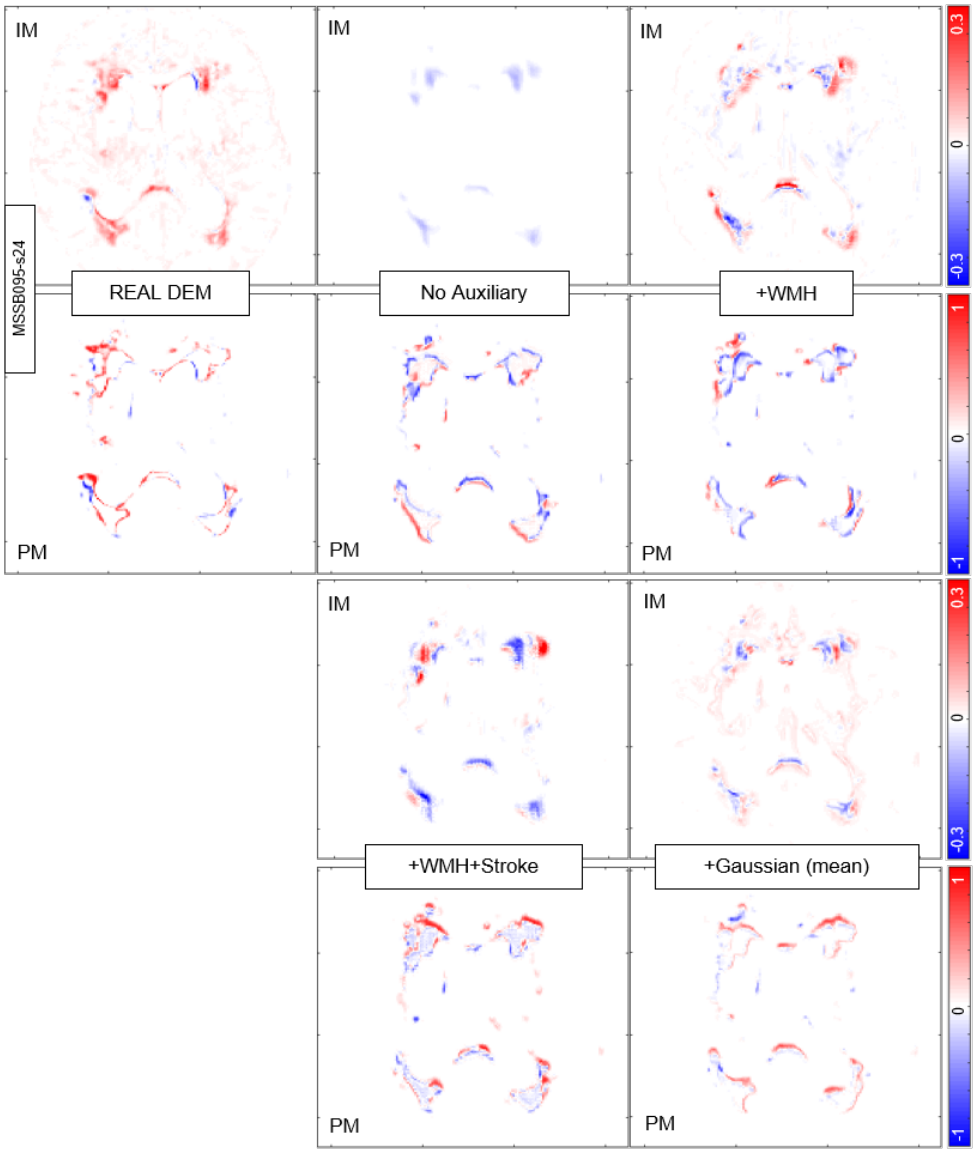


Figure 5.11: Qualitative (visual) assessment of DEM produced by the unsupervised and indirectly supervised DEP models; DEP-GAN using IM and DEP-GAN using PM, with different types/modalities of auxiliary input. The corresponding T2-FLAIR (input data) can be seen in Figure 5.12.

5.8.2.3 Qualitative (visual) analysis

It is worth to mention first that the growing and shrinking regions of WMH are considerably smaller than those unchanged (stable) as depicted in Figure 5.10. Note that it is very difficult to discern the borders between growing and shrinking regions when SL coalesce with WMH even though SL were removed from the analysis as previously explained. Nevertheless, inaccuracies while determining the borders between coalescent

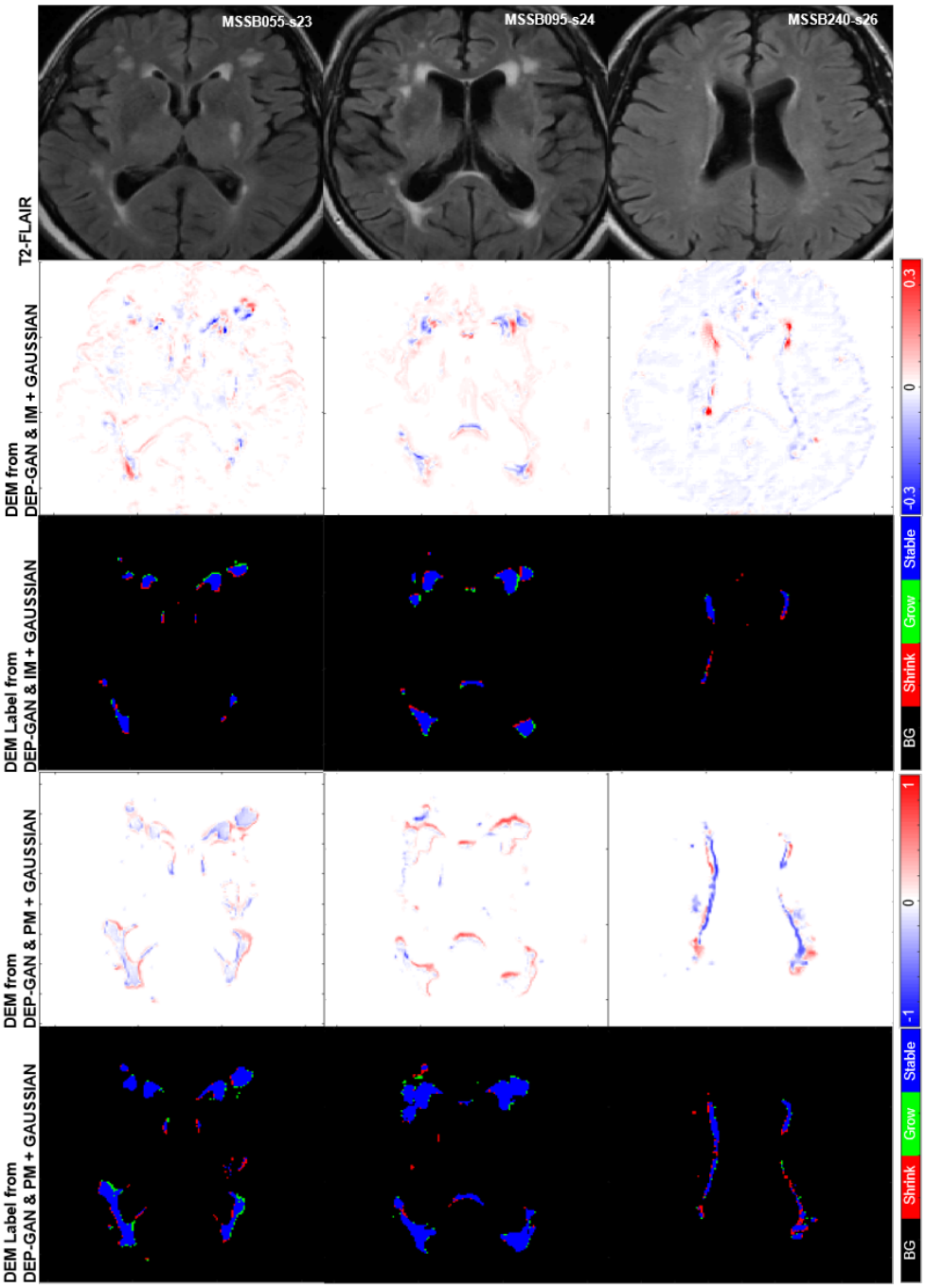


Figure 5.12: Qualitative (visual) assessment of DEM and its corresponding DEM label produced by the unsupervised and indirectly supervised DEP models; DEP-GAN using IM and DEP-GAN using PM respectively, with different types/modalities of auxiliary input. The corresponding golden standard of DEM label can be seen in Figure 5.10.

WMH and SL and the small size of the volume changes in each WMH cluster might have influenced in the low DSC values obtained in the regions that experienced change as seen in Table 5.4. It is also worth to note that most regions of WMH are stable, and

DEP-UResNet and DEP-GAN using PM did not have any problem on segmenting these regions as depicted in Figures 5.10 and 5.12. Furthermore, the small spatial changes of growing and shrinking WMH might not influence the outcome of clinical diagnosis because volumetric changes of WMH did not change drastically in total when different auxiliary inputs were used as depicted in Figures 5.7 and 5.8 (described in Section 5.8.1.2). However, clinical diagnosis outcome might be influenced by the types of DEP model as each of them produced different result characteristics of volumetric agreement and correlation (see Section 5.8.1.2).

Based on qualitative (visual) assessment of DEM produced by DEP-GAN using IM/PM depicted in Figure 5.11, auxiliary input improved the quality of the generated DEMs where they had more correct details than the ones generated without using auxiliary input. However, good details of the generated DEM from IM/PM did not necessarily translate to good three-class DEM label (i.e., three labels of growing, shrinking, and stable WMH) as depicted in Figure 5.12. Some reasons that might have caused this are; 1) the generated DEM from IM/PM is result of a regression process from the baseline IM/PM using DEP-GAN and 2) the three-class DEM label itself is generated from the resulted regression, where WMH is defined by having irregularity/probability values greater than or equal to 0.178 for IM (see Section 4.6.1) and 0.5 for PM. Note that regression of the whole brain using IM/PM is harder than direct segmentation of three regions of WMH (i.e., stable, shrinking, and growing WMH). Furthermore, small changes in IM/PM did not necessarily change the state of voxel from WMH to non-WMH or vice versa. These are the challenges of performing prediction of WMH evolution using DEP-GAN and IM/PM instead of DEP-UResNet.

5.8.2.4 Clinical plausibility analysis

From Table 5.6, it can be seen that the use of expert-delineated binary WMH masks and WMH maps obtained from thresholding IM or PM (see the second to the fourth rows), all produced the same ANCOVA model's results; none of the covariates of the model had an effect in the 1-year WMH volume change, yielding almost identical numerical results in the first two decimal places. Therefore, the use of LOTS-IM and UResNet, generators of the IM and PM respectively, for producing WMH maps in clinical studies of mild to moderate stroke seems plausible.

