Automatic Spatial Estimation of White Matter Hyperintensities Evolution in Brain MRI using Disease Evolution Predictor Deep Neural Networks

Muhammad Febrian Rachmadi[a,b], Maria del C. Valdés-Hernández[b], Stephen Makin[b], Joanna Wardlaw[b], Taku Komura[a]

[a] *School of Informatics, University of Edinburgh, Edinburgh, UK*
[b] *Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK*

Abstract

Previous studies have indicated that white matter hyperintensities (WMH) may evolve, i.e., shrink, grow, or stay stable, over a period of time. However, predicting the evolution of WMH is challenging because the rate and direction of WMH evolution varies greatly across studies. Evolution of WMH also has a non-deterministic nature because some clinical factors that possibly influence it are still not known. In this study, we attempt to predict the evolution of WMH from baseline to follow-up (i.e., 1-year later) using deep learning. We name this proposed model “Disease Evolution Predictor” (DEP). The DEP model receives a baseline image as input and produces a map called “Disease Evolution Map” (DEM), which represents the evolution of WMH from baseline to follow-up. Two models of DEP are proposed, i.e., DEP-UResNet and DEP-GAN, which represent supervised and unsupervised deep learning algorithms respectively. To simulate the non-deterministic and unknown parameters involved in WMH evolution, we propose modulating a Gaussian noise array to the DEP model as auxiliary input. This forces the DEP model to imitate a wider spectrum of alternatives in the results. The alternatives of using other types of auxiliary input instead, such as baseline WMH and stroke lesion loads were also tested. Based on our experiments, the supervised DEP-UResNet regularly performed better than the DEP-GAN. However, DEP-GAN using probability map (PM) yielded similar performances to the DEP-UResNet and performed best in clinical analysis. Furthermore, ablation study showed that auxiliary input, especially the Gaussian noise, improved the performance of DEP models regardless the model’s architecture. To the best of our knowledge, this is the first extensive study on modelling WMH evolution using deep learning algorithms and dealing with the non-deterministic nature of WMH evolution.

Keywords: white matter hyperintensities (WMH), WMH evolution, disease evolution predictor (DEP), DEP Generative Adversarial Network (DEP-GAN), DEP U-Residual Network (DEP-UResNet).

1. Introduction

White matter hyperintensities (WMH) are neuroradiological features seen in T2-weighted and T2-fluid attenuated inversion recovery (T2-FLAIR) brain magnetic resonance images (MRI). Clinically, WMH have been commonly associated with stroke, ageing, and dementia progression (Wardlaw et al., 2013; Prins and Scheltens, 2015). Furthermore, recent studies have shown that WMH may decrease (i.e., shrink/regress), stay unchanged (i.e., stable), or increase (i.e., grow/progress) over a period of time (Ramirez et al., 2016; Chappell et al., 2017; Wardlaw et al., 2017). In this study, we refer to theses changes as “evolution of WMH”.

Predicting the evolution of WMH is challenging because the rate and direction of WMH evolution varies considerably across studies (Schmidt et al., 2016; van Leijsen et al., 2017a,b) and several risk factors, either commonly or not fully known, could be involved in their progression (Wardlaw et al., 2017). For example, some risk factors and predictors that have been commonly associated with WMH progression are baseline WMH volume (Schmidt et al., 2003; Sachdev et al., 2007;

van Dijk et al., 2008; Wardlaw et al., 2017; Chappell et al., 2017), blood pressure or hypertension (Veldink et al., 1998; Schmidt et al., 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., 2008; Wardlaw et al., 2017), and genetic properties (Schmidt et al., 2002a, 2011; Godin et al., 2009; Luo et al., 2017). Surrounding regions of WMH that may appear like normal appearing white matter (NAWM) with less structural integrity, usually called the "penumbra of WMH" (Maillard et al., 2011), have also been reported as having a high risk of becoming WMH over time (Maillard et al., 2014; Pasi et al., 2016). On the other hand, regression of WMH volume has been reported in several radiological observations on MRI, such as after cerebral infraction (Moriya et al., 2009), strokes (Durand-Birchenall et al., 2012; Cho et al., 2015; Wardlaw et al., 2017), improved hepatic encephalopathy (Mínguez et al., 2007), lower blood pressure (Wardlaw et al., 2017), liver transplantation (Rovira Cañellas et al., 2007), and carotid artery stenting (Yamada et al., 2010). While a recent study suggested that areas of shrinking WMH were actually still damaged (Jiaerken et al., 2018), a more recent study showed that WMH regression did not accompany brain atrophy and suggested that WMH regression follows a relatively benign clinical course (van Leijsen et al., 2019).

In this study, we propose 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 called "Disease Evolution Predictor" (DEP) model (discussed in Section 2.2). 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 2.1). For this study we have chosen deep neural networks due to their exceptional performance on WMH segmentation (Rachmadi et al., 2017; Li et al., 2018; Kuijf et al., 2019). We use a Generative Adversarial Network (GAN) (Goodfellow et al., 2014) and the U-Residual Network (UResNet) (Guerrero et al., 2018) 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 we are interested in predicting the evolution of specific neuroradiological MRI features (i.e., WMH in T2-FLAIR), 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 mild cognitive impairment to Alzheimer's disease (Spasov et al., 2019) and progression of cognitive decline in Alzheimer's disease patients (Choi et al., 2018). Instead, our 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 (Rachmadi et al., 2019a). Results and evaluations can be seen in Section 4.

This study is an extension of our previous work (Rachmadi et al., 2019a) in MICCAI 2019. The main contributions of this study, not addressed in our previous work are as follows.

1. We propose and evaluate the use of three different modalities for the DEM: 1) irregularity map (IM) (Rachmadi et al., 2019a), 2) probability map (PM), and 3) binary WMH label (LBL).
2. We performed an ablation study of using different GAN architectures for DEP-GAN model, namely 1) Wasserstein GAN with gradient penalty (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).
3. We propose three different end-to-end DEP learning approaches: 1) supervised learning using DEP-UResNet, 2) unsupervised learning using DEP-GAN and IM (Rachmadi et al., 2019a), and 3) indirectly supervised learning using DEP-GAN and PM.
4. We performed an ablation study of four different types of auxiliary input for DEP model: 1) no auxiliary input, 2) baseline WMH load, 3) baseline WMH and stroke lesions (SL) loads, and 4) Gaussian noise.
5. We performed clinical plausibility analysis of the WMH volumetric changes predicted by the DEP models and risk factors of WMH evolution using analysis of covariance (ANCOVA).

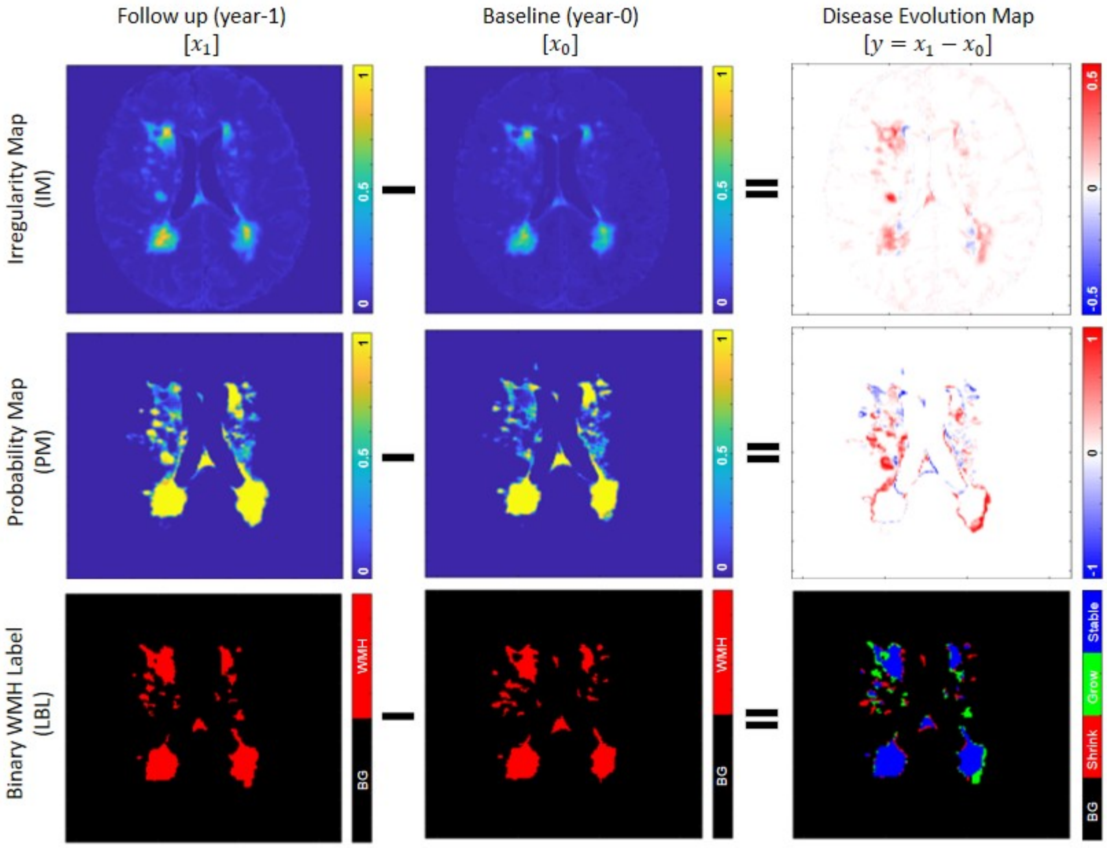


Figure 1: "Disease evolution map" (DEM) (**right**) is produced by subtracting baseline images (**middle**) from follow-up image (**left**). In DEM produced by irregularity map (IM) (**first row**) and probability map (PM) (**second row**), bright yellow pixels represent positive values (i.e., progression) while dark blue pixels represent negative values (i.e., regression). On the other hand, DEM produced by binary WMH label (LBL) (**third row**) has three foreground labels which represent progression or "Grow" (green), regression or "Shrink" (red), and "Stable" (blue). We named this special DEM as three-class DEM label (LBL-DEM).

