Probabilistic Deep Learning with Adversarial Training and Volume Interval Estimation - Better Ways to Perform and Evaluate Predictive Models for White Matter Hyperintensities Evolution

Citation for published version:
Rachmadi, MF, Valdés-hernández, MDC, Maulana, R, Wardlaw, J, Makin, S & Skibbe, H 2021, Probabilistic Deep Learning with Adversarial Training and Volume Interval Estimation - Better Ways to Perform and Evaluate Predictive Models for White Matter Hyperintensities Evolution. in *Predictive Intelligence in Medicine.*, Chapter 16, Predictive Intelligence in Medicine, vol. 12928, pp. 168-180. https://doi.org/10.1007/978-3-030-87602-9_16

Digital Object Identifier (DOI):
10.1007/978-3-030-87602-9_16

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Predictive Intelligence in Medicine

Publisher Rights Statement:
This is the author's peer-reviewed manuscript as accepted for publication.

Probabilistic Deep Learning with Adversarial Training and Volume Interval Estimation - Better Ways to Perform and Evaluate Predictive Models for White Matter Hyperintensities Evolution

Muhammad Febrian Rachmadi*[1][0000−0003−1672−9149], Maria del C. Valdés-Hernández[2][0000−0003−2771−6546], Rizal Maulana[3], Joanna Wardlaw[2], Stephen Makin[4], and Henrik Skibbe[1]

[1] Brain Image Analysis Unit, RIKEN Center for Brain Science, Wako, Japan
febrian.rachmadi@riken.jp
[2] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
[3] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[4] University of Aberdeen, Aberdeen, UK

Abstract. Predicting disease progression always involves a high degree of uncertainty. White matter hyperintensities (WMHs) are the main neuroradiological feature of small vessel disease and a common finding in brain scans of dementia patients and older adults. In predicting their progression previous studies have identified two main challenges: 1) uncertainty in predicting the areas/boundaries of shrinking and growing WMHs and 2) uncertainty in the estimation of future WMHs volume. This study proposes the use of a probabilistic deep learning model called Probabilistic U-Net trained with adversarial loss for capturing and modelling spatial uncertainty in brain MR images. This study also proposes an evaluation procedure named volume interval estimation (VIE) for improving the interpretation of and confidence in the predictive deep learning model. Our experiments show that the Probabilistic U-Net with adversarial training improved the performance of non-probabilistic U-Net in Dice similarity coefficient for predicting the areas of shrinking WMHs, growing WMHs, stable WMHs, and their average by up to 3.35%, 2.94%, 0.47%, and 1.03% respectively. It also improved the volume estimation by 11.84% in the “Correct Prediction in Estimated Volume Interval” metric as per the newly proposed VIE evaluation procedure.

Keywords: Progression prediction · White matter hyperintensities · Volume interval estimation.

1 Introduction

White matter hyperintensities (WMHs) are neuroradiological features often seen in T2-FLAIR brain MRI, characteristic of small vessel disease (SVD), which are

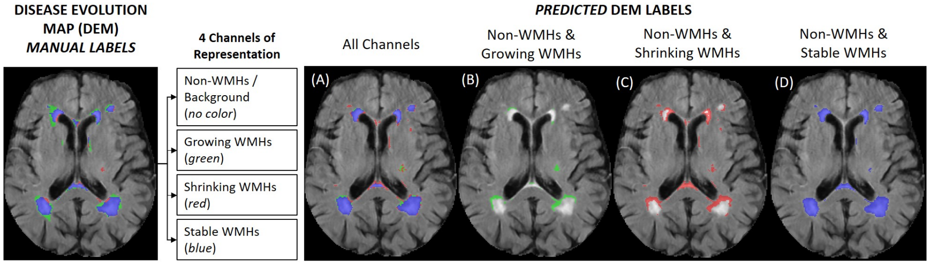

Fig. 1. (Left) Example of Disease evolution map (DEM) produced by subtracting manually generated labels of WMHs at baseline (t0) from manually generated labels of WMHs at follow-up (t1). Green regions are for growing WMHs, red regions are for shrinking WMHs, and blue regions are for stable WMHs (i.e., no changes from t0 to t1). Note there is another channel used to represent the non-WMHs/background in the supervised deep learning model. **(Right)** Different visualizations can be produced based on which channels are used in the testing/inference. *From left to right*: (A) All predicted channels are used to visualize the whole segmentation, (B) only the predicted non-WMHs and growing WMHs channels are used to visualize the segmentation of growing WMHs, (C) only the predicted non-WMHs and shrinking WMHs channels are used to visualize the segmentation of shrinking WMHs, and (D) only the predicted non-WMHs and stable WMHs are used to visualize the segmentation of stable WMHs.

associated with stroke and dementia progression [12]. Clinical studies indicate that the volume of WMHs on a patient may decrease (i.e., regress), stay the same, or increase (i.e., progress) over time [2,12].