As discussed in Section 5.1, baseline WMH volume has been recognised the main predictor of WMH change over time (Chappell et al., 2017; Wardlaw et al., 2017), although the existence of previous SL and hypertension have been acknowledged as

Table 5.6: Results from the ANCOVA models that investigate the effect of several clinical variables (i.e. stroke subtype, stroke-related imaging markers, and vascular risk factors) in the WMH volume change from baseline to one year after. The first column at the left hand side refers to the models/methods used to obtain the follow-up WMH volume used in the ANCOVA models as outcome variable. The rest of the columns show the coefficient estimates B and the significance level given by the p-value (i.e. B(p)), for each covariate included in the models.

Reference (binary mask)	**Stroke lacunar**	**BG PVS scores**	**Diabetes (y/n)**	**Hypertension (y/n)**	**Smoker (y/n)**	**Index SL (% in ICV)**	**Old SL (% in ICV)**
Expert-delineated	-0.04(0.65)	0.07(0.25)	-0.10(0.48)	-0.05(0.66)	-0.07(0.42)	-0.03(0.46)	0.13(0.15)
Thresholded IM	-0.04(0.66)	0.08(0.19)	-0.12(0.44)	-0.04(0.71)	-0.09(0.38)	-0.03(0.43)	0.14(0.14)
Thresholded PM	-0.04(0.66)	0.08(0.19)	-0.12(0.44)	-0.04(0.71)	-0.09(0.38)	-0.03(0.43)	0.14(0.14)
Supervised (DEP-UResNet)	**Stroke lacunar**	**BG PVS scores**	**Diabetes (y/n)**	**Hypertension (y/n)**	**Smoker (y/n)**	**Index SL (% in ICV)**	**Old SL (% in ICV)**
No Auxiliary	-0.12(0.11)	0.10(0.03)	-0.06(0.57)	0.03(0.73)	-0.08(0.29)	-0.04(0.14)	0.30(<0.001)
+WMH	-0.10(0.13)	0.11(0.006)	0.04(0.65)	0.01(0.87)	-0.05(0.38)	-0.04(0.13)	0.20(<0.001)
+WMH+Stroke	-0.07(0.29)	0.06(0.14)	0.07(0.48)	-0.02(0.75)	-0.10(0.15)	-0.05(0.10)	0.32(<0.001)
+Gaussian (mean)	-0.09(0.26)	0.11(0.04)	0.06(0.61)	0.02(0.81)	-0.10(0.21)	-0.06(0.08)	0.36(<0.001)
Unsupervised (DEP-GAN & IM)	**Stroke lacunar**	**BG PVS scores**	**Diabetes (y/n)**	**Hypertension (y/n)**	**Smoker (y/n)**	**Index SL (% in ICV)**	**Old SL (% in ICV)**
No Auxiliary	0.03(0.68)	-0.03(0.58)	-0.07(0.54)	0.0006(0.99)	-0.08(0.33)	-0.11(0.001)	0.25(0.001)
+WMH	0.22(0.09)	0.08(0.36)	-0.004(0.98)	0.12(0.40)	-0.08(0.54)	-0.06(0.25)	0.32(0.01)
+WMH+Stroke	-0.11(0.45)	-0.08(0.40)	0.03(0.88)	0.10(0.53)	0.11(0.47)	-0.02(0.77)	0.34(0.02)
+Gaussian (mean)	-0.02(0.86)	-0.07(0.24)	-0.06(0.69)	-0.05(0.62)	-0.07(0.43)	-0.14(0.0004)	0.20(0.03)
Indirectly Spv. (DEP-GAN & PM)	**Stroke lacunar**	**BG PVS scores**	**Diabetes (y/n)**	**Hypertension (y/n)**	**Smoker (y/n)**	**Index SL (% in ICV)**	**Old SL (% in ICV)**
No Auxiliary	-0.10(0.24)	0.14(0.009)	0.10(0.45)	0.04(0.67)	-0.03(0.70)	-0.05(0.18)	0.18(0.03)
+WMH	-0.03(0.72)	0.09(0.09)	-0.14(0.31)	-0.04(0.68)	-0.06(0.46)	-0.04(0.30)	0.19(0.03)
+WMH+Stroke	-0.10(0.28)	0.17(0.006)	0.10(0.50)	0.10(0.36)	-0.02(0.81)	-0.08(0.05)	0.24(0.01)
+Gaussian (mean)	-0.09(0.25)	0.10(0.04)	0.02(0.87)	-0.0001(0.99)	-0.08(0.27)	-0.04(0.17)	0.14(0.05)

contributed factors. However, from the results of the ANCOVA models (Table 5.6), none of the DEP models that used these (i.e WMH and/or SL volumes) as auxiliary inputs showed similar performance (i.e. in terms of strength and significance in the effect of all the covariates in the WMH change) as the reference WMH maps. The only DEP model that shows promise in reflecting the effect of the clinical factors selected as covariates in WMH progression was the DEP-GAN that used as input the PM of baseline WMH and Gaussian noise (i.e. written in bold and underlined in the left hand side column of Table 5.6).

Some factors might have adversely influenced the performance of these predictive models. First, all deep-learning schemes require a very large amount of balanced (e.g.

in terms of the appearance, frequency and location of the feature of interest, i.e. WMH in this case) data, generally not available. The lack of data available imposed the use of 2D model configurations, which generated unbalance in the training: for example, not all axial slices have the same probability of WMH occurrence, also WMH are known to be less frequent in temporal lobes and temporal poles are a common site of artefacts affecting the IM and PM, error that might propagate or even be accentuated when these modalities are used as inputs. Second, the combination of hypertension, age and the extent, type, lapse of time since occurrence and location of the stroke might be influential on the WMH evolution, therefore rather than a single value, the incorporation of a model that combines these factors would be beneficial. However, such model is still to be developed also due to lack of data available. Third, the tissue properties have not been considered. A model to reflect the brain tissue properties in combination with vascular and inflammatory risk factors is still to be developed. Lastly, the deep-learning models as we know them, although promising, are reproductive, not creative. The development of more advanced inference systems is paramount before these schemes can be used in clinical practice.

5.8.2.5 Prediction error analysis and discussion

From Table 5.4 (columns 2-4), it can be seen that all DEP models tested in this ablation study could correctly predict the progression/regression of WMH volume better than a random guess system ($\geq 50\%$). Furthermore, it also can be seen that DEP models with auxiliary input, either Gaussian noise or known risk factors of WMH evolution (i.e., WMH and SL loads), produced better performances in most cases and evaluation analyses than the DEP models without any auxiliary input. These results show the importance of auxiliary input, especially Gaussian noise which simulates the non-deterministic nature of WMH evolution. Furthermore, it is clear now that PM is better for representing the evolution of WMH than IM when DEP-GAN is used, especially if ones would like to have good volumetric agreement and correlation, spatial agreement, and clinical plausibility of the WMH evolution.

5.8.3 Ablation study of the DEP-GAN's regularisation terms

In this study, three regularisation terms are proposed for DEP-GAN (i.e., intensity, DSC, and volume) instead of one term (i.e., only intensity) like in the VA-GAN. Table 5.7 shows prediction results where the weights of each term are set to 0 to investigate how

Table 5.7: Results from ablation study of the DEP-GAN's regularisation terms tested using PM (see Equation 5.4). The prediction error of WMH change, volumetric agreement of WMH volume, and spatial agreement of WMH evolution were calculated and compared to the gold standard expert-delineated WMH masks (i.e., LBL-DEM). "Vol." stands for volumetric and "G" and "S" stand for percentage of subjects correctly predicted as having growing and shrinking WMH by DEP models. The best value for each learning approaches and evaluation measurements is written in bold.

DEP-GAN (PM)			**Grow**	**Shrink**	**Avg.** [%]	**Vol. Bias** [*ml*]	**Lower**	**Upper**	**Entire**	**Change**	**Stable**	**Shrink**	**Grow**	**Avg.** ((St+
λ_1	λ_2	λ_3	**(G)** [%]	**(S)** [%]	((G+S)/2)	*mean(SD)*	**LoA** [*ml*]	**LoA** [*ml*]	**WMH**	**(C)**	**(St)**	**(Sr)**	**(Gr)**	Sr+Gr)/3)
0	0	0	64.29	85.19	**74.74**	3.03(7.65)	-11.9684	**18.0372**	0.6131	0.1667	0.6178	0.1045	0.0813	0.2679
0	0	100	65.31	79.63	72.47	2.28(8.16)	-13.7197	18.2747	**0.6132**	0.1749	0.6166	0.1009	0.0909	0.2695
0	1	0	50.00	83.33	66.67	4.32(8.18)	-11.7181	20.3473	0.6093	0.1919	0.6063	0.1366	0.0706	0.2712
100	0	0	57.14	83.33	70.24	3.79(7.83)	-11.5525	19.1234	0.6075	0.1827	0.6143	0.1312	0.0741	0.2732
0	1	100	**67.35**	75.93	71.64	2.37(8.50)	-14.2904	19.0237	0.6101	0.1889	0.6177	0.1203	0.0922	0.2767
100	1	0	58.16	77.78	67.97	**2.23(8.85)**	-15.1197	19.5748	0.6096	0.1912	0.6079	0.1209	**0.0925**	0.2738
100	0	100	57.14	**88.89**	73.02	4.51(8.15)	**-11.4546**	20.4778	0.6078	**0.1993**	0.5996	**0.1446**	0.0760	0.2734
100	1	100	56.12	81.48	68.80	3.46(8.26)	-12.7218	19.6500	0.6107	0.1801	**0.6245**	0.1216	0.0868	**0.2776**

each of these three terms affect the prediction results. Note that λ_1 is the weight for intensity loss, λ_2 is the weight for DSC loss, and λ_3 is the weight for volumetric loss (see Equation 5.4). This ablation study was performed using DEP-GAN-2C using PM.

From this ablation study, the use of more terms in the regularisation had a positive impact in the prediction results. It is expected because multiple terms forced the DEP-GAN's generator to generalise and perform well on all important measurements used in the evaluation of the prediction of WMH evolution, i.e., intensities in the regression of PM's values, WMH segmentation correctness in DSC, and volumetric prediction of WMH. However, it is worth mentioning that the improvements were limited and still could be improved in the future.

5.9 Conclusion and Future Work

In this study, an end-to-end training scheme was proposed to predict the evolution of WMH using deep learning algorithms called DEP model. To the best of our knowledge, this is the first extensive study on modelling WMH evolution using deep learning algorithms. Different configurations of DEP models (i.e., unsupervised (DEP-GAN using IM), indirectly supervised (DEP-GAN using PM), and supervised (DEP-UResNet)) with different types of auxiliary input (i.e., Gaussian noise, WMH load, and WMH and SL loads) were evaluated. These configurations were designed and evaluated to find

the best approach to automatically predict and delineate the evolution of WMH from a baseline measurement to a follow-up visit.