2. Proposed Methods

2.1. Disease Evolution Map (DEM)

To produce a standard representation of WMH evolution, a simple subtraction operation between two irregularity maps from two time points (i.e., baseline assessment from follow-up assessment) named "Disease Evolution Map" (DEM) was proposed in our previous work (Rachmadi et al., 2019a). In the present study, we evaluate the use of three different modalities in the subtraction operation: irregularity map (i.e. as per (Rachmadi et al., 2019a)), probability map, and binary WMH label.

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 (Rachmadi et al., 2018b). 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. IM is also independent from a human rater or training data, as it is produced using an unsupervised method (i.e., LOTS-IM) (Rachmadi et al., 2019b). Furthermore, previous studies have shown that IM can also be used for WMH segmentation (Rachmadi et al., 2018b), data augmentation of supervised WMH segmentation (Jeong et al., 2019), and simulation of WMH progression and regression (Rachmadi et al., 2018c). DEM resulted from the subtraction of two IMs has values ranging from -1 to 1 (first row of Figure 1). Note how both regression and progression (i.e. blue for negative values and red for positive values) 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 centres of (big) WMH clusters mostly 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 1).

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., 2015). 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. We refer this DEM as three-class DEM label (LBL-DEM), and its depiction can be seen in the bottom-right of Figure 1.

2.2. *Disease Evolution Predictor (DEP) Model using Deep Neural Networks*

In this study, two learning approaches of Disease Evolution Predictor (DEP) model are proposed and evaluated: 1) non-supervised DEP model based on generative adversarial networks (DEP-GAN) (Rachmadi et al., 2019a) and 2) supervised DEP model based on UResNet (DEP-UResNet). DEP-GAN uses either IM or PM to represent the WMH while DEP-UResNet uses T2-FLAIR and three-class DEM label (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 3.2). Finally, DEP-UResNet is categorised as supervised learning as it simply learns DEM labels from LBL-DEM.

2.2.1. *DEP Generative Adversarial Network (DEP-GAN)*

DEP Generative Adversarial Network (DEP-GAN) (Rachmadi et al., 2019a) is based on a 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 visual attribution GAN (VA-GAN), originally proposed to detect atrophy in T2-weighted MRI of Alzheimer's disease (Baumgartner et al., 2017). DEP-GAN consists of a generator based on a U-Residual Network (UResNet) (Guerrero et al., 2018) and two separate convolutional networks used as discriminators (hereinafter will be referred as critics). The schematic of DEP-GAN can be seen in Figure 2.

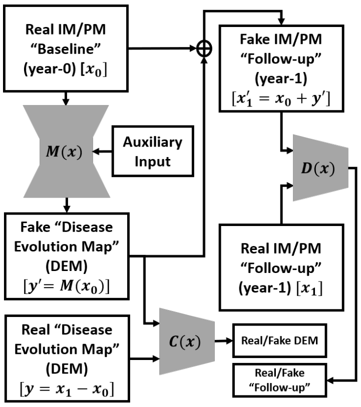

Real IM/PM "Baseline" (year-0) $[x_0]$
Fake IM/PM "Follow-up" (year-1) $[x'_1 = x_0 + y']$
$M(x)$
Auxiliary Input
$D(x)$
Fake "Disease Evolution Map" (DEM) $[y' = M(x_0)]$
Real IM/PM "Follow-up" (year-1) $[x_1]$
Real "Disease Evolution Map" (DEM) $[y = x_1 - x_0]$
$C(x)$
Real/Fake DEM
Real/Fake "Follow-up"

Figure 2: Schematic of the proposed DEP-GAN with 2 discriminators (critics). DEP-GAN can take either irregularity map (IM) or probability map (PM) as input. DEP-GAN also has an auxiliary input to deal with the non-deterministic factors in WMH evolution (see Section 2.3 for full explanation).

Let x_0 be the baseline (year-0) image and x_1 be the follow-up (year-1) image. Then, the "real" DEM (y) can be produced by a simple subtraction ($y = x_1 - x_0$). To generate the "fake" DEM (y'), i.e. without x_1, a generator function ($M(x)$) is used: $y' = M(x_0)$. Thus, a "fake" follow-up image (x'_1) can be produced by $x'_1 = x_0 + y'$. Once $M(x)$ is well/fully trained, the "fake" follow-up (x'_1) and the "real" follow-up (x_1) should be indistinguishable by a critic function $D(x)$, while "fake" DEM (y') and "real" DEM (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 3.

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

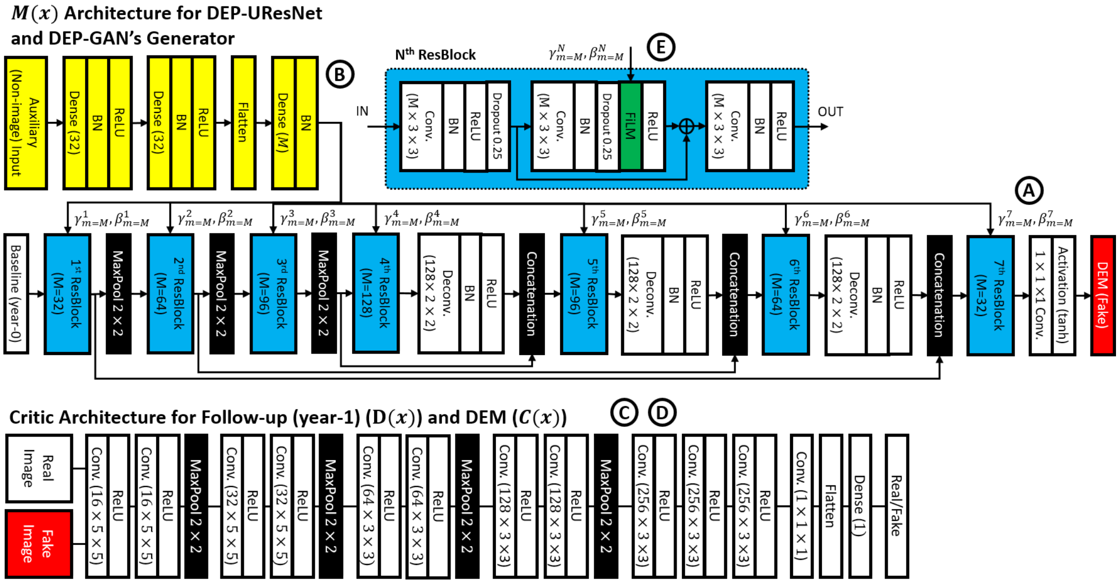


Figure 3: Architecture of DEP-GAN, which consists of one generator (**upper side**, "A") and two critics (**lower side**, "C" and "D"). Note how the proposed auxiliary input is feed-forwarded to convolutional layers (yellow, "B") and then modulated to the generator using FiLM layer (green) inside residual block (ResBlock) (light blue, "E"). Please see Section 2.3 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*).

baseline IM/PM (x_0) is feed-forwarded to this generator to generate a "fake" DEM (y'). There is also an auxiliary input modulated into the generator using a FiLM layer (Perez et al., 2018) inside the residual block (ResBlock) to deal with non-deterministic factors of WMH evolution. This auxiliary input and its modulation will be fully discussed in Section 2.3. The architecture of the DEP-GAN's generator is depicted in the upper side of Figure 3 (with "A", "B", and "E" annotations for UResNet-based generator of $M(x)$, auxiliary input, and residual block (ResBlock) respectively).