Previous studies have proposed various unsupervised and supervised deep learning models to predict the progression (i.e., evolution) of WMHs [8,9]. In the supervised approaches, a deep learning model learns to perform multi-class segmentation of non-WMHs, shrinking WMHs, growing WMHs, and stable WMHs from the namely *disease evolution map* (DEM). The DEM is produced by subtracting manually generated labels of WMHs at baseline (t0) from manually generated labels of WMHs at follow-up (t1) (see Fig. 1).

One study [8] exposed two big challenges in predicting the progression of WMHs: 1) spatial uncertainty in predicting regions of WMHs dynamic changes and their boundaries (i.e., voxels of growth and shrinkage), and 2) uncertainty in the estimation of future WMHs volume (i.e., closeness between the predicted volume of WMHs and the true future volume of WMHs). In relation to the first challenge, it was observed that it is difficult to distinguish the intensities/textures of shrinking and growing WMHs in the MRI sequence used by the study (i.e., T2-FLAIR). This type of uncertainty is commonly known as *aleatoric uncertainty* [4]. In relation to the second challenge, the study showed that different predictive models produced similar error and correlation values in estimating the future volume of WMHs, making it harder to determine the best predictive model.

Our main contributions are listed as follows. Firstly, we propose a combination of probabilistic deep learning model with adversarial training to capture spatial uncertainties to predict WMHs evolution. Secondly, we propose a new evalu-

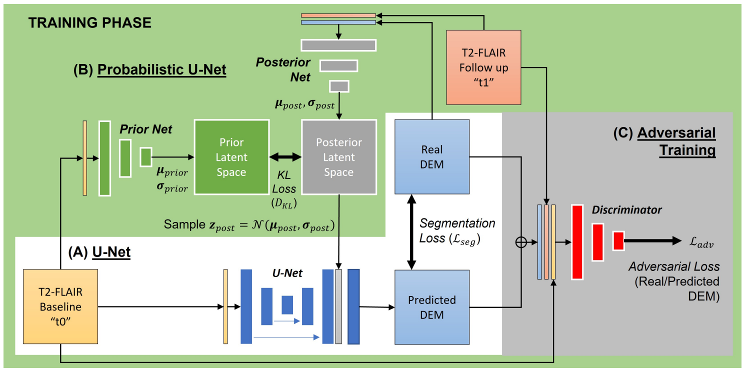
TRAINING PHASE
(B) Probabilistic U-Net
Posterior Net
$\boldsymbol{\mu}_{post}, \boldsymbol{\sigma}_{post}$
T2-FLAIR Follow up "t1"
Prior Net
Prior Latent Space
$\boldsymbol{\mu}_{prior}$
$\boldsymbol{\sigma}_{prior}$
KL Loss (D_{KL})
Posterior Latent Space
Real DEM
(C) Adversarial Training
Sample $\boldsymbol{z}_{post} = \mathcal{N}(\boldsymbol{\mu}_{post}, \boldsymbol{\sigma}_{post})$
Segmentation Loss ($\mathcal{L}_{seg}$)
Discriminator
$\mathcal{L}_{adv}$
$\oplus$
(A) U-Net
U-Net
T2-FLAIR Baseline "t0"
Predicted DEM
Adversarial Loss (Real/Predicted DEM)

Fig. 2. Illustration of the deep learning models' training phase used in this study. We investigate three different training schemes, which are (A) deterministic training using U-Net [10], (B) probabilistic training using Probabilistic U-Net [5], and (C) adversarial training using a GAN discriminator [3,7], all of which can be combined together. Symbol $\oplus$ stands for OR operation. Full schematics (i.e., figures) of all networks are available in the Supplementary Materials.