Based on the two ablation analyses done as part of the present study, DEP-GAN-2C performed better than WGAN-GP, VA-GAN, and DEP-GAN using 1 critic. Furthermore, Gaussian noise successfully improved all DEP models in almost all evaluation measurements when used as auxiliary input. This shows that there are indeed some unknown factors that influence the evolution of WMH. These unknown factors make the problem of predicting/delineating WMH evolution non-deterministic, and Gaussian noise were proposed to simulate this scenario. The intuition behind this approach is that Gaussian noise fills in the missing (unavailable) risks factors or their combination, which could influence the evolution of WMH. Note that it is very challenging to collect and compile all risk factors of WMH evolution in a longitudinal study.

From the experiments, on average, supervised DEP-UResNet yielded the best results in almost every evaluation measurement. However, it is worth to mention that it did not perform well in the clinical plausibility test. The indirectly supervised DEP-GAN yielded similar average performance to the supervised DEP-UResNet's performance and yielded the best results out of all schemes in the clinical plausibility test. Moreover, results from DEP-UResNet and DEP-GAN using PM were not statistically different to each other on delineating the WMH clusters.

If we consider the results, time, and resources spent in this study, then DEP-GAN using PM showed the biggest and strongest potential of all DEP models. Not only did it perform similarly to the supervised DEP-UResNet but it also did not need manual WMH labels on two MRI scans for training (i.e., baseline and follow-up scans). The PM needed as input for this model can be efficiently produced by any supervised deep/machine learning model. Moreover, the development of automatic WMH segmentation for producing better PM could be done separately and independently from the development of the DEP model. If a better PM model is available in the future, then the DEP-GAN model can be retrained using the newly produced PM for better performance. Also, DEP-GAN using PM could be used for other (neuro-degenerative) pathologies, as long as a set of PM from these other pathologies could be produced and used to (re-)train the DEP-GAN.

There are several shortcomings anticipated from the results of this study. Firstly, manual WMH labels of two MRI scans (i.e., baseline and follow-up scans) are necessary for training the DEP-UResNet. In many scenarios, this is not applicable and efficient in terms of time and resources. Secondly, the unsupervised DEP-GAN using IM is

computationally very demanding as it involves regressing IM values across the whole brain tissue. This resulted in low performances of DEP-GAN using IM in almost all evaluation measurements. Thirdly, the schemes' performances depend on the accuracy of the quality of input. For example, the PM generated in this study are slightly biased towards overestimating the WMH in the optical radiation and underestimating WMH in the frontal lobe. This could be caused by the absence of correcting the T2-FLAIR images for b1 magnetic field inhomogeneities. However, a previous study on small vessel disease images demonstrated this procedure might affect the results underestimating the subtle white matter abnormalities characteristics of this disease, and recommends this procedure to be used in T1- and T2-W structural images but not in T2-FLAIR images for WMH segmentation tasks (Valdés Hernández et al., 2016) Hence, the biggest challenge of using DEP-GAN using PM is its highly dependency on the quality of initial PM. Fourthly, volumetric agreement analyses suggest that there are still large differences in absolute volume and in change estimates produced by the proposed DEP models. While this study is intended as a "proof-of-principle" study to advance the field of white matter - and ultimately brain- health prediction, it is worth to mention that better reliability in the WMH assessment is necessary so as DEP models can be used in clinical practice. Furthermore, better understanding of what DEP models extract to estimate WMH evolution would be very useful in clinical practice. Lastly, the limitation of using (Gaussian) random noise in DEP models is the fact that we do not really know which set of Gaussian random noise should be used to generate the best result for each subject. Note that, in this study, all DEP models that used Gaussian noise as auxiliary input were tested 10 times to calculate the mean and the "best" set of Gaussian noise which produced the best automatic delineation of WMH evolution overall. In conclusion, DEP models suffer similar problems and limitations to any machine learning based medical image analysis methods.

The DEP models proposed in this study open up several possible future avenues to further improve their performances. Firstly, multi-channel (e.g., PM and T2-FLAIR) input could be used instead of single channel input. In this study, single channel input was used to draw a fair comparison between DEP-UResNet which uses T2-FLAIR and DEP-GAN which uses either IM or PM. Secondly, 3D architecture of DEP-GAN could be employed when more subjects are accessible in the future. 3D deep neural networks have been reported to have better performances than the 2D ones, but they are more difficult to train (Çiçek et al., 2016; Baumgartner et al., 2018). Thirdly, Gaussian noise and known risk factors (e.g., WMH and SL loads) could be modulated

together instead of modulating them separately in different models. By modulating them together, DEP model would be influenced by both known (available) risk factors and unknown (missing) factors represented by Gaussian noise. Lastly, different random noise distribution could be used instead of Gaussian distribution. Note that each risk factors of WMH evolution (e.g., WMH load, age, and blood pressure) could have different data distribution, not only Gaussian distribution. If a specific data distribution (i.e., the same or similar to the real risk factor's data distribution) could be used for a specific risk factor, then the real data could replace the random noise if available in the testing.

Chapter 6

Conclusion and Future Work

The preceding chapters of this thesis have described the development of segmentation, characterisation, and evolution prediction methods for WMH in structural brain MRI. In this chapter, a general summary of this thesis is provided. Furthermore, contributions, impacts, and possible future investigations for each chapter (study) are also discussed.

6.1 Summary

WMH are neuroradiological features seen in T2-FLAIR and have been commonly associated with stroke, ageing, and dementia progression (Wardlaw et al., 2013). Recent studies have shown that WMH may shrink (i.e., regress), stay unchanged (i.e., stable), or grow (i.e., progress) over a period of time (Ramirez et al., 2016; Wardlaw et al., 2017). The objective of this thesis is to propose automatic methods for segmentation, characterisation, and evolution prediction of WMH that can be used in clinical research to estimate the size and location of WMH in time to study their progression/regression in relation to clinical health and disease indicators, for ultimately designing more effective therapeutic interventions.

Chapter 3 tackles the problem of segmenting early (i.e., small and subtle) WMH using CNNs and GSI. Segmenting early WMH is crucial for early detection of dementia and AD GSIand longitudinal study of dementia's progression. In this study, synthetic GSI is incorporated to the patch-based CNNs as additional input channels, and it successfully improves the performance of CNNs to segment small WMH. Thus, showing that spatial information is important for the segmentation of WMH.

Chapter 4 describes a novel unsupervised method to characterise the WMH named LOTS-IM. LOTS-IM produces an IM which describes the intensities of WMH by real

values between 0 to 1 instead of binary label which describes the hard border of WMH. IM is also better than the well known PM as it can well characterise both non-WMH and WMH regions, including the WMH "penumbra". The WMH penumbra is especially important for the study of WMH progression (Kapeller et al., 2003; Bendfeldt et al., 2009; Callisaya et al., 2013). This chapter also describes the simulation of progression and regression of WMH over a period of time using IM. In the evaluation of WMH segmentation, it is shown that LOTS-IM outperforms LST-LGA (i.e., the current state-of-the-art unsupervised WMH segmentation method), conventional supervised machine learning algorithms (i.e., SVM and RF), and some supervised deep learning algorithms (i.e., DBM and CEN). Furthermore, the results also show that LOTS-IM has comparable performance with the state-of-the-art supervised deep learning algorithms (DeepMedic, UResNet, and UNet). Whereas, the biggest limitation of the proposed simulation of WMH progression and regression using IM is that it is not based on real longitudinal data (i.e., not a data-driven method).

Finally, **Chapter 5** describes novel deep learning models for predicting and estimating the evolution (i.e., progression and regression) of WMH using longitudinal data, addressing the limitation of our previous study on simulation of brain abnormalities using IM described in Chapter 4. In this experiment, an end-to-end model called DEP model which uses deep learning and an auxiliary input module is proposed and evaluated for the prediction of WMH evolution from baseline to follow-up while addressing the non-deterministic nature of this process. Two models of DEP are proposed, which are DEP-UResNet and DEP-GAN, representing supervised and unsupervised deep learning algorithms respectively. DEP-UResNet uses baseline T2-FLAIR as the main input while DEP-GAN uses either baseline IM or PM instead. To simulate the non-deterministic and unknown parameters involved in WMH evolution, a modulation of Gaussian noise array to the DEP model as an auxiliary input is proposed. This forces the DEP model to imitate a wider spectrum of alternatives in the results. Based on the results, DEP-GAN using PM and Gaussian noise as an auxiliary input yielded one of the best results in almost all evaluations, including clinical plausibility. The DEP-UResNet regularly performed better than the DEP-GAN using PM and Gaussian noise in some evaluations, but eventually it did not show promise in the clinical evaluation.

6.2 Contributions of this thesis

The main contribution of this doctoral thesis is the development of methods for early WMH segmentation, characterisation of WMH, and prediction of WMH evolution using machine learning algorithms. This thesis is arranged such that each chapter contributes to one of the contributions. The contributions of each chapter (i.e. study, set of experiments) in this thesis are discussed below.

Spatial information for early WMH segmentation: From previous studies, it has been suggested that WMH are only the "tip of the iceberg" where they represent the extreme end of a continuous spectrum of white matter injuries (Zhang et al., 2013; Lockhart et al., 2012; Wardlaw et al., 2015). This is especially a challenge for early WMH segmentation as they appear very subtle and are indistinguishable from the non-WMH (i.e., healthy) regions in T2-FLAIR brain MRI. In this thesis (i.e., Chapter 3), it has been shown that spatial information is a good additional prior knowledge to the textures of T2-FLAIR brain MRI for achieving a good segmentation of early WMH (i.e. small and subtle WMH).

Irregularity map for characterisation of WMH: The newly proposed IM, described in Chapter 4, is unique, and differs from the PM and binary WMH labels as it is able to represent not only the WMH but also the non-WMH regions in T2-FLAIR brain MRI. Unlike PM and binary WMH labels, IM is also able to represent the "penumbra of WMH" (Maillard et al., 2011), which has been suggested to be important for the study of WMH progression (Maillard et al., 2014; Pasi et al., 2016). Furthermore, IM has a good performance as unsupervised WMH segmentation approach and can be used for the simulation of WMH progression and regression.

Automatic spatial estimation of WMH Evolution: Predicting the evolution (i.e., progression and regression) of WMH is a challenging task, especially because it involves both commonly known and unknown clinical risk factors. In other words, evolution of WMH is a non-deterministic (probabilistic) process. In Chapter 5 where DEP models are described, it has been shown that the non-deterministic nature of WMH evolution can be well simulated by using Gaussian noise as an auxiliary input of the DEP model. Furthermore, it has been shown that the unsupervised model of DEP-GAN using PM performed statistically similar to the supervised model of DEP-UResNet on estimating the spatial evolution of WMH. It also has been suggested in this thesis that DEP-GAN using PM and Gaussian noise performed the best in clinical plausibility test, outperforming the DEP-UResNet. To the best of our knowledge, DEP model is the first

predictive model using deep learning to estimate the evolution of pathology in medical image analysis research field.