Unlike VA-GAN that uses only one critic (i.e., only $D(x)$) (Baumgartner et al., 2017), 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., 2017) and encode realistic plausibility in the modifier (i.e., DEM) (Rachmadi et al., 2019a). 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, we argue that an anatomically realistic DEM is 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 3 (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 Wasserstein GAN (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

$$\begin{aligned} M^* = \arg\min_M \max_{D\in\mathcal{D}} \mathcal{L}_{critic}(M, D) + \\ \arg\min_M \max_{C\in\mathcal{C}} \mathcal{L}_{critic}(M, C) + \mathcal{L}_{reg}(M) \end{aligned} \tag{1}$$

where

$$\begin{aligned} \mathcal{L}_{critic}(M, D) = \mathbb{E}_{x_1\sim\mathbb{P}_1}[D(x_1)] - \\ \mathbb{E}_{x_0\sim\mathbb{P}_0}[D(x_0 + M(x_0))], \end{aligned} \tag{2}$$

$$\begin{aligned} \mathcal{L}_{critic}(M, C) = \mathbb{E}_{x_0,x_1\sim\mathbb{P}_0,\mathbb{P}_1}[C(x_1 - x_0)] - \\ \mathbb{E}_{x_0\sim\mathbb{P}_0}[C(M(x_0))], \end{aligned} \tag{3}$$

$$\begin{aligned} \mathcal{L}_{reg}(M) = [\lambda_1 \left\|x_1' - x_1\right\|_1 + \\ \lambda_2(1 - DSC(x_1', x_1)) + \\ \lambda_3 \left\|vol(x_1') - vol(x_1)\right\|_2], \end{aligned} \tag{4}$$

x_0 is the baseline image that has an underlying distribution $\mathbb{P}_0$, x_1 is the follow-up image that has an underlying distribution $\mathbb{P}_1$, $M(x_0)$ represents the "fake" DEM, $x'_1 = x_0 + M(x_0)$ is the "fake" follow-up image, $\mathcal{D}$ and $\mathcal{C}$ are the critics (i.e. a set of 1-Lipschitz functions (Baumgartner et al., 2017; Gulrajani et al., 2017)), and $\|\cdot\|_1$ and $\|\cdot\|_2$ are the L1 and L2 norms 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., 2017; Gulrajani et al., 2017).

In summary, to optimise the generator ($M(x)$), we need to optimise Equation 1, which optimises both critics ($D(x)$ and $C(x)$) using Equations 2 and 3 respectively based on WGAN-GP's optimisation process (Gulrajani et al., 2017), and use the regularisation function described in Equation 4. Each term in the Equation 4 simply says:

1. Intensities of "fake" follow-up images (x'_1) have to be similar to the "real" follow-up images (x_1) based on L1 norm.
2. The WMH segmentation estimated from x'_1 has to be spatially similar to the WMH segmentation estimated from x_1 based on the Dice similarity coefficient (DSC) (see Equation 6).
3. The WMH volume (in ml) estimated from x'_1 has to be similar to the WMH volume estimated from x_1 based on L2 norm.

The WMH segmentation of x'_1 and x_1 is estimated by either thresholding IM values (i.e., irregularity values) to be above 0.178 (Rachmadi et al., 2019b) or PM values (i.e., probability values) to be above 0.5. Furthermore, each term in Equation 4 is weighted by λ_1, λ_2, and λ_3 which equals to 100 (Baumgartner et al., 2017), 1 and 100 respectively.

2.2.2. DEP U-Residual Network (DEP-UResNet)

In the case of WMH binary labels (LBL) for both time points (i.e., baseline and follow-up in longitudinal data set) are available, a simple supervised deep neural network method can be used to automatically estimate WMH evolution. As previously described in Section 2.1, DEM produced from LBL (i.e., three-class DEM label (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 1.

In this case, the DEP-GAN's generator is detached from the critics and modified into DEP U-Residual Network (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, whcih can be seen in Figure 3 (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. Furthermore, the auxiliary input proposed in this study can be also applied to the DEP-UResNet.

2.3. 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 our previous work, we proposed an auxiliary input module which modulates random noises from normal (Gaussian) distribution to every layer of the DEP-GAN's generator to simulate the unknown/missing factors (i.e., non-image features) involved in WMH evolution and the non-deterministic property of WMH evolution (Rachmadi et al., 2019a). To modulate the auxiliary input to every layer of the DEP-GAN's generator we used Feature-wise Linear Modulation (FiLM) layer (Perez et al., 2018). The FiLM layer is depicted as the green block inside the residual block (ResBlock) in Figure 3 (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

$$FiLM(F_m|\gamma_m, \beta_m) = \gamma_m F_m + \beta_m. \quad (5)$$

where γ_m and β_m for each ResBlock in each layer are automatically determined by convolutional layers (depicted as yellow blocks in Figure 3 with "B" annotation). Note that the proposed auxiliary input module can be easily applied to any deep neural network model. Thus, we applied the auxiliary input module to the two DEP models proposed in the present study: DEP-GAN and DEP-UResNet.

We perform an ablation study of auxiliary input modalities for DEP model by using: 1) no auxiliary input, 2) baseline WMH volume, 3) both baseline WMH and SL volumes, and 4) Gaussian noise. The WMH and SL volumes were obtained from WMH and SL labels/masks (see Section 3.1). Whereas, an array of 32 random noises which follow Gaussian distribution (Gaussian noise) of $z \sim \mathcal{N}(0, 1)$ was used as per our previous work (Rachmadi et al., 2019a). 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.

3. Data and Experiments

3.1. Subjects and Data

We used MRI data 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). 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, 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., 2015). Ground truth segmentations were performed using a multi-spectral semi-automatic method (Valdés Hernández et al., 2015) only from baseline and 1-year follow-up scan visits in the image space of the T1-weighted scan of the second visit, in $n = 152$ (out of 264) patients. T2-weighted, FLAIR, gradient echo, and T1-weighted 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 slice thickness of $0.9375 \times 0.9375 \times 4$ mm. We used data from all patients who had the three scan visits and ground truth generated as per above. Hence, our sample consists on MRI data (i.e., $s = n \times 2 = 304$ MRI scans) for *baseline* and *1-year follow-up* data. Out of all patients, there are 70 of them that have stroke subtype lacunar (46%) with median small vessel disease (SVD) score of 1. Other demographics and clinical characteristics of the patients that provided data for this study can be seen in Table 1.

The primary study that provided the data used a semi-automatic multi-spectral method to produce several brain masks including intracranial volume (ICV), cerebrospinal fluid (CSF), stroke lesions (SL), and WMH, all which were visually checked and manually edited by an expert (Valdés Hernández et al., 2015). The image processing protocol followed to generate these masks is fully explained in (Valdés Hernández et al., 2015). 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., 2015) 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 and decreasing volume of WMH respectively.

Table 1: Demographics and clinical characteristics of the samples used in this study ($n = 152$). SVD and PV stand for small vessel disease and periventricular respectively.

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 [IQR])	1 [0 2]
	PV WMH Fazekas score (median [IQR])	1 [1 2]
	Deep WMH Fazekas score (median [IQR])	1 [1 2]

As previously explained, IM and PM are needed for DEP-GAN (i.e., the non-supervised learning approach of DEP model). We used LOTS-IM with 128 target patches (Rachmadi et al., 2019b) to generate IM from each MRI data. To generate PM, we trained a 2D UResNet (Guerrero et al., 2018) with gold standard WMH and SL masks for WMH and SL segmentation. For this training, we used all subjects in our data set and a 4-fold cross validation training scheme. See Section 3.2 to see how the 4-fold cross validation is done for this study. Furthermore, note that this UResNet is different from the DEP-UResNet, which is newly proposed in this study. Notice that we affix “DEP” key-word to any model’s name used for prediction and delineation of WMH evolution.

3.2. Experiment Setup

For the present study, we opted to use 2D architectures for all our networks rather than 3D ones. This includes the DEP models (i.e., DEP-GAN and DEP-UResNet) for estimating WMH evolution and UResNet for WMH and stroke lesions segmentation. The main reason of this decision was the few data available (i.e. only 152 subjects) in this study. 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., 2017). The 2D version of VA-GAN has been previously tested on synthetic data (Baumgartner et al., 2017).

To train DEP models (i.e., DEP-GAN and DEP-UResNet) and also UResNet (i.e., for generating PM), 4-fold cross validation was performed. Note that cross validation was not used in the previous study that introduced DEP-GAN (Rachmadi et al., 2019a). In each fold, out of 304 MRI data (152 subjects × 2 scans), 228 MRI data (114 subjects × 2 scans) were used for training and 76 MRI data (38 subjects × 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, we omitted slices without any brain tissues. Thus, around 4,000 slices were used in the training process in each fold. Values of IM/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, we first performed an ablation study using different GAN architectures for DEP model, which are based on WGAN-GP, VA-GAN, DEP-GAN with 1 critic (DEP-GAN-1C), and DEP-GAN with 2 critics (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 ($C(x)$). 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 ($D(x)$). The difference between VA-GAN and DEP-GAN-1C is that DEP-GAN-1C has additional losses for optimisation in the training (see Section 2.2.1). Lastly, DEP-GAN-2C has two critics ($C(x)$ and $D(x)$) and additional losses for the training.