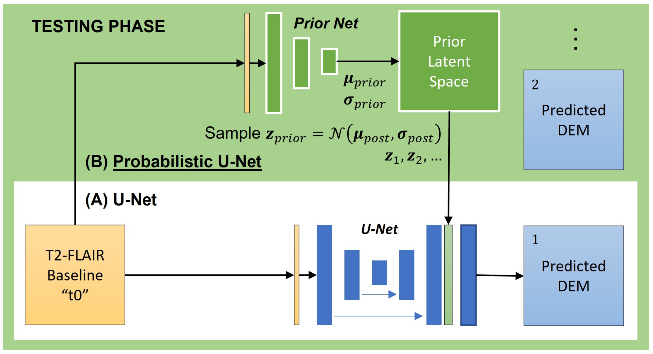
TESTING PHASE
Prior Net
Prior Latent Space
⋮
$\boldsymbol{\mu}_{prior}$
$\boldsymbol{\sigma}_{prior}$
2
Predicted DEM
Sample $\boldsymbol{z}_{prior} = \mathcal{N}(\boldsymbol{\mu}_{post}, \boldsymbol{\sigma}_{post})$
$\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$
(B) Probabilistic U-Net
(A) U-Net
U-Net
T2-FLAIR Baseline "t0"
1
Predicted DEM

Fig. 3. Illustration of the testing/inference phase of the deep learning model used in this study. In this study, we perform two types of inference (based on the training phase previously performed): (A) deterministic inference using U-Net and (B) probabilistic inference using Probabilistic U-Net.

ation procedure, which we name volume interval estimation (VIE), for achieving better interpretation and higher confidence in our predictive models in estimating the future volume of WMHs. The codes and trained model are available on our GitHub page (https://github.com/febrianrachmadi/probunet-gan-vie).

2 Proposed Approach

2.1 Probabilistic Model for Capturing Spatial Uncertainty

Uncertainties are unavoidable when predicting the progression of WMHs, and a previous study showed that incorporating uncertainties into a deep learning model produced the best prediction results [8]. However, the models evaluated in [8] only incorporate external uncertainties (i.e., non-image factors of stroke lesions' volume and unrelated Gaussian noise) and not primary/secondary information coming from brain MRI scans (e.g. statistical spatial maps showing the association of specific WMHs voxels with clinical variables like smoking status).

In this study, we propose the use of the Probabilistic U-Net [5] to capture uncertainties from the brain MR images when predicting the progression of WMHs. The Probabilistic U-Net combines a U-Net [10] with an auxiliary decoder network called Prior Net. The Prior Net models uncertainty in the data as a multivariate Gaussian distribution called prior latent space. The Prior Net learns the prior latent space from another decoder network called Posterior Net that generates a posterior latent space from training data (Fig. 2(B)). The posterior latent space and the Posterior Net are only available during training. *Kullback-Leibler* Divergence ($\mathcal{D}_{KL}$) score is used during training to make the prior latent space similar to the posterior latent space. In testing/inference (Fig. 3), the learned prior latent space is used to sample $\boldsymbol{z}$, which are broadcasted and concatenated to the original U-Net for generating some variations in the predicted segmentation for the same input image. While variations of prediction are inferred from a few samples from a low-dimensional latent space (i.e., sample $\boldsymbol{z}$), most information used for predicting the evolution of WMHs in spatial space still comes from the U-Net (i.e., U-Net's feature maps that are concatenated with the samples).

2.2 Adversarial Training for the Predictive Deep Learning Model

A previous study [8] also showed that adversarial training can help producing good predictions by ensuring that each prediction (i.e., predicted DEM) "looks" similar to the real DEM. However, adversarial training was only used for a GAN-based model (i.e., without any manual DEM). In this study, we propose adding adversarial training/loss in the supervised approach where the GAN's discriminator tries to distinguish the "real" manual DEM from the"fake" predicted DEM produced by the U-Net/Probabilistic U-Net. Adding adversarial loss in the training phase is advantageous because it uses information from the entire image space (i.e., global context information) rather than local (i.e., pixel-wise) information usually given by the traditional segmentation loss. Fig. 2(C) shows how the GAN's discriminator is used in the training phase.

2.3 Volume Interval Estimation for Better Interpretation

One of the many challenges in predicting the progression of WMHs is to ascertain the quality of the prediction, especially when estimating the future volume of WMHs. Despite the existence of several metrics for quality control of an image estimation machine-learning algorithm [1], predictive deep learning models normally use the mean square error (MSE) to evaluate how close the predicted future volumes of WMHs are to the true future volumes after the training phase. However, how can we calculate the MSE in a real world scenario where the real future volume of WMHs is unknown?