6.3 Future Work

In this section, a number of possible future investigations is discussed.

Development of better irregularity map: While IM has shown promising results for the characterisation of WMH, its accuracy on WMH segmentation and estimation of WMH evolution are still worse than the PM produced by deep learning algorithms. Better means of target patch sampling and distance value calculation might improve the quality of IM for WMH segmentation. Furthermore, the use of 3D voxels might also improve IM quality.

Main input of the DEP model: The performance of DEP models, both DEP-UResNet and DEP-GAN, still can be improved by using multi-channel inputs instead of one-channel input. While DEP-GAN using PM and Gaussian noise performed better than other DEP models, its main input (i.e., PM) might be not enough for representing the actual patient/subject clinical condition. Thus, combination of MRI modalities (e.g., T2-FLAIR, T2-W, and T1-W), IM, and PM as main input channels might improve the performance of DEP-GAN.

3D convolutional layer for DEP model: In this thesis, 2D convolutional layer is used for DEP model because the number of available longitudinal data is limited. However, multiple previous studies have reported that 3D convolutional layer improved the performance of deep neural networks in medical image analysis (Çiçek et al., 2016; Kamnitsas et al., 2017). The biggest challenge of this future work would be the availability of longitudinal data for training, as 3D deep neural networks are more difficult to train (Baumgartner et al., 2018).

Better representation of clinical risk factors of WMH evolution: A good predictive modelling of WMH evolution should be able to represent at least two things: 1) the non-deterministic nature and unknown (or missing) risk factors of WMH evolution and 2) commonly known clinical risk factors of WMH evolution. This thesis has shown that the non-deterministic nature and unknown risk factors of WMH evolution can be represented by using Gaussian noise. However, in this thesis, the Gaussian noise was used in substitution of the clinical risk factors. In the future, both of known and unknown risk factors of WMH evolution should be considered in the DEP model.

6.4 List of Publications

In this section, all publications (i.e., paper and software) authored and co-authored by the author of this thesis while doing his doctoral study in medical image analysis research field are listed.

6.4.1 Papers in international journals

1. **Rachmadi, M. F.**, Valdés-Hernández, M. D. C., Li, H., Guerrero, R., Meijboom, R., Wiseman, S., Waldman, A., Zhang, J., Rueckert, D., Wardlaw, J., and Komura, T. (2020). Limited One-time Sampling Irregularity Map (LOTS-IM) for automatic unsupervised assessment of white matter hyperintensities and multiple sclerosis lesions in structural brain magnetic resonance images. *Computerized Medical Imaging and Graphics*, 79:101685.

2. Malla, P., Uziel, C., Valdés-Hernández, M. D. C., **Rachmadi, M. F.**, & Komura, T. (2019). Evaluation of enhanced learning techniques for segmenting ischaemic stroke lesions in brain magnetic resonance perfusion images using a convolutional neural network scheme. *Frontiers in Neuroinformatics*, 13, 33.

3. Jeong, Y., **Rachmadi, M. F.**, Valdés-Hernández, M. D. C., & Komura, T. (2019). Dilated saliency U-Net for white matter hyperintensities segmentation using irregularity age map. *Frontiers in Aging Neuroscience*, 11, 150.

4. **Rachmadi, M. F.**, Valdés-Hernández, M. D. C., Agan, M. L. F., Di Perri, C., Komura, T., & Alzheimer's Disease Neuroimaging Initiative. (2018). Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Computerized Medical Imaging and Graphics*, 66, 28-43.

5. **Rachmadi, M. F.**, Valdés-Hernándezez, M. d. C., Agan, M. L. F., and Komura, T. (2017a). Deep learning vs. conventional machine learning: Pilot study of WMH segmentation in brain MRI with absence or mild vascular pathology. Journal of Imaging, 3(4):66.

Submitted:

1. **Rachmadi, M. F.**, Valdés-Hernández, M. D. C., Makin, S., Wardlaw, J. M., & Komura, T. (2019). Automatic spatial estimation of white matter hyperintensities

evolution in brain MRI using disease evolution predictor deep neural networks. *bioRxiv*, 738641. Submitted to *Medical Image Analysis* (in revision).

6.4.2 Papers in conference proceedings

1. **Rachmadi, M. F.**, del C. Valdés-Hernández, M., Makin, S., Wardlaw, J. M., and Komura, T. (2019a). Predicting the evolution of white matter hyperintensities in brain MRI using generative adversarial networks and irregularity map. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 146–154, Cham. Springer International Publishing.

2. **Rachmadi, M. F.**, Valdés-Hernández, M. d. C., and Komura, T. (2018c). Automatic irregular texture detection in brain MRI without human supervision. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 506–513, Cham. Springer International Publishing.

3. **Rachmadi, M. F.**, Valdés-Hernández, M. D. C.,M., and Komura, T. (2018a). Transfer learning for task adaptation of brain lesion assessment and prediction of brain abnormalities progression/regression using irregularity age map in brain MRI. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 85–93, Cham. Springer International Publishing.

4. **Rachmadi, M. F.**, Valdés-Hernández, M. d. C., and Komura, T. (2017c). Voxel-based irregularity age map (IAM) for brain's white matter hyperintensities in MRI. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 321–326. IEEE.

5. **Rachmadi, M. F.**, Valdés-Hernández, M. d. C., Agan, M. L. F., and Komura, T. (2017b). Evaluation of four supervised learning schemes in white matter hyperintensities segmentation in absence or mild presence of vascular pathology. In *Medical Image Understanding and Analysis*, pages 482–493, Cham. Springer International Publishing.

6.4.3 Publicly published software

1. **DBM for MRI**
   URL: `https://github.com/febrianrachmadi/boltzmannmachine`

2. **LOTS-IM**

   URL: `https://github.com/febrianrachmadi/lots-iam-gpu`

3. **DEP model**

   URL: `https://github.com/febrianrachmadi/dep-gan-im`

Appendix A

Supplementary Materials

In this appendix, we attached proformas filled in by a neuroradiologist for neuroradiological evaluation described in Section 3.6.11.

018_S_2133_2012 CLEAN u

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM: Anterior WM tract	DWM: Central WM tract	DWM: Posterior WM tract	Pons	PVWM	BG	DWM: Anterior WM tract	DWM: Central WM tract	DWM: Posterior WM tract	Pons		PVWM			Basal Ganglia		DWM: Anterior WM tract			DWM: Central WM tract			DWM: Posterior WM tract		
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N
1													8						8	16	2		17					10
2													6						4	13	2		18					9
3													3						7	12	2		18					10
4													7						4	11	2		12					9
5													5						9	14			9					7
6													5						6	14			10					7

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

009_5_0751_2012 CLEAN 4

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N		
	1																		1	2			3			4	4			
	2			1															1	2			2			*	3			
	3																		2	2			2				2			
	4																		1	2			2				2			
	5																		1	2			4				3			
	6																		1	2			2				3			

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

002_S_0413_2012 CLEAN 4

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N	C	A	N
1													6					7			2	6	15					5		
2													5					6			2	6	12					5		
3													4					4			1	8	11					5		
4													3					7				7	9					4		
5				1									3					9				13	10					3		
6				1									4					4				9	13					4		

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

002_S_2073_2012 CLEAN 4

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N		
1				3									8	8					7	8	1			19		2	5			
2				3										10					8	8*	1			19		1	4			
3				3										3					11	8	1			21		2	5			
4				5										6					12	5	1			22		2	4			
5				4										4					9	4	2			14		2	3			
6				4										6					10	2				15		2	4			

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

002_S_4799_2012 CLEAN4

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N		
1														10					12	11	2		15			15	2			
2														8					12	10	2		14			8	2			
3														5					8	11	2		12			0	2			
4														2					10	8	4		13			8	2			
5														2					10	8	2		13			7	2			
6														4					10	7	1		14			5	4			

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

035_S_2074_2012 clean 4

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N		
1													1					1		2	2			6			7			
2													1					2		2	2			9			2			
3													0					1		2	1			7			6			
4													1					3		2	1			7			5			
5													1					3		2	2			8			4			
6													1					3		3	1			9			4			

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

035_S_2061_2012 elaine

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM									
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract		
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N
1													3					5			6			8			5	
2													3					3			5			4			5	
3													0					2			1			8			5	
4													1					2			3			7			8	
5													2					4		1	2			3			6	
6													1					5			3			3			2	

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

031_S_4005 2012 Clerk

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N	C	A	N
1													2					2			2				1					5
2													1					2			2				1					5
3													3					2			4				0					4
4													1					1			2				0					3
5													2					2			4				4					4
6													1					1			1				1					1

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

123_S_0301_2012

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM									
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract			
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N	C	A	N
1				1														1				7			4					
2				5														1				3			4					
3				5														1							3					
4				6																3										
5				2														1		3										
6				4																		1								

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

123_S_2055_2012

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N	C	A	N
1				8									4			1					1									
2				~~6~~ 7									2								1									
3				10									4																	
4				15									4																	
5				10									5																	
6				13									2												1					

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

038_S_2073_2012

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM									
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract			
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N	C	A	N
1																		1			1									
2																		1			1									
3																	2	1			1									
4																		1		1	1									
5																		1		1	1									
6													1					1			2									

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

072_S_2307_2012

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N	C	A	N
1																	1					2			1			1		
2																	0					0			1			1		
3																	6					1			0			0		
4																	0					0			0			0		
5				2													0					0			0			0		
6				3													0					0			0			0		

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

041_S_4004_2012

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM									
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract		
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N
1	2																	5						6			1	
2	2																	6						6			2	
3	2																	3						3			1	
4	2																	5						2			2	
5	2																	4						1			2	
6	2																	5						1			2	

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

350_S_4114_2012

Table for evaluating the results of the methods tested for identifying White Matter Lesions

WML method	WMLs not identified						WMLs missed partially						WMLs misclassified																	
	Pons	PVWM	BG	DWM			Pons	PVWM	BG	DWM			Pons		PVWM			Basal Ganglia		DWM										
				Anterior WM tract	Central WM tract	Posterior WM tract				Anterior WM tract	Central WM tract	Posterior WM tract								Anterior WM tract			Central WM tract			Posterior WM tract				
													A	N	C	A	N	A	N	C	A	N	C	A	N	C	A	N		
1				1									4					5			3			9			3			
2				1									3					5		2	3			8			3			
3				1									3					6			3			9			3			
4													2					9		1	3			12			6			
5													2					8		3	4			10			4			
6				1									1					6		3	4			12			3			

Note: C = Cortex (gray matter)
A = Artifact
N = Normal WM

PVWM = Periventricular White Matter
BG = Basal Ganglia

* = old stroke lesion

Bibliography

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Baumgartner, C. F., Koch, L. M., Tezcan, K. C., and Ang, J. X. (2018). Visual feature attribution using Wasserstein GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8309–8319.