Furthermore, we also performed an ablation study using different types of auxiliary input their effects 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. 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 stroke lesion volumes. Note that WMH and SL loads for the whole data set (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 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 Dice similarity coefficient (DSC) was also reported.

3.3. Evaluation Metrics

In this study, we used the following tests 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 Dice similarity coefficient (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 metric 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 agreements (i.e., mean ± 1.96 × standard deviation) between ground truth and predicted WMH volumes of the follow-up assessment. We also calculated the **Volumetric correlation** between ground truth and follow-up predicted WMH volumes, complementary to the Bland-Altman plot. Whereas, for evaluating the **spatial agreement** between ground truth and automatic delineation results, we used the Dice similarity coefficient (DSC). Higher DSC means better performance, and it can be computed as follow:

$$DSC = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{6}$$

where TP is true positive, FP is false positive and FN is false negative.

In addition, we performed **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. For this, analyses of covariance (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 stroke lesion (abbreviated as "index SL"), and 7) volume of old stroke lesions (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.

4. Results and Discussions

4.1. *Ablation study of different GAN architectures for DEP model*

In this ablation study, we used different GAN architectures for DEP model to see the impact of number of critics, location of critic(s), and additional losses. As previously described in Section 3.2, 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 critic for the follow-up image (i.e., $D(x)$) and additional losses for optimisation in the training (see Section 2.2.1), 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.

4.1.1. *Spatial agreement (DSC) and qualitative (visual) analyses*

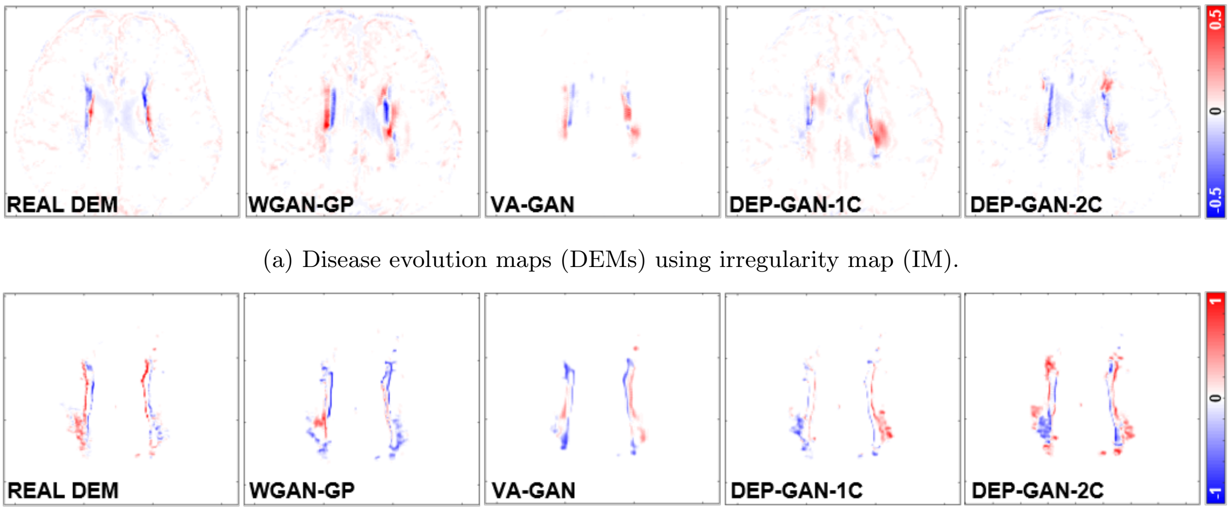
Based on Table 2 (columns 8-13), we can see that DEP-GAN-2C produced better spatial agreement (i.e., higher DSC score) than WGAN-GP, VA-GAN, and DEP-GAN-1C, especially for changing and growing WMH. Qualitative (visual) assessment of generated DEM depicted in Figure 4 also shows that DEP-GAN-2C produced 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, we also can see from both Table 2 and Figure 4 that the use of PM produced better spatial agreement than IM, regardless of the GAN architecture.

4.1.2. *Volumetric agreement (Bland-Altman) and correlation analyses*

From Table 3, we can see 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 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

Table 2: Results from ablation study of different GAN architectures for DEP models. We calculated the prediction error of WMH change, volumetric agreement of WMH volume, and spatial agreement of WMH evolution, compared to the gold standard expert-delineated WMH masks (i.e., three-class DEM labels), using the Dice similarity coefficient (DSC). "Vol." stands for volumetric, "LoA" stands for limit of agreement, "gr" and "sh" stand for number of subjects that have increasing and decreasing WMH volume (i.e., 98 and 54 respectively), and "G" and "S" stand for percentage of subjects correctly predicted as having growing and shrinking WMH by DEP models. Thus, $G = p_{gr}/gr$ and $S = p_{sh}/sh$ where "p_{gr}" and "p_{sh}" stand for number of subjects predicted as having growing and shrinking WMH respectively. The best value for each learning approaches and evaluation metrics is written in bold.

Unsupervised	**Grow**	**Shrink**	**Avg.** [%]	**Vol. Bias** [*ml*]	**Lower**	**Upper**	**Entire**	**Change**	**Stable**	**Shrink**	**Grow**	**Avg.** ((St+
(IM)	**(G)** [%]	**(S)** [%]	((G+S)/2)	*mean(std)*	**LoA** [*ml*]	**LoA** [*ml*]	**WMH**	**(C)**	**(St)**	**(Sr)**	**(Gr)**	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	**Grow**	**Shrink**	**Avg.** [%]	**Vol. Bias** [*ml*]	**Lower**	**Upper**	**Entire**	**Change**	**Stable**	**Shrink**	**Grow**	**Avg.**((St+
Supervised (PM)	**(G)** [%]	**(S)** [%]	((G+S)/2)	*mean(std)*	**LoA** [*ml*]	**LoA** [*ml*]	**WMH**	**(C)**	**(St)**	**(Sr)**	**(Gr)**	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) Disease evolution maps (DEMs) using irregularity map (IM).

(b) Disease evolution maps (DEMs) using probability map (PM).

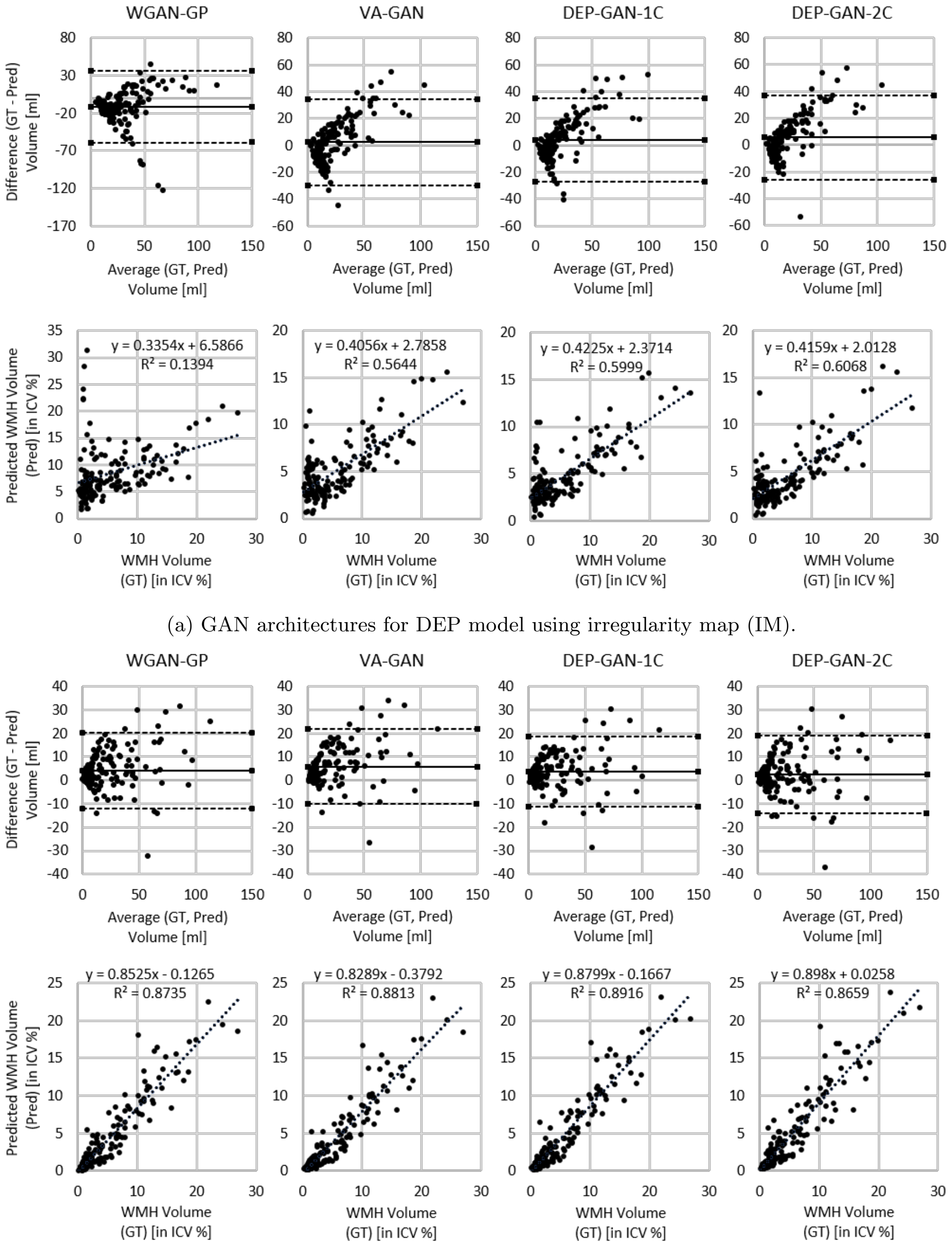
Figure 4: Examples of real DEM and generated DEMs produced by different GAN architectures for DEP model. From left to right: real DEM and generated DEMs produced by WGAN-GP, VA-GAN, DEP-GAN with 1 critic (DEP-GAN-1C), and DEP-GAN with 2 critics (DEP-GAN-2C) respectively.