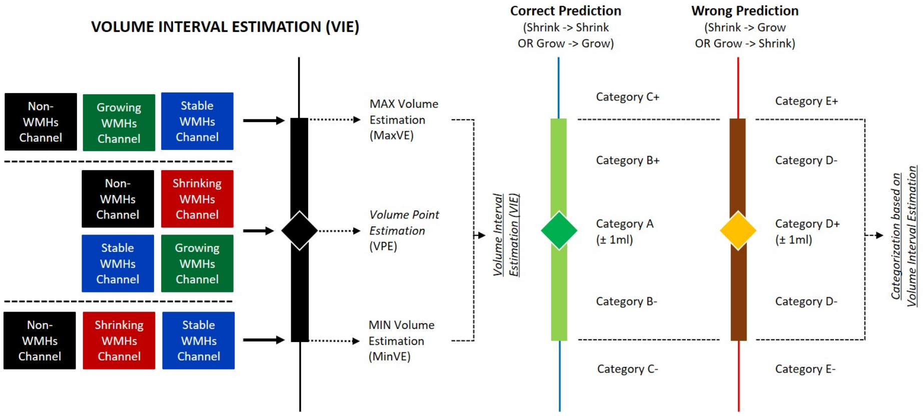

Fig. 4. (Left) Visualization of Volume Interval Estimation (VIE) produced by using subsets of predicted channels of non-WMHs, shrinking WMHs, growing WMHs, and stable WMHs. Note that the normal volume point estimation (VPE) is done by using all predicted channels. **(Right)** By using volume interval estimation, we can categorize prediction results more accurately (i.e., not only correct and wrong predictions). Detailed categorization scheme is shown in Table 1.

For better interpretation and confidence in our prediction model, we propose using the Volume Interval Estimation (VIE). Instead of evaluating how close the predicted volume point estimation (VPE) is to the true volume of future WMHs at time point "1" (True time-point 1 Volume, or Tt1V), we evaluate where Tt1V lies within the VIE, i.e., the interval bounded by the maximum (MaxVE) and minimum (MinVE) volume estimations. VIE's interval is bounded by two extreme assumptions of WMHs progression: 1) there are no shrinking WMHs (which produces MaxVE) and 2) there are no growing WMHs (which produces MinVE). Note that the normal assumption for the WMHs progression (i.e., WMHs can be stable, growing, or shrinking) is located between these two extreme assumptions considering the stable WMHs to be regions of chronic damage (i.e., otherwise MinVE would be equal to zero). Thus, VPE is located between the MinVE and MaxVE. As illustrated in Fig. 4 (left), MinVE is produced by dropping the growing WMHs channel in the predicted DEM while MaxVE is produced by dropping the shrinking WMHs channel.

We can further categorize VIE according to 1) the location of Tt1V within VIE and 2) whether the volume estimation is correctly predicted or not (i.e., patient with growing WMHs is correctly predicted to have growing WMHs, and so on). Fig. 4 (right) and Table 1 illustrate and describe each VIE's category.

3 Dataset and Experimental Setting

3.1 Dataset and Cross Validation

We use MRI data from all stroke patients ($n = 152$) enrolled in a study of stroke mechanisms [12], imaged at three time points (i.e., first time (baseline scan), at

Table 1. Categorization of the proposed volume interval estimation (VIE) based on the position of true future (follow-up) Total WMHs volume (Tt1V) in the predicted volume interval between maximum volume estimation (MaxVE), minimum volume estimation (MinVE), and volume point estimation (VPE). Visualization of the proposed volume interval estimation can be seen in Fig. 4. (For the dataset used in this study 1 *ml* is approximately 284 voxels, as 1 voxel represents a volume of 0.00351 ml.)