Bellini, R., Kleiman, Y., and Cohen-Or, D. (2016). Time-varying weathering in texture space. *ACM Transactions on Graphics (TOG)*, 35(4):141.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons.

Bendfeldt, K., Kuster, P., Traud, S., Egger, H., Winklhofer, S., Mueller-Lenke, N., Naegelin, Y., Gass, A., Kappos, L., Matthews, P. M., et al. (2009). Association of regional gray matter volume loss and progression of white matter lesions in multiple sclerosis—a longitudinal voxel-based morphometry study. *NeuroImage*, 45(1):60–67.

Bengio, Y. et al. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1):1–127.

Biesbroek, J. M., Kuijf, H. J., van der Graaf, Y., Vincken, K. L., Postma, A., Mali, W. P. T. M., Biessels, G. J., Geerlings, M. I., and Group, o. b. o. t. S. S. (2013). Association between subcortical vascular lesion location and cognition: A voxel-based and tract-based lesion-symptom mapping study. The SMART-MR study. *PLOS ONE*, 8(4):1–7.

Birdsill, A. C., Koscik, R. L., Jonaitis, E. M., Johnson, S. C., Okonkwo, O. C., Hermann, B. P., LaRue, A., Sager, M. A., and Bendlin, B. B. (2014). Regional white matter hyperintensities: Aging, Alzheimer's disease risk, and cognitive function. *Neurobiology of Aging*, 35(4):769–776.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Bland, J. M. and Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307 – 310. Originally published as Volume 1, Issue 8476.

Brosch, T., Tang, L. Y., Yoo, Y., Li, D. K., Traboulsee, A., and Tam, R. (2016). Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration

applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, 35(5):1229–1239.

Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., and Cherubini, A. (2015). Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review. *Neuroinformatics*, 13(3):261–276.

Callisaya, M. L., Beare, R., Phan, T. G., Blizzard, L., Thrift, A. G., Chen, J., and Srikanth, V. K. (2013). Brain structural change and gait decline: A longitudinal population-based study. *Journal of the American Geriatrics Society*, 61(7):1074–1079.

Castellino, R. A. (2005). Computer aided detection (CAD): An overview. *Cancer Imaging: the Official Publication of the International Cancer Imaging Society*, 5:17—19.

Chappell, F. M., del Carmen Valdés Hernández, M., Makin, S. D., Shuler, K., Sakka, E., Dennis, M. S., Armitage, P. A., Muñoz Maniega, S., and Wardlaw, J. M. (2017). Sample size considerations for trials using cerebral white matter hyperintensity progression as an intermediate outcome at 1 year after mild stroke: Results of a prospective cohort study. *Trials*, 18(1):1–10.

Chen, X. and Konukoglu, E. (2018). Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In *MIDL Conference book*. MIDL.

Cho, A.-H., Kim, H.-R., Kim, W., and Yang, D. W. (2015). White matter hyperintensity in ischemic stroke patients: It may regress over time. *Journal of Stroke*, 17(1):60.

Choi, H. and Jin, K. H. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural Brain Research*, 344:103 – 109.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham. Springer International Publishing.

Clayden, J., Maniega, S., Storkey, A., King, M., Bastin, M., and Clark, C. (2011). TractoR: Magnetic resonance imaging and tractography with R. *Journal of Statistical Software, Articles*, 44(8):1–18.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Criminisi, A. and Shotton, J. (2013). *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media.

Dauphin, Y. N., Vries, H. d., and Bengio, Y. (2015). Equilibrated adaptive learning rates for non-convex optimization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1504–1512, Cambridge, MA, USA. MIT Press.

de Brébisson, A. and Montana, G. (2015). Deep neural networks for anatomical brain segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20–28.

de Leeuw, F.-E., de Groot, J. C., Achten, E., Oudkerk, M., Ramos, L. M. P., Heijboer, R., Hofman, A., Jolles, J., van Gijn, J., and Breteler, M. M. B. (2001). Prevalence of cerebral white matter lesions in elderly people: A population based magnetic resonance imaging study. The Rotterdam Scan Study. *Journal of Neurology, Neurosurgery & Psychiatry*, 70(1):9–14.

Debette, S. and Markus, H. S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ*, 341:c3666.

DeCarli, C., Fletcher, E., Ramey, V., Harvey, D., and Jagust, W. J. (2005). Anatomical mapping of white matter hyperintensities (WMH) exploring the relationships between periventricular WMH, deep WMH, and total WMH burden. *Stroke*, 36(1):50–55.

Dheeba, J., Singh, N. A., and Selvi, S. T. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of Biomedical Informatics*, 49:45–52.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Dickie, D. A., Job, D. E., Gonzalez, D. R., Shenkin, S. D., Ahearn, T. S., Murray, A. D., and Wardlaw, J. M. (2014). Correction: Variance in brain volume with advancing age: Implications for defining the limits of normality. *PloS one*, 9(1):10–1371.

Dickie, D. A., Job, D. E., Gonzalez, D. R., Shenkin, S. D., and Wardlaw, J. M. (2015). Use of brain MRI atlases to determine boundaries of age-related pathology: The importance of statistical method. *PloS one*, 10(5):e0127939.

Dickie, D. A., Ritchie, S. J., Cox, S. R., Sakka, E., Royle, N. A., Aribisala, B. S., Valdés Hernández, M. d. C., Maniega, S. M., Pattie, A., Corley, J., et al. (2016). Vascular risk factors and progression of white matter hyperintensities in the Lothian Birth Cohort 1936. *Neurobiology of Aging*, 42:116 – 123.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211.

Dugas-Phocion, G., Ballester, M. A. G., Malandain, G., Lebrun, C., and Ayache, N. (2004). Improved EM-based tissue segmentation and partial volume effect quantification in multi-sequence brain MRI. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, pages 26–33, Berlin, Heidelberg. Springer Berlin Heidelberg.

Durand-Birchenall, J., Leclercq, C., Daouk, J., Monet, P., Godefroy, O., and Bugnicourt, J.-M. (2012). Attenuation of brain white matter lesions after lacunar stroke. *International Journal of Preventive Medicine*, 3(2):134–138.

Fan, H., Su, H., and Guibas, L. (2017). A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2463–2471.

Fazekas, F., Chawluk, J., Alavi, A., Hurtig, H., and Zimmerman, R. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *AJR. American Journal of Roentgenology*, 149(2):351—356.

Firbank, M., Minett, T., and O'brien, J. (2003). Changes in DWI and MRS associated with white matter hyperintensities in elderly subjects. *Neurology*, 61(7):950–954.

Firmino, M., Angelo, G., Morais, H., Dantas, M. R., and Valentim, R. (2016). Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *Biomedical Engineering Online*, 15(1):2.

Forbes, F., Doyle, S., Garcia-Lorenzo, D., Barillot, C., and Dojat, M. (2010). Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 69–72.

Freifeld, O., Greenspan, H., and Goldberger, J. (2009). Multiple sclerosis lesion detection using constrained GMM and curve evolution. *International Journal of Biomedical Imaging*, 2009:715124.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical Image Analysis*, 17(1):1 – 18.

Ge, Y. (2006). Multiple sclerosis: The role of MR imaging. *American Journal of Neuroradiology*, 27(6):1165–1176.

Ghafoorian, M. (2018). *Machine Learning for Quantification of Small Vessel Disease Imaging Biomarkers*. PhD thesis, Radboud University Nijmegen. [http://hdl.handle.net/2066/183226].

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

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I. W. M., Sanchez, C. I., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., and Platel, B. (2017b). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific reports*, 7(1):5110.

Gibson, E., Gao, F., Black, S. E., and Lobaugh, N. J. (2010). Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T. *Journal of Magnetic Resonance Imaging : JMRI*, 31(6):1311—1322.

Godin, O., Tzourio, C., Maillard, P., Alpérovitch, A., Mazoyer, B., and Dufouil, C. (2009). Apolipoprotein E genotype is related to progression of white matter lesion load. *Stroke*, 40(10):3186–3190.

Godin, O., Tzourio, C., Maillard, P., Mazoyer, B., and Dufouil, C. (2011). Antihypertensive treatment and change in blood pressure are associated with the progression of white matter lesion volumes. *Circulation*, 123(3):266–273.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Gouw, A. A., van der Flier, W. M., Fazekas, F., van Straaten, E. C., Pantoni, L., Poggesi, A., Inzitari, D., Erkinjuntti, T., Wahlund, L. O., Waldemar, G., Schmidt, R., Scheltens, P., and Barkhof, F. (2008a). Progression of white matter hyperintensities and incidence of new lacunes over a 3-year period. *Stroke*, 39(5):1414–1420.

Gouw, A. A., Van Der Flier, W. M., Van Straaten, E. C., Pantoni, L., Bastos-Leite, A. J., Inzitari, D., Erkinjuntti, T., Wahlund, L. O., Ryberg, C., Schmidt, R., Fazekas, F., Scheltens, P., and Barkhof, F. (2008b). Reliability and sensitivity of visual scales versus volumetry for evaluating white matter hyperintensity progression. *Cerebrovascular Diseases*, 25(3):247–253.

Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783.

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

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, Red Hook, NY, USA. Curran Associates Inc.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264.

Gwo, C.-Y., Zhu, D. C., and Zhang, R. (2019). Brain white matter hyperintensity lesion characterization in T2 Fluid-Attenuated Inversion Recovery Magnetic Resonance Images: Shape, Texture, and Potential Growth. *Frontiers in Neuroscience*, 13:353.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18 – 31.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hinton, G. (2010). A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1):926.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116.

Hong, Y., Joshi, S., Sanchez, M., Styner, M., and Niethammer, M. (2012). Metamorphic geodesic regression. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, pages 197–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Huang, C.-C., Yang, A. C., Chou, K.-H., Liu, M.-E., Fang, S.-C., Chen, C.-C., Tsai, S.-J., and Lin, C.-P. (2018). Nonlinear pattern of the emergence of white matter hyperintensity in healthy Han Chinese: An adult lifespan study. *Neurobiology of Aging*, 67:99 – 107.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863.

Ithapu, V., Singh, V., Lindner, C., Austin, B. P., Hinrichs, C., Carlsson, C. M., Bendlin, B. B., and Johnson, S. C. (2014). Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. *Human Brain Mapping*, 35(8):4219–4235.

Jeerakathil, T., Wolf, P. A., Beiser, A., Massaro, J., Seshadri, S., D'agostino, R. B., and DeCarli, C. (2004). Stroke risk profile predicts white matter hyperintensity volume: the Framingham Study. *Stroke*, 35(8):1857–1861.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825 – 841.

Jeong, Y., Rachmadi, M. F., Valdés-Hernández, M. d. C., and Komura, T. (2019). Dilated saliency U-Net for white matter hyperintensities segmentation using Irregularity Age Map. *Frontiers in Aging Neuroscience*, 11:150.