Table 3: Volumetric correlation analysis in ablation study of GAN architectures for DEP model. The best value for each correlation metric is written in bold.

Unsupervised (IM)	WGAN-GP	VA-GAN	DEP-GAN-1C	DEP-GAN-2C
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
R^2	0.8735	0.8813	**0.8916**	0.8659
Trend	$y = 0.8525x - 0.1265$	$y = 0.8289x - 0.3792$	$y = 0.8799x - 0.1667$	$\mathbf{y = 0.898x + 0.0258}$

when IM was used. However, VA-GAN's good performance in the volumetric agreement analysis did not translate to good spatial agreement as previously described in Section 4.1.1.

Based on the Bland-Altman and correlation plots depicted in Figure 5, we can see that PM is better than IM for representing the volumetric change of WMH. From the correlation plots, we can see that



(a) GAN architectures for DEP model using irregularity map (IM).

(b) GAN architectures for DEP model using probability map (PM).

Figure 5: Volumetric agreement (in *ml*) and correlation (in ICV %) analyses between ground truth (GT) 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.

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.

4.1.3. Prediction error analysis and discussion

From Table 2 (columns 2-4), we can see that most the GAN-based 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, we can conclude that DEP-GAN with 2 critics (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-based deep learning methods are used.

4.2. Ablation study of auxiliary input in DEP models

In this ablation study, we used different types (modalities) of auxiliary input to see how they affect the performance of DEP models for predicting the evolution of WMH. We tested 4 modalities of auxiliary input, 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 reported.

4.2.1. Volumetric agreement (Bland-Altman) and correlation analyses

From Table 4 (columns 5-7), we can see 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, we also can see almost all DEP-UResNet 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 standard deviation 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 6, 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 4.1.2. 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 and Figure 7).

4.2.2. 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 for the entire WMH 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 4 columns 8-13). Furthermore, it also outperformed the rest of the models on changing, shrinking, and growing WMH clusters. Compared to the "vanilla" DEP-UResNet with No Auxiliary, Wilcoxon tests yielded p-values of 0.1563, 0.0425, 0.0625, 0.0313, 0.0313, and 0.0425 for the entire WMH, changing WMH, stable WMH, shrinking WMH, growing WMH, and average respectively. 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 im-

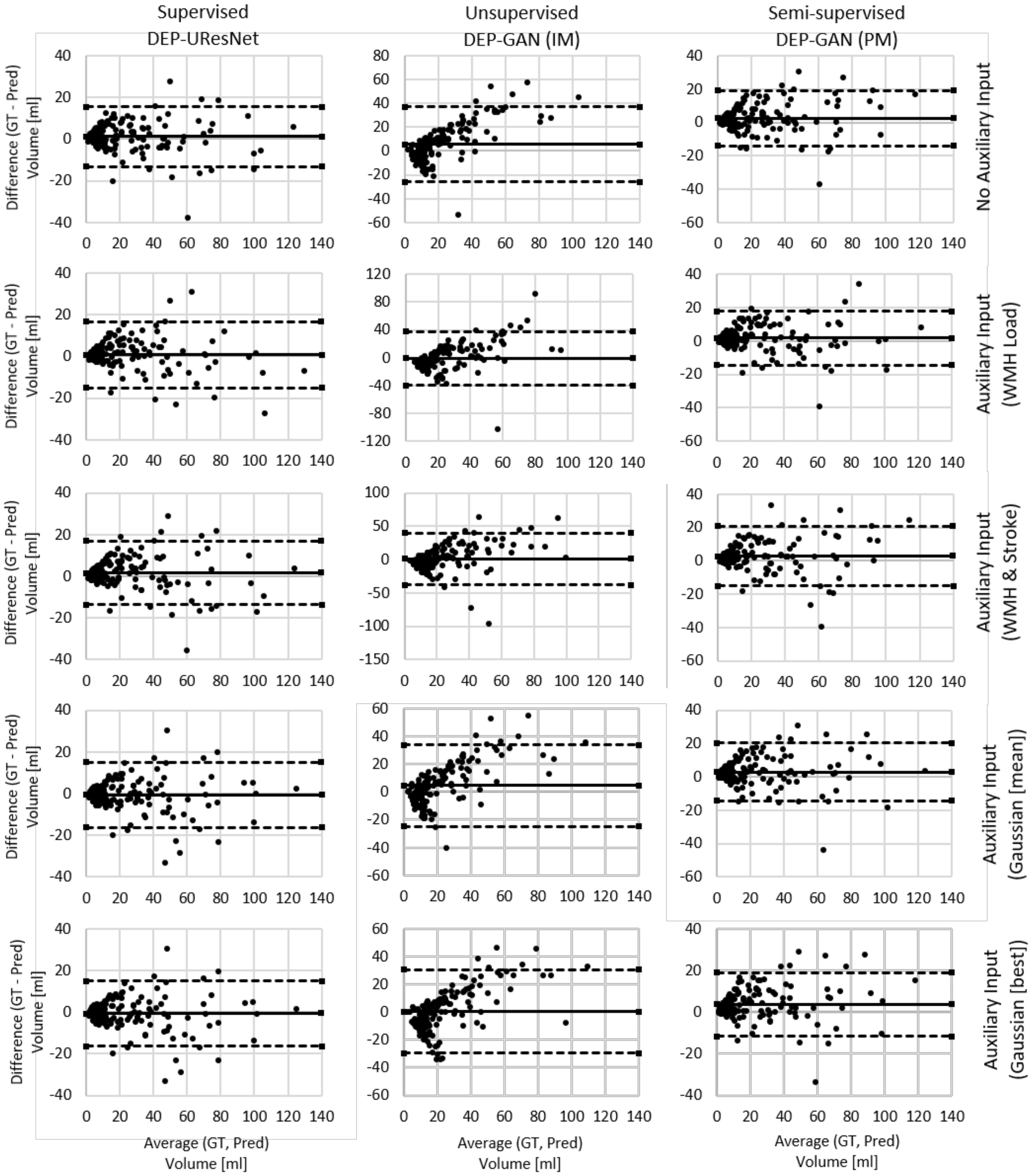


Figure 6: Volumetric agreement analysis (in *ml*) between ground truth (GT) and predicted volume of WMH with different types/modalities of auxiliary input (Pred) using Bland-Altman plot which correspond to data presented in Table 4. Solid lines correspond to “Vol. Bias” while dashed lines correspond to either “Lower LoA” or “Upper LoA” of the same table. “LoA” stands for limit of agreement.