Category	**Description**
A	*Correct prediction* (VPE - 1 *ml* <= Tt1V <= VPE + 1 *ml*)
B+	*Correct prediction* (VPE + 1 *ml* <= Tt1V <= MaxVE)
B-	*Correct prediction* (MinVE <= Tt1V <= VPE - 1 *ml*)
C+	*Correct prediction* (Tt1V > MaxVE)
C-	*Correct prediction* (Tt1V < MinVE)
D+	*Wrong prediction* (VPE - 1 *ml* <= Tt1V <= VPE + 1 *ml*)
D-	(VPE + 1 *ml* <= Tt1V <= MaxVE **OR** VPE + 1 *ml* <= Tt1V <= MaxVE)
E+	*Wrong prediction* (Tt1V > MaxVE)
E-	*Wrong prediction* (Tt1V < MinVE)

approximately 3 months, and a year after). This study uses the baseline (t0) and 1-year follow-up (t1) MRI data ($s = n \times 2 = 304$), both acquired at a GE 1.5T scanner following the same imaging protocol, explained in [11]. These data are pre-processed (co-registered, brain-extracted, filtered, and normalised) as explained in [9,8]. The spatial resolution of the images used in this study is $256 \times 256 \times 42$ with slice thickness of $0.9375 \times 0.9375 \times 4$ cubic mm. To make sure data from all patients are used in the testing and evaluation, we perform 4-fold cross validation where each fold uses 114 and 38 patients for training and testing respectively. Each model is trained for 64 epochs in one experiment.

3.2 Segmentation Loss ($\mathcal{L}_{seg}$)

In this study, we use the non-linear *softmax* function at the segmentation layer; see Eq. 1. The parameter $\boldsymbol{s}$ is the output of the segmentation layer. The network classifies each voxel either as non-WMHs, shrinking WMHs, growing WMHs, or stable WMHs. Thus, the number of output classes is set to $C = 4$.

$$p_i = \sigma(\boldsymbol{s})_i = \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}} \text{ for } i = 1, ..., C \tag{1}$$

We tested two different segmentation losses ($\mathcal{L}_{seg}$): 1) weighted cross entropy (WCE) (Eq. 2), and 2) *alpha* weighted focal loss (FL) [6] (Eq. 3). In both equations, tar_i is the true target class for each voxel and p_i is the probability of each voxel to be of the target class i. Whereas, w_i is the weight loss of class i in WCE and α_i is the weight loss of class i in FL. A larger weight loss for class i indicates that class i is predominant, contributing a larger loss value in total. Finally, γ is FL's hyperparameter, which is set to $\gamma = 2$ following the recommendation of the original paper [6]. Based on our preliminary experiments, the best weights for both WCE (i.e., $\boldsymbol{w} = (w_{i=1}, w_{i=2}, w_{i=3}, w_{i=4})$) and FL (i.e., $\boldsymbol{\alpha} = (\alpha_{i=1}, \alpha_{i=2}, \alpha_{i=3}, \alpha_{i=4})$) are 0.25, 0.75, 0.75, and 0.5 for non-WMHs ($i = 1$), shrinking WMHs ($i = 2$), growing WMHs ($i = 3$), or stable WMHs ($i = 4$) respectively.

$$\mathcal{L}_{seg}^{WCE} = WCE = -w_i \, tar_i \, log\,(p_i) \tag{2}$$

$$\mathcal{L}_{seg}^{FL} = FL = -\alpha_i \, tar_i \, (1 - p_i)^{\gamma} \, log\,(p_i) \tag{3}$$

3.3 *Kullback-Leibler* Divergence ($\mathcal{D}_{KL}$) for Probabilistic Loss

An additional *Kullback-Leibler* Divergence score ($\mathcal{D}_{KL}$) is used in the training if Probabilistic U-Net setting is used [5]. In this setting, Prior Net and Posterior Net are trained together with the generator (i.e., U-Net) for predicting the DEM. Let Q be the posterior distribution from the Posterior Net and P be the prior distribution from the Prior Net. The difference between the posterior distribution Q and the prior distribution P is penalized by Eq. 4 where X_{post} is the T2-FLAIR at t1, Y_{post} is the true DEM, and X_{prior} is the T2-FLAIR at t0. Following the original paper [5], the dimension for both $\boldsymbol{z}_{post}$ and $\boldsymbol{z}_{prior}$ is 6.

$$\mathcal{D}_{KL}(Q \parallel P) = \mathbb{E}_{\boldsymbol{z}_{post} \sim Q, \boldsymbol{z}_{prior} \sim P}[\log Q(X_{post}, Y_{post}) - \log P(X_{prior})] \tag{4}$$

In the training phase of the Probabilistic U-Net, each segmentation prediction