Jiaerken, Y., Luo, X., Yu, X., Huang, P., Xu, X., Zhang, M., and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (2019). Microstructural and metabolic changes in the longitudinal progression of white matter hyperintensities. *Journal of Cerebral Blood Flow & Metabolism*, 39(8):1613–1622. PMID: 29519198.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61 – 78.

Kapeller, P., Barber, R., Vermeulen, R., Ader, H., Scheltens, P., Freidl, W., Almkvist, O., Moretti, M., Del Ser, T., Vaghfeldt, P., et al. (2003). Visual rating of age-related white matter changes on magnetic resonance imaging: Scale comparison, interrater agreement, and correlations with quantitative measurements. *Stroke*, 34(2):441–445.

Kempton, M. J., Geddes, J. R., Ettinger, U., Williams, S. C., and Grasby, P. M. (2008). Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Archives of general psychiatry*, 65(9):1017–1032.

Khademi, A., Venetsanopoulos, A., and Moody, A. R. (2012). Robust white matter lesion segmentation in FLAIR MRI. *IEEE Transactions on Biomedical Engineering*, 59(3):860–871.

Khayati, R., Vafadust, M., Towhidkhah, F., and Nabavi, M. (2008). Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and markov random field model. *Computers in Biology and Medicine*, 38(3):379 – 390.

Kikinis, R., Guttmann, C., Metcalf, D., Wells, W., Ettinger, G., Weiner, H., and Jolesz, F. (1999). Quantitative follow-up of patients with multiple sclerosis using MRI: Technical aspects. *Journal of Magnetic Resonance Imaging: JMRI*, 9(4):519—530.

Kim, K. W., MacFall, J. R., and Payne, M. E. (2008). Classification of white matter lesions on magnetic resonance imaging in elderly persons. *Biological Psychiatry*, 64(4):273–280.

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

Klöppel, S., Abdulkadir, A., Hadjidemetriou, S., Issleib, S., Frings, L., Thanh, T. N., Mader, I., Teipel, S. J., Hüll, M., and Ronneberger, O. (2011). A comparison of different automated methods for the detection of white matter lesions in MRI data. *NeuroImage*, 57(2):416–422.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M. J., Casamitjana, A., Collins, D. L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B., Park, H., Park, S. H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C. H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M. A., and Biessels, G. J. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568.

Lambert, C., Benjamin, P., Zeestraten, E., Lawrence, A. J., Barrick, T. R., and Markus, H. S. (2016). Longitudinal patterns of leukoaraiosis and brain atrophy in symptomatic small vessel disease. *Brain*, 139(4):1136–1151.

Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., Bryan, R. N., and Davatzikos, C. (2008). Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Academic Radiology*, 15(3):300–313.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., Simard, P., and Vapnik, V. (1995). Comparison of learning algorithms for handwritten digit recognition. In *International Conference on Artificial Neural Networks, Paris*, pages 53–60. EC2 & Cie.

Leite, M., Rittner, L., Appenzeller, S., Ruocco, H. H., and Lotufo, R. (2015). Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging. *Journal of Medical Imaging*, 2(1):014002–014002.

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., and Menze, B. (2018). Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage*, 183:650 – 665.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88.

Liu, M., Zhang, D., Yap, P.-T., and Shen, D. (2012). Hierarchical ensemble of multi-level classifiers for diagnosis of Alzheimer's disease. In *International Workshop on Machine Learning in Medical Imaging*, pages 27–35. Springer.

Lockhart, S., Mayda, A. B., Roach, A., Fletcher, E., Carmichael, O., Maillard, P., Schwarz, C., Yonelinas, A., Ranganath, C., and DeCarli, C. (2012). Episodic memory

function is associated with multiple measures of white matter integrity in cognitive aging. *Frontiers in Human Neuroscience*, 6:56.

Longstreth, W., Manolio, T. A., Arnold, A., Burke, G. L., Bryan, N., Jungreis, C. A., Enright, P. L., O'Leary, D., and Fried, L. (1996). Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people. *Stroke*, 27(8):1274–1282.

Longstreth Jr, W. T., Arnold, A. M., Beauchamp Jr, N. J., Manolio, T. A., Lefkowitz, D., Jungreis, C., Hirsch, C. H., O'Leary, D. H., and Furberg, C. D. (2005). Incidence, manifestations, and predictors of worsening white matter on serial cranial magnetic resonance imaging in the elderly: The Cardiovascular Health Study. *Stroke*, 36(1):56–61.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S. Y., AlMazroa, M. A., Alvarado, M., Anderson, H. R., Anderson, L. M., Andrews, K. G., Atkinson, C., Baddour, L. M., Barker-Collo, S., Bartels, D. H., Bell, M. L., Benjamin, E. J., Bennett, D., Bhalla, K., Bikbov, B., Abdulhak, A. B., Birbeck, G., Blyth, F., Bolliger, I., Boufous, S., Bucello, C., Burch, M., Burney, P., Carapetis, J., Chen, H., Chou, D., Chugh, S. S., Coffeng, L. E., Colan, S. D., Colquhoun, S., Colson, K. E., Condon, J., Connor, M. D., Cooper, L. T., Corriere, M., Cortinovis, M., de Vaccaro, K. C., Couser, W., Cowie, B. C., Criqui, M. H., Cross, M., Dabhadkar, K. C., Dahodwala, N., Leo, D. D., Degenhardt, L., Delossantos, A., Denenberg, J., Jarlais, D. C. D., Dharmaratne, S. D., Dorsey, E. R., Driscoll, T., Duber, H., Ebel, B., Erwin, P. J., Espindola, P., Ezzati, M., Feigin, V., Flaxman, A. D., Forouzanfar, M. H., Fowkes, F. G. R., Franklin, R., Fransen, M., Freeman, M. K., Gabriel, S. E., Gakidou, E., Gaspari, F., Gillum, R. F., Gonzalez-Medina, D., Halasa, Y. A., Haring, D., Harrison, J. E., Havmoeller, R., Hay, R. J., Hoen, B., Hotez, P. J., Hoy, D., Jacobsen, K. H., James, S. L., Jasrasaria, R., Jayaraman, S., Johns, N., Karthikeyan, G., Kassebaum, N., Keren, A., Khoo, J.-P., Knowlton, L. M., Kobusingye, O., Koranteng, A., Krishnamurthi, R., Lipnick, M., Lipshultz, S. E., Ohno, S. L., Mabweijano, J., MacIntyre, M. F., Mallinger, L., March, L., Marks, G. B., Marks, R., Matsumori, A., Matzopoulos, R., Mayosi, B. M., McAnulty, J. H., McDermott, M. M., McGrath, J., Memish, Z. A., Mensah, G. A., Merriman, T. R., Michaud, C., Miller, M., Miller, T. R., Mock, C., Mocumbi, A. O., Mokdad, A. A., Moran, A., Mulholland, K., Nair, M. N., Naldi, L., Narayan, K. M. V., Nasseri, K., Norman, P., O'Donnell, M., Omer, S. B., Ortblad, K., Osborne, R., Ozgediz, D., Pahari, B., Pandian, J. D., Rivero, A. P., Padilla, R. P., Perez-Ruiz, F., Perico, N., Phillips, D., Pierce, K., Pope, C. A., Porrini, E., Pourmalek, F., Raju, M., Ranganathan, D., Rehm, J. T., Rein, D. B., Remuzzi, G., Rivara, F. P., Roberts, T., León, F. R. D., Rosenfeld, L. C., Rushton, L., Sacco, R. L., Salomon, J. A., Sampson, U., Sanman, E., Schwebel, D. C., Segui-Gomez, M., Shepard, D. S., Singh, D., Singleton, J., Sliwa, K., Smith, E., Steer, A., Taylor, J. A., Thomas, B., Tleyjeh, I. M., Towbin, J. A., Truelsen, T., Undurraga, E. A., Venketasubramanian, N., Vijayakumar, L., Vos, T., Wagner, G. R., Wang, M., Wang, W., Watt, K., Weinstock, M. A., Weintraub, R., Wilkinson, J. D., Woolf, A. D., Wulf, S., Yeh, P.-H., Yip, P., Zabetian, A., Zheng, Z.-J., Lopez, A. D., and Murray, C. J. (2012). Global and

regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859):2095 – 2128.

Luo, X., Jiaerken, Y., Yu, X., Huang, P., Qiu, T., Jia, Y., Li, K., Xu, X., Shen, Z., Guan, X., Zhou, J., Zhang, M., and Adni, F. T. A. D. N. I. (2017). Associations between APOE genotype and cerebral small-vessel disease: A longitudinal study. *Oncotarget*, 8(27):44477–44489.

Lutkenhoff, E. S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J. D., Owen, A. M., and Monti, M. M. (2014). Optimized brain extraction for pathological brains (optiBET). *PloS one*, 9(12):e115551.

Lyksborg, M., Puonti, O., Agn, M., and Larsen, R. (2015). An ensemble of 2D convolutional neural networks for tumor segmentation. In *Image Analysis*, pages 201–211, Cham. Springer International Publishing.

Lyu, S. and Farid, H. (2003). Detecting hidden messages using higher-order statistics and support vector machines. In Petitcolas, F. A. P., editor, *Information Hiding*, pages 340–354, Berlin, Heidelberg. Springer Berlin Heidelberg.

Maillard, P., Crivello, F., Dufouil, C., Tzourio-Mazoyer, N., Tzourio, C., and Mazoyer, B. (2009). Longitudinal follow-up of individual white matter hyperintensities in a large cohort of elderly. *Neuroradiology*, 51(4):209–220.

Maillard, P., Fletcher, E., Harvey, D., Carmichael, O., Reed, B., Mungas, D., and Decarli, C. (2011). White matter hyperintensity penumbra. *Stroke*, 42(7):1917–1922.

Maillard, P., Fletcher, E., Lockhart, S. N., Roach, A. E., Reed, B., Mungas, D., Decarli, C., and Carmichael, O. T. (2014). White matter hyperintensities and their penumbra lie along a continuum of injury in the aging brain. *Stroke*, 45(6):1721–1726.

Malik, J., Belongie, S., Shi, J., and Leung, T. (1999). Textons, contours and regions: Cue integration in image segmentation. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 918–925 vol.2.

Mendrik, A. M., Vincken, K. L., Kuijf, H. J., Breeuwer, M., Bouvy, W. H., de Bresser, J., Alansary, A., de Bruijne, M., Carass, A., El-Baz, A., and et al. (2015). MRBrainS challenge: Online evaluation framework for brain image segmentation in 3T MRI scans. *Computational Intelligence and Neuroscience*, 2015.