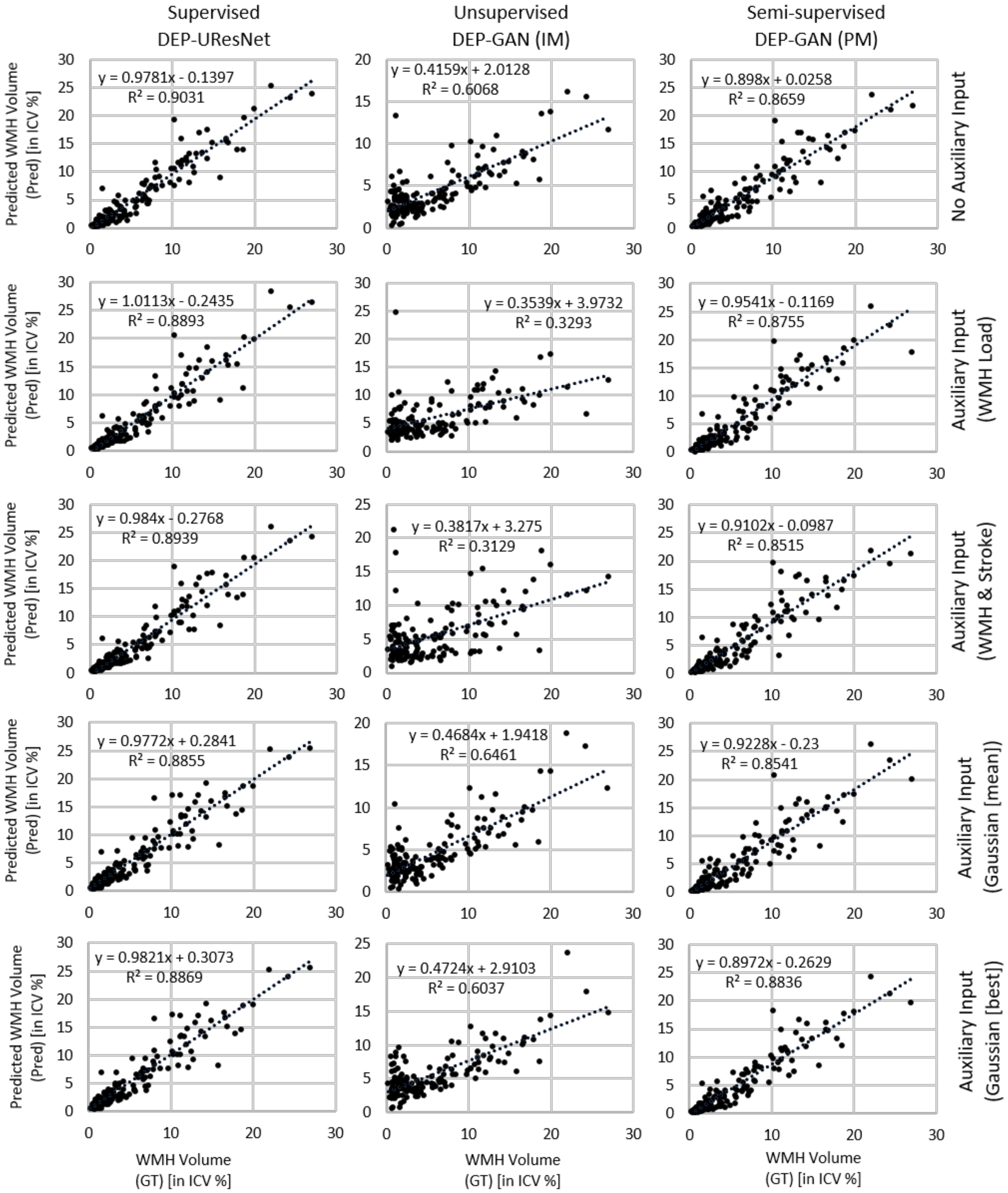


Figure 7: Correlation plots between manual WMH volume produced by the expert (GT) and predicted WMH volume by various DEP models with different types/modalities of auxiliary input (Pred). WMH volume is in the percentage of intracranial volume (ICV) to remove any potential bias associated with head size.

Table 4: Results from ablation study of auxiliary input in DEP models. Prediction errorof WMH change, volumetric agreement of WMH volume, and spatial agreement ofWMH evolution in Dice similarity coefficient (DSC) were calculated to the gold standard expert-delineated WMH masks (i.e., three-class DEM labels). "Vol." stands for volumetric, "LoA" stands for limit of agreement, "gr" and "sh" stand for number of subjects that have increasing and decreasing WMH volume (i.e., 98 and 54 respectively), and "G" and "S" stand for percentage of subjects correctly predicted as having growing and shrinking WMH by DEP models. Thus, $G = p_{gr}/gr$ and $S = p_{sh}/sh$ where "p_{gr}" and "p_{sh}" stand for number of subjects predicted as having growing and shrinking WMH. The best value for each machine learning approaches and evaluation metrics is written in bold. Furthermore, the best value of all learning approaches for each evaluation metrics is underlined and written in bold.

Supervised (DEP-UResNet)	**Grow (G)**[%]	**Shrink (S)** [%]	**Avg.** [%] ((G+S)/2)	**Vol. Bias** [*ml*] *mean(std)*	**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(std)*	**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(std)*	**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

Table 5: Volumetric correlation analysis of DEP models with different types/modalities of auxiliary input in ablation study of auxiliary input.

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	$y = 0.9821x + 0.3073$
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	$y = 0.4724x + 2.9103$
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	$y = 0.8972x - 0.2629$

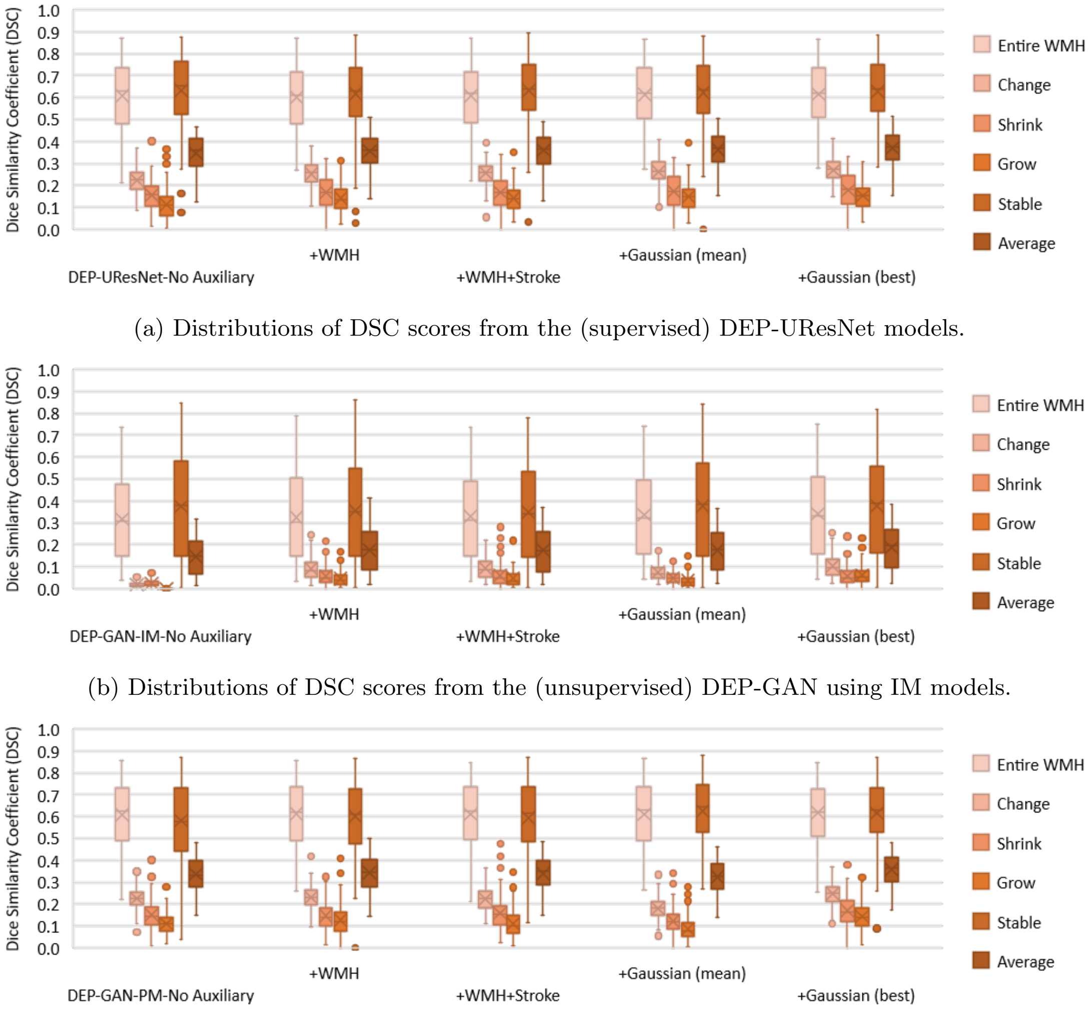
proved if the "best" Gaussian noise is used and evaluated (see the "DEP-UResNet+Gaussian (best)" in Table 4)

Based on Table 4 results, the (indirectly supervised) DEP-GAN using PM had close performance to the (supervised) DEP-UResNet in all performed analyses, especially in the spatial agreement analysis (columns 8-13). To give a better visualisation of the spread of the performances, we plotted 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 using box-plot in Figure 8. Furthermore, we also conducted Wilcoxon test 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 8, we can see that performances of DEP-GAN using PM and 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.

4.2.3. 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 10. Furthermore, it is very difficult to discern the borders between growing and shrinking re-



(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 8: Distributions of DSC scores from all evaluated DEP models in auxiliary input ablation study. These distributions correspond to the Table 4, columns 8-13.

gions when stroke lesions coalesce with WMH even though stroke lesions were removed from the analysis as previously explained. Nevertheless, inaccuracies while determining the borders between coalescent WMH and stroke lesions and the small size of the volume changes in each WMH cluster (Rachmadi et al., 2018a) might have influenced in the low DSC values obtained in the regions that experienced change as seen in Table 4. Furthermore, 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 10 and 11.