Mínguez, B., Rovira, A., Alonso, J., and Córdoba, J. (2007). Decrease in the volume of white matter lesions with improvement of hepatic encephalopathy. *American Journal of Neuroradiology*, 28(8):1499–1500.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Moeskops, P., de Bresser, J., Kuijf, H. J., Mendrik, A. M., Biessels, G. J., Pluim, J. P., and Išgum, I. (2018). Evaluation of a deep learning approach for the segmentation of

brain tissues and white matter hyperintensities of presumed vascular origin in MRI. *NeuroImage: Clinical*, 17:251–262.

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., and Išgum, I. (2016). Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261.

Moriya, Y., Kozaki, K., Nagai, K., and Toba, K. (2009). Attenuation of brain white matter hyperintensities after cerebral infarction. *American Journal of Neuroradiology*, 30(3):3174.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877.

Myers, J. L., Well, A., and Lorch, R. F. (2010). *Research design and statistical analysis*. Routledge.

Newell, D., Nie, K., Chen, J.-H., Hsu, C.-C., Hon, J. Y., Nalcioglu, O., and Su, M.-Y. (2010). Selection of diagnostic features on breast MRI to differentiate between malignant and benign lesions using computer-aided diagnosis: Differences in lesions presenting as mass and non-mass-like enhancement. *European Radiology*, 20(4):771–781.

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11(1):169–198.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27.

Pantoni, L. (2010). Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *The Lancet Neurology*, 9(7):689–701.

Pasi, M., Van Uden, I. W., Tuladhar, A. M., De Leeuw, F. E., and Pantoni, L. (2016). White matter microstructural damage on diffusion tensor imaging in cerebral small vessel disease: Clinical consequences. *Stroke*, 47(6):1679–1684.

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI Conference on Artificial Intelligence*, New Orleans, United States.

Power C, M., Deal A, J., Sharrett Richey, A., Jack Jr, C. R., Knopman, D., Mosley H, T., and Gottesman F, R. (2015). Smoking and white matter hyperintensity progression: The ARIC-MRI Study. *Neurology*, 84(8):841–848.

Prins, N. D. and Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: An update. *Nature reviews. Neurology*, 11(3):157–65.

Prins, N. D., van Straaten, E. C. W., van Dijk, E. J., Simoni, M., van Schijndel, R. A., Vrooman, H. A., Koudstaal, P. J., Scheltens, P., Breteler, M. M. B., and Barkhof, F. (2004). Measuring progression of cerebral white matter lesions on MRI. *Neurology*, 62(9):1533 LP – 1539.

Promjunyakul, N., Lahna, D., Kaye, J., Dodge, H., Erten-Lyons, D., Rooney, W., and Silbert, L. (2015). Characterizing the white matter hyperintensity penumbra with cerebral blood flow measures. *NeuroImage: Clinical*, 8:224 – 229.

Rachmadi, M. F., del C. Valdés-Hernández, M., and Komura, T. (2018a). Transfer learning for task adaptation of brain lesion assessment and prediction of brain abnormalities progression/regression using irregularity age map in brain MRI. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 85–93, Cham. Springer International Publishing.

Rachmadi, M. F., del C. Valdés-Hernández, M., Makin, S., Wardlaw, J. M., and Komura, T. (2019a). Predicting the evolution of white matter hyperintensities in brain mri using generative adversarial networks and irregularity map. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 146–154, Cham. Springer International Publishing.

Rachmadi, M. F., del C. Valdés-Hernández, M., Li, H., Guerrero, R., Meijboom, R., Wiseman, S., Waldman, A., Zhang, J., Rueckert, D., Wardlaw, J., and Komura, T. (2020). Limited One-time Sampling Irregularity Map (LOTS-IM) for automatic unsupervised assessment of white matter hyperintensities and multiple sclerosis lesions in structural brain magnetic resonance images. *Computerized Medical Imaging and Graphics*, 79:101685.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Agan, M. L. F., Di Perri, C., Komura, T., Initiative, A. D. N., et al. (2018b). Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Computerized Medical Imaging and Graphics*, 66:28–43.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Agan, M. L. F., and Komura, T. (2017a). Deep learning vs. conventional machine learning: Pilot study of WMH segmentation in brain MRI with absence or mild vascular pathology. *Journal of Imaging*, 3(4):66.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Agan, M. L. F., and Komura, T. (2017b). Evaluation of four supervised learning schemes in white matter hyperintensities segmentation in absence or mild presence of vascular pathology. In *Medical Image Understanding and Analysis*, pages 482–493, Cham. Springer International Publishing.

Rachmadi, M. F., Valdés-Hernández, M. d. C., and Komura, T. (2017c). Voxel-based irregularity age map (IAM) for brain's white matter hyperintensities in MRI. In *2017*

International Conference on Advanced Computer Science and Information Systems (ICACSIS), pages 321–326. IEEE.

Rachmadi, M. F., Valdés-Hernández, M. d. C., and Komura, T. (2018c). Automatic irregular texture detection in brain MRI without human supervision. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 506–513, Cham. Springer International Publishing.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Makin, S., Wardlaw, J., and Komura, T. (2019b). Automatic spatial estimation of white matter hyperintensities evolution in brain mri using disease evolution predictor deep neural networks. *bioRxiv*.

Ramirez, J., McNeely, A. A., Berezuk, C., Gao, F., and Black, S. E. (2016). Dynamic progression of white matter hyperintensities in alzheimer's disease and normal aging: Results from the sunnybrook dementia study. *Frontiers in Aging Neuroscience*, 8:62.

Reginold, W., Sam, K., Poublanc, J., Fisher, J., Crawley, A., and Mikulis, D. J. (2018). Impact of white matter hyperintensities on surrounding white matter tracts. *Neuroradiology*, 60(9):933–944.

Reijmer, Y. D., Brundel, M., de Bresser, J., Kappelle, L. J., Leemans, A., Biessels, G. J., and (2013). Microstructural white matter abnormalities and cognitive functioning in type 2 diabetes. *Diabetes Care*, 36(1):137–144.

Rekik, I., Allassonnière, S., Carpenter, T. K., and Wardlaw, J. M. (2014). Using longitudinal metamorphosis to examine ischemic stroke lesion dynamics on perfusion-weighted images and in relation to final outcome on T2-w images. *NeuroImage: Clinical*, 5:332–340.

Riedmiller, M. and Braun, H. (1992). Rprop: A fast adaptive learning algorithm. In *Proc. of the Int. Symposium on Computer and Information Science VII*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Rovira Cañellas, A., Mínguez, B., Aymerich, F. X., Jacas, C., Huerga, E., Córdoba, J., and Alonso, J. (2007). Decreased white matter lesion volume and improved cognitive function after liver transplantation. *Hepatology*, 46(5):1485–1490.

Roy, P. K., Bhuiyan, A., Janke, A., Desmond, P. M., Wong, T. Y., Abhayaratna, W. P., Storey, E., and Ramamohanarao, K. (2015). Automatic white matter lesion segmentation using contrast enhanced FLAIR intensity and Markov Random Field. *Computerized Medical Imaging and Graphics*, 45:102–111.

Sachdev, P., Wen, W., Chen, X., and Brodaty, H. (2007). Progression of white matter hyperintensities in elderly individuals over 3 years. *Neurology*, 68(3):214–222.

Salakhutdinov, R. and Hinton, G. (2009). Deep boltzmann machines. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

Salakhutdinov, R. and Larochelle, H. (2010). Efficient learning of deep Boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 693–700.

Scheltens, P., Barkhof, F., Leys, D., Pruvo, J., Nauta, J., Vermersch, P., Steinling, M., and Valk, J. (1993). A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. *Journal of the Neurological Sciences*, 114(1):7 – 12.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., and Shen, D., editors, *Information Processing in Medical Imaging*, pages 146–157, Cham. Springer International Publishing.

Schmidt, H., Zeginigg, M., Wiltgen, M., Freudenberger, P., Petrovic, K., Cavalieri, M., Gider, P., Enzinger, C., Fornage, M., Debette, S., Rotter, J. I., Ikram, M. A., Launer, L. J., and Schmidt, R. (2011). Genetic variants of the NOTCH3 gene in the elderly and magnetic resonance imaging correlates of age-related cerebral small vessel disease. *Brain*, 134(11):3384–3397.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V. J., Zimmer, C., Hemmer, B., and Mühlau, M. (2012a). An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, 59(4):3774 – 3783.

Schmidt, R., Berghold, A., Jokinen, H., Gouw, A. A., van der Flier, W. M., Barkhof, F., Scheltens, P., Petrovic, K., Madureira, S., Verdelho, A., Ferro, J. M., Waldemar, G., Wallin, A., Wahlund, L.-O., Poggesi, A., Pantoni, L., Inzitari, D., Fazekas, F., and Erkinjuntti, T. (2012b). White matter lesion progression in LADIS. *Stroke*, 43(10):2643–2647.

Schmidt, R., Enzinger, C., Ropele, S., Schmidt, H., and Fazekas, F. (2003). Progression of cerebral white matter lesions: 6-Year results of the Austrian Stroke Prevention Study. *Lancet*, 361(9374):2046–2048.

Schmidt, R., Fazekas, F., Enzinger, C., Ropele, S., Kapeller, P., and Schmidt, H. (2002a). Risk factors and progression of small vessel disease-related cerebral abnormalities. In *Ageing and Dementia Current and Future Concepts*, pages 47–52, Vienna. Springer Vienna.

Schmidt, R., Fazekas, F., Kapeller, P., Schmidt, H., and Hartung, H.-P. (1999). MRI white matter hyperintensities: Three-year follow-up of the Austrian Stroke Prevention Study. *Neurology*, 53(1):132–132.

Schmidt, R., Ropele, S., Enzinger, C., Petrovic, K., Smith, S., Schmidt, H., Matthews, P. M., and Fazekas, F. (2005). White matter lesion progression, brain atrophy, and cognitive decline: The Austrian stroke prevention study. *Annals of Neurology*, 58(4):610–616.

Schmidt, R., Scheltens, P., Erkinjuntti, T., Pantoni, L., Markus, H. S., Wallin, A., Barkhof, F., Fazekas, F., Pantoni, L., Schmidt, R., Barkhof, F., Scheltens, P., Erkinjuntti, T., Fazekas, F., and Wallin, A. (2004). White matter lesion progression. *Neurology*, 63(1):139 LP – 144.

Schmidt, R., Schmidt, H., Fazekas, F., Kapeller, P., Roob, G., Lechner, A., Kostner, G. M., and Hartung, H. P. (2000). MRI cerebral white matter lesions and paraoxonase PON1 polymorphisms: Three-year follow-up of the Austrian stroke prevention study. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 20(7):1811–1816.

Schmidt, R., Schmidt, H., Kapeller, P., Lechner, A., and Fazekas, F. (2002b). Evolution of white matter lesions. *Cerebrovascular Diseases (Basel, Switzerland)*, 13 Suppl 2:16—20.

Schmidt, R., Seiler, S., and Loitfelder, M. (2016). Longitudinal change of small-vessel disease-related brain abnormalities. *Journal of Cerebral Blood Flow and Metabolism*, 36(1):26–39.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., and Arbel, T. (2011). Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis*, 15(2):267–282.