Based on qualitative (visual) assessment of DEM produced by DEP-GAN using IM/PM depicted in Figure 9, 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 11. 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 ir-

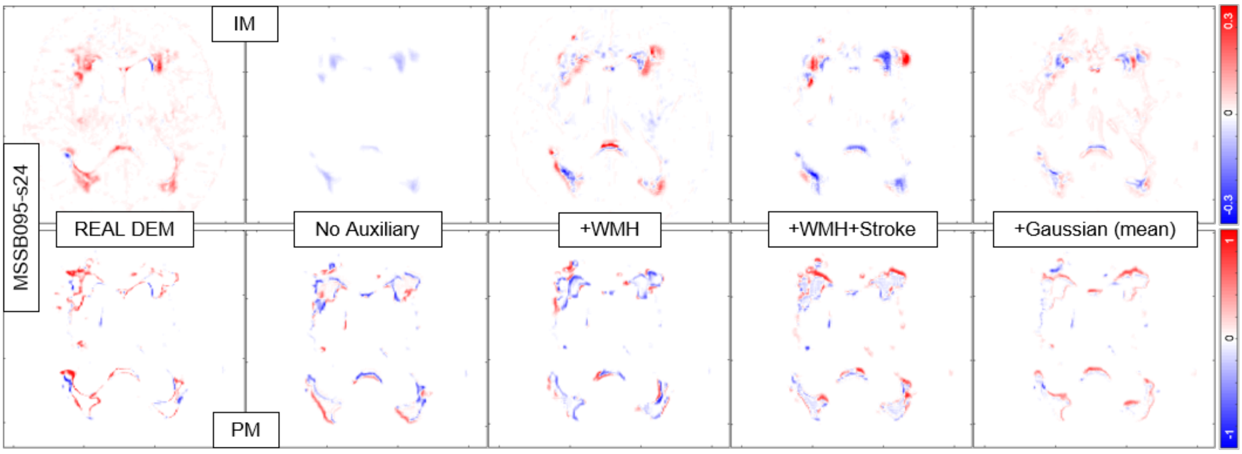


Figure 9: Qualitative (visual) assessment of DEM produced by the unsupervised and indirectly supervised DEP models; DEP-GAN using irregularity map (IM) and DEP-GAN using probability map (PM), with different types/modalities of auxiliary input. The corresponding T2-FLAIR (input data) can be seen in Figure 11.

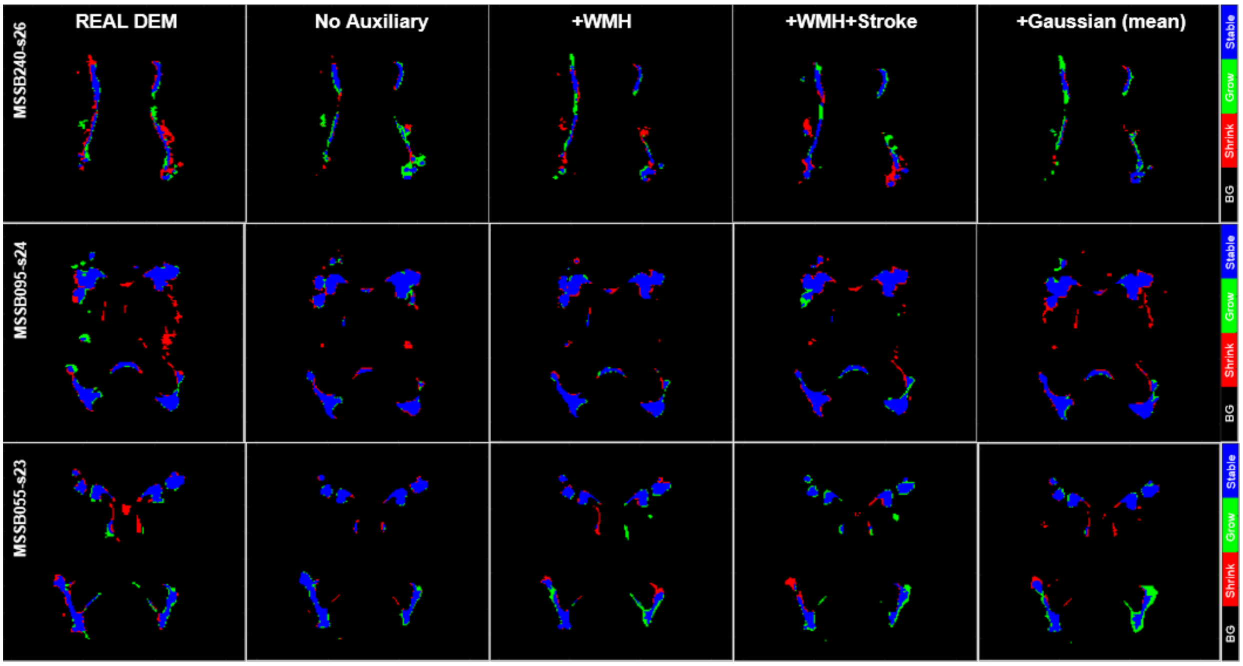


Figure 10: Qualitative (visual) assessment of DEM label produced by the supervised DEP model, DEP-UResNet, with different types/modalities of auxiliary input. The corresponding T2-FLAIR (input data) can be seen in Figure 11.

regularity/probability values greater than or equal to 0.178 for IM and 0.5 for PM (Rachmadi et al., 2019b). 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.

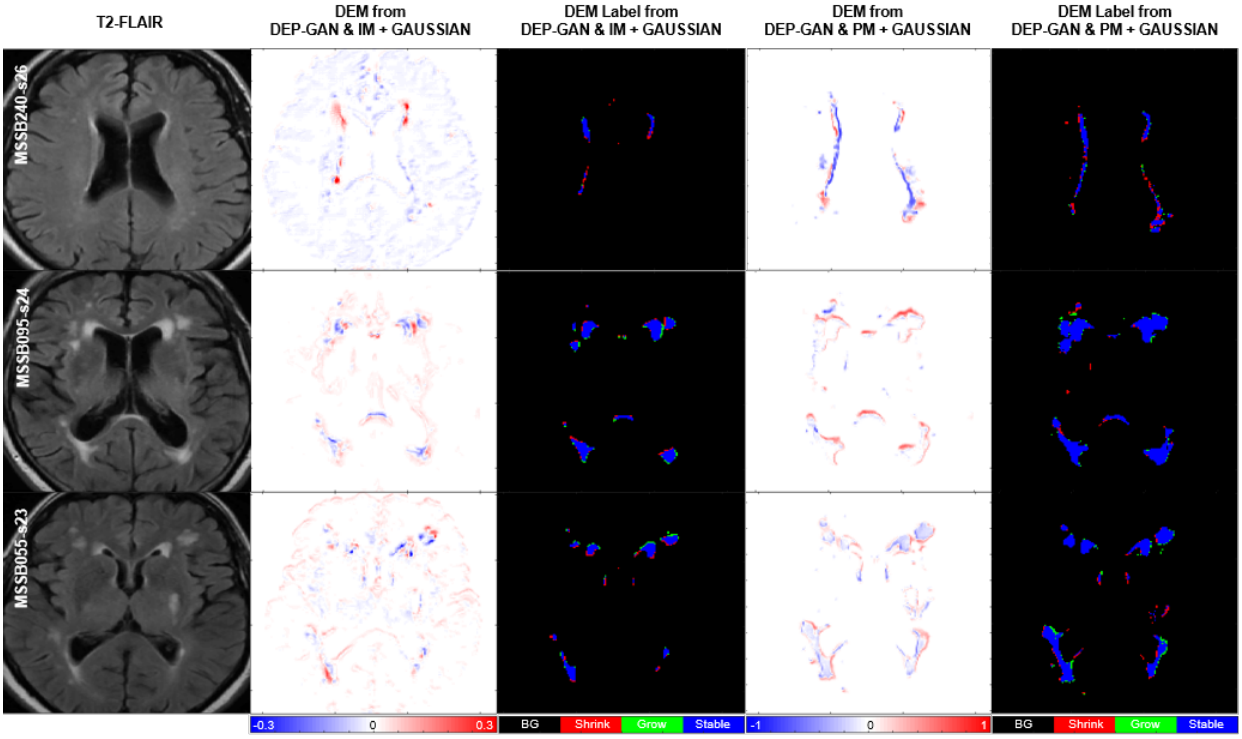

T2-FLAIR
DEM from DEP-GAN & IM + GAUSSIAN
DEM Label from DEP-GAN & IM + GAUSSIAN
DEM from DEP-GAN & PM + GAUSSIAN
DEM Label from DEP-GAN & PM + GAUSSIAN
MSSB240-s26
MSSB095-s24
MSSB055-s23
-0.3
0
0.3
BG
Shrink
Grow
Stable
-1
0
1
BG
Shrink
Grow
Stable

Figure 11: Qualitative (visual) assessment of DEM and its corresponding DEM label produced by the unsupervised and indirectly supervised DEP models; DEP-GAN using irregularity map (IM) and DEP-GAN using probability map (PM) respectively, with different types/modalities of auxiliary input. The corresponding golden standard of DEM label can be seen in Figure 10.