Shi, Y. and Wardlaw, J. M. (2016). Update on cerebral small vessel disease: A dynamic whole-brain disease. *Stroke and Vascular Neurology*, 1(3):83–92.

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D. S., Calabresi, P. A., and Pham, D. L. (2010). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524 – 1535.

Shiraishi, J., Li, Q., Appelbaum, D., and Doi, K. (2011). Computer-aided diagnosis and artificial intelligence in clinical imaging. *Seminars in Nuclear Medicine*, 41(6):449 – 462. Image Perception in Nuclear Medicine.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, E. E., Egorova, S., Blacker, D., Killiany, R. J., Muzikansky, A., Dickerson, B. C., Tanzi, R. E., Albert, M. S., Greenberg, S. M., and Guttmann, C. R. G. (2008). Magnetic resonance imaging white matter hyperintensities and brain volume in the prediction of mild cognitive impairment and dementia. *Archives of Neurology*, 65(1):94–100.

Spasov, S., Passamonti, L., Duggento, A., Liò, P., and Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. *NeuroImage*, 189:276 – 287.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Steenwijk, M. D., Pouwels, P. J., Daams, M., van Dalen, J. W., Caan, M. W., Richard, E., Barkhof, F., and Vrenken, H. (2013). Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage: Clinical*, 3:462 – 469.

Stollenga, M. F., Byeon, W., Liwicki, M., and Schmidhuber, J. (2015). Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2998–3006, Cambridge, MA, USA. MIT Press.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28(3) of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA. PMLR.

Suzuki, K. (2012). A review of computer-aided diagnosis in thoracic and colonic imaging. *Quantitative Imaging in Medicine and Surgery*, 2(3):163—176.

Theodoridou, A. and Settas, L. (2006). Demyelination in rheumatic diseases. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(3):290–295.

Tin Kam Ho (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Valdés Hernández, M., Piper, R., Bastin, M., Royle, N., Maniega, S. M., Aribisala, B., Murray, C., Deary, I., and Wardlaw, J. (2014). Morphologic, distributional, volumetric, and intensity characterization of periventricular hyperintensities. *American Journal of Neuroradiology*, 35(1):55–62.

Valdés Hernández, M. d. C., Ferguson, K. J., Chappell, F. M., and Wardlaw, J. M. (2010). New multispectral MRI data fusion technique for white matter lesion segmentation: method and comparison with thresholding in FLAIR images. *European Radiology*, 20(7):1684–1691.

Valdés Hernández, M. d. C., González-Castro, V., Ghandour, D. T., Wang, X., Doubal, F., Maniega, S. M., Armitage, P. A., and Wardlaw, J. M. (2016). On the computational assessment of white matter hyperintensity progression: Difficulties in method selection and bias field correction performance on images with significant white matter pathology. *Neuroradiology*, 58(5):475–485.

Valdés Hernández, M. D. C., Armitage, P. A., Thrippleton, M. J., Chappell, F., Sandeman, E., Muñoz Maniega, S., Shuler, K., and Wardlaw, J. M. (2015a). Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. *Brain and Behavior*, 5(12):e00415.

Valdés Hernández, M. d. C., Maconick, L. C., Muñoz Maniega, S., Wang, X., Wiseman, S., Armitage, P. A., Doubal, F. N., Makin, S., Sudlow, C. L. M., Dennis, M. S., Deary, I. J., Bastin, M., and Wardlaw, J. M. (2015b). A comparison of location of acute symptomatic vs. 'silent' small vessel lesions. *International Journal of Stroke: Official Journal of the International Stroke Society*, 10(7):1044—1050.

Valdés Hernández, M. d. C., Morris, Z., Dickie, D. A., Royle, N. A., Muñoz Maniega, S., Aribisala, B. S., Bastin, M. E., Deary, I. J., and Wardlaw, J. M. (2013). Close correlation between quantitative and qualitative assessments of white matter lesions. *Neuroepidemiology*, 40(1):13—22.

Valdés Hernández, M. D. C., Qiu, X., Wang, X., Wiseman, S., Sakka, E., Maconick, L. C., Doubal, F., Sudlow, C. L. M., and Wardlaw, J. M. (2017). Interhemispheric characterization of small vessel disease imaging markers after subcortical infarct. *Brain and Behavior*, 7(1):e00595.

van der Holst, H. M., van Uden, I. W. M., Tuladhar, A. M., de Laat, K. F., van Norden, A. G. W., Norris, D. G., van Dijk, E. J., Rutten-Jacobs, L. C., and de Leeuw, F.-E. (2016). Factors associated with 8-year mortality in older patients with cerebral small vessel disease: the Radboud University Nijmegen Diffusion Tensor and Magnetic Resonance Cohort (RUN DMC) Study. *JAMA Neurology*, 73(4):402–409.

Van Dijk, E. J., Prins, N. D., Vrooman, H. A., Hofman, A., Koudstaal, P. J., and Breteler, M. M. B. (2008). Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences: Rotterdam scan study. *Stroke*, 39(10):2712–2719.

van Leijsen, E. M., Bergkamp, M. I., van Uden, I. W., Cooijmans, S., Ghafoorian, M., van der Holst, H. M., Norris, D. G., Kessels, R. P., Platel, B., Tuladhar, A. M., and de Leeuw, F. E. (2019). Cognitive consequences of regression of cerebral small vessel disease. *European Stroke Journal*, 4(1):85–89.

van Leijsen, E. M., Bergkamp, M. I., van Uden, I. W., Ghafoorian, M., van der Holst, H. M., Norris, D. G., Platel, B., Tuladhar, A. M., and de Leeuw, F.-E. (2018). Progression of white matter hyperintensities preceded by heterogeneous decline of microstructural integrity. *Stroke*, 49(6):1386–1393.

van Leijsen, E. M., de Leeuw, F.-E., and Tuladhar, A. M. (2017). Disease progression and regression in sporadic small vessel disease–insights from neuroimaging. *Clinical Science*, 131(12):1191–1206.

Van Leijsen, E. M., Van Uden, I. W., Ghafoorian, M., Bergkamp, M. I., Lohner, V., Kooijmans, E. C., Van Der Holst, H. M., Tuladhar, A. M., Norris, D. G., Van DIjk, E. J., Rutten-Jacobs, L. C., Platel, B., Klijn, C. J., and De Leeuw, F. E. (2017). Nonlinear temporal dynamics of cerebral small vessel disease. *Neurology*, 89(15):1569–1577.

Van Nguyen, H., Zhou, K., and Vemulapalli, R. (2015). Cross-domain synthesis of medical images using efficient location-sensitive deep network. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 677–684, Cham. Springer International Publishing.

van Straaten, E. C. W., Harvey, D., Scheltens, P., Barkhof, F., Petersen, R. C., Thal, L. J., Jack, C. R., and DeCarli, C. (2008). Periventricular white matter hyperintensities increase the likelihood of progression from amnestic mild cognitive impairment to dementia. *Journal of Neurology*, 255(9):1302.

Veldink, J. H., Scheltens, P., Jonker, C., and Launer, L. J. (1998). Progression of cerebral white matter hyperintensities on MRI is related to diastolic blood pressure. *Neurology*, 51(1):319–320.

Verhaaren, B. F. J., Vernooij, M. W., De Boer, R., Hofman, A., Niessen, W. J., Van Der Lugt, A., Ikram, M. A., F.J., V. B., W., V. M., Renske, d. B., Albert, H., J., N. W., Aad, v. d. L., and Arfan, I. M. (2013). High blood pressure and cerebral white matter lesion progression in the general population. *Hypertension*, 61(6):1354–1359.

Videbech, P. (1997). MRI findings in patients with affective disorder: A meta-analysis. *Acta Psychiatrica Scandinavica*, 96(3):157–168.

Wahlund, L. O., Barkhof, F., Fazekas, F., Bronge, L., Augustin, M., Sjögren, M., Wallin, A., Ader, H., Leys, D., Pantoni, L., Pasquier, F., Erkinjuntti, T., Scheltens, P., and European Task Force on Age-Related White Matter Changes (2001). A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke*, 32(6):1318–22.

Wardlaw, J. M., Chappell, F. M., Valdés Hernández, M. D. C., Makin, S. D., Staals, J., Shuler, K., Thrippleton, M. J., Armitage, P. A., Munõz-Maniega, S., Heye, A. K., Sakka, E., and Dennis, M. S. (2017). White matter hyperintensity reduction and outcomes after minor stroke. *Neurology*, 89(10):1003–1010.

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R. I., O'Brien, J. T., Barkhof, F., Benavente, O. R., Black, S. E., Brayne, C., Breteler, M., Chabriat, H., Decarli, C., de Leeuw, F.-E., Doubal, F., Duering, M., Fox, N. C., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., Oostenbrugge, R. v., Pantoni, L., Speck, O., Stephan, B. C. M., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P. B., Dichgans, M., and STandards for ReportIng Vascular changes on nEuroimaging (STRIVE v1) (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet. Neurology*, 12(8):822–38.

Wardlaw, J. M., Valdés Hernández, M. C., and Muñoz-Maniega, S. (2015). What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. *Journal of the American Heart Association*, 4(6):e001140.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., Morris, J. C., Petersen, R. C., Saykin, A. J., Schmidt, M. E., Shaw, L., Shen, L., Siuciak, J. A., Soares, H., Toga, A. W., and Trojanowski, J. Q. (2013). The alzheimer's disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111 – e194.

Xia, T., Chartsias, A., and Tsaftaris, S. A. (2019). Adversarial pseudo healthy synthesis needs pathology factorization. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 512–526, London, United Kingdom. PMLR.

Yamada, K., Sakai, K., Owada, K., Mineura, K., and Nishimura, T. (2010). Cerebral white matter lesions may be partially reversible in patients with carotid artery stenosis. *American Journal of Neuroradiology*, 31(7):1350–1352.

Yoshita, M., Fletcher, E., Harvey, D., Ortega, M., Martinez, O., Mungas, D. M., Reed, B. R., and DeCarli, C. S. (2006). Extent and distribution of white matter hyperintensities in normal aging, MCI, and AD. *Neurology*, 67(12):2192–2198.

Yu, R., Xiao, L., Wei, Z., and Fei, X. (2015). Automatic Segmentation of White Matter Lesions Using SVM and RSF Model in Multi-channel MRI. In *Image and Graphics*, pages 654–663, Cham. Springer International Publishing.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57.

Zhang, Y., Schuff, N., Camacho, M., Chao, L. L., Fletcher, T. P., Yaffe, K., Woolley, S. C., Madison, C., Rosen, H. J., Miller, B. L., and Weiner, M. W. (2013). MRI markers for mild cognitive impairment: comparisons between white matter integrity and gray matter volume measurements. *PloS one*, 8(6):e66367.

Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.