Table 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)

4.2.4. Clinical plausibility analysis

From Table 6, we can see 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 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 stroke lesions (SL) and hypertension have been acknowledged as contributed factors. However, from the results of the ANCOVA models (Table 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 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.

4.2.5. Prediction error analysis and discussion

From Table 4 (columns 2-4), we can see 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, we also can see 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. Conclusion and Future Work

In this study, we proposed a training scheme to predict the evolution of WMH using deep learning algorithms called Disease Evolution Predictor (DEP) model. To the best of our knowledge, this is the first extensive study on modelling WMH evolution using deep learning algorithms. Furthermore, we evaluated different configurations of DEP models: unsupervised, indirectly supervised, and supervised (i.e., DEP-GAN using irregularity map (IM), DEP-GAN using probability map (PM), and DEP-UResNet) with different types of auxiliary input (i.e., Gaussian noise, WMH load, and WMH and stroke lesions (SL) loads). 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 with 2 critics 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 metrics 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 our experiments, on average, supervised DEP-UResNet yielded the best results in almost every evaluation metric. 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 (neurodegenerative) 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 metrics. 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 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-weighted structural images but not in FLAIR images for WMH segmentation tasks (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, we only used single channel 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., 2017). 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.

Acknowledgements

The first author (MFR) would like to thank Indonesia Endowment Fund for Education (LPDP) of Ministry of Finance, Republic of Indonesia, for funding his study at School of Informatics, the University of Edinburgh. Funds from Row Fogo Charitable Trust (Grant No. BRO-D.FID3668413) (MCVH) are also gratefully acknowledged. The primary study that provided data for this study was funded by the Wellcome Trust (Ref No. 088134/Z/0). Authors also thank support from the 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.

References

Baumgartner, C. F., Koch, L. M., Tezcan, K. C., Ang, J. X., Konukoglu, E., 2017. Visual feature attribution using wasserstein gans. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit.

Bland, J. M., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet 327 (8476), 307–310.

Chappell, F. M., Hernández, M. d. C. V., Makin, S. D., Shuler, K., Sakka, E., Dennis, M. S., Armitage, P. A., Maniega, S. M., 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), 78.

Cho, A.-H., Kim, H.-R., Kim, W., Yang, D. W., 2015. White Matter Hyperintensity in Ischemic Stroke Patients: It May Regress Over Time. Journal of Stroke 17 (1), 60.

Choi, H., Jin, K. H., Initiative, A. D. N., et al., 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., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 424–432.

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

Durand-Birchenall, J., Leclercq, C., Daouk, J., Monet, P., Godefroy, O., Bugnicourt, J.-M., 2 2012. Attenuation of brain white matter lesions after lacunar stroke. International journal of preventive medicine 3 (2), 134–138.

Fan, H., Su, H., Guibas, L. J., 2017. A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613.

Godin, O., Tzourio, C., Maillard, P., Alperovitch, A., Mazoyer, B., 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., 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., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680.

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., Barkhof, F., 2008. Progression of White Matter Hyperintensities and Incidence of New Lacunes Over a 3-Year Period. Stroke 39 (5), 1414–1420.

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

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C., 2017. Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5767–5777.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A., 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264.

Hernández, M. d. C. V., González-Castro, V., Ghandour, D. T., Wang, X., Doubal, F., Maniega, S. M., Armitage, P. A., 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.

Jenkinson, M., Bannister, P., Brady, M., 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., 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., 2018. Microstructural and metabolic changes in the longitudinal progression of white matter hyperintensities.

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

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

Luo, X., Jiaerken, Y., Yu, X., Huang, P., Qiu, T., Jia, Y., Li, K., Xu, X., Shen, Z., Guan, X., Zhou, J., Zhang, M., Adni, F. T. A. D. N. I., 5 2017. Associations between APOE genotype and cerebral small-vessel disease: a longitudinal study. Oncotarget 8 (27), 44477–44489.

Maillard, P., Fletcher, E., Harvey, D., Carmichael, O., Reed, B., Mungas, D., 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., 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.

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

Moriya, Y., Kozaki, K., Nagai, K., Toba, K., 2009. Attenuation of brain white matter hyperintensities after cerebral infarction. American Journal of Neuroradiology 30 (3), 3174.

Pasi, M., Van Uden, I. W., Tuladhar, A. M., De Leeuw, F. E., Pantoni, L., 2016. White Matter Microstructural Damage on Diffusion Tensor Imaging in Cerebral Small Vessel Disease: Clinical Consequences. Stroke 47 (6), 1679–1684.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., 2018. Film: Visual reasoning with a general conditioning layer. In: Thirty-Second AAAI Conference on Artificial Intelligence.

Power C, M., Deal A, J., Sharrett Richey, A., Jack Jr, C. R., Knopman, D., Mosley H, T., Gottesman F, R., 2015. Smoking and white matter hyperintensity progression: The ARIC-MRI Study. Neurology 84 (8), 841–848.

Prins, N. D., Scheltens, P., 2015. White matter hyperintensities, cognitive impairment and dementia: an update. Nature reviews. Neurology 11 (3), 157–65.

Rachmadi, M., Valdés-Hernández, M., Agan, M., Komura, T., 2017. 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., del C. Valdés-Hernández, M., Makin, S., Wardlaw, J. M., Komura, T., 2019a. Predicting the evolution of white matter hyperintensities in brain mri using generative adversarial networks and irregularity map. In: Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 146–154.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Agan, M. L. F., Di Perri, C., Komura, T., Initiative, A. D. N., et al., 2018a. 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., Komura, T., 2018b. Automatic irregular texture detection in brain mri without human supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 506–513.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Komura, T., 2018c. 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. Springer, pp. 85–93.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Li, H., Guerrero, R., Meijboom, R., Wiseman, S., Waldman, A., Zhang, J., Rueckert, D., Wardlaw, J., et al., 2019b. 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. BioRxiv, 334292.

Ramirez, J., McNeely, A. A., Berezuk, C., Gao, F., 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 (MAR), 1–9.

Rovira Cañellas, A., Mínguez, B., Aymerich, F. X., Jacas, C., Huerga, E., Córdoba, J., Alonso, J., 2007. Decreased white matter lesion volume and improved cognitive function after liver transplantation. Hepatology 46 (5), 1485–1490.

Sachdev, P., Wen, W., Chen, X., Brodaty, H., 2007. Progression of white matter hyperintensities in elderly individuals over 3 years. Neurology 68 (3), 214–222.

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., 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, R., Enzinger, C., Ropele, S., Schmidt, H., 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., Schmidt, H., 2002a. Risk factors and progression of small vessel disease-related cerebral abnormalities. In: Ageing and Dementia Current and Future Concepts. Springer, pp. 47–52.

Schmidt, R., Schmidt, H., Kapeller, P., Lechner, A., Fazekas, F., 2002b. Evolution of white matter lesions. Cerebrovascular Diseases 13 (SUPPL. 2), 16–20.

Schmidt, R., Seiler, S., Loitfelder, M., 2016. Longitudinal change of small-vessel disease-related brain abnormalities. Journal of Cerebral Blood Flow and Metabolism 36 (1), 26–39.

Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., Initiative, A. D. N., et al., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. Neuroimage 189, 276–287.

Valdés Hernández, M. d. C., Armitage, P. A., Thrippleton, M. J., Chappell, F., Sandeman, E., Muñoz Maniega, S., Shuler, K., Wardlaw, J. M., 2015. 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.

van Dijk, E. J., Prins, N. D., Vrooman, H. A., Hofman, A., Koudstaal, P. J., 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., 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., de Leeuw, F.-E., Tuladhar, A. M., 2017a. Disease progression and regression in sporadic small vessel diseaseinsights 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., De Leeuw, F. E., 2017b. Nonlinear temporal dynamics of cerebral small vessel disease. Neurology 89 (15), 1569–1577.

Veldink, J. H., Scheltens, P., Jonker, C., 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., Arfan, I. M., 6 2013. High Blood Pressure and Cerebral White Matter Lesion Progression in the General Population. Hypertension 61 (6), 1354–1359.

Wardlaw, J. M., Chappell, F. M., Valdés Hernández, M. D. C., Makin, S. D., Staals, J., Shuler, K., Thrippleton, M. J., Armitage, P. A., Munz-Maniega, S., Heye, A. K., Sakka, E., 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., 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.

Yamada, K., Sakai, K., Owada, K., Mineura, K., 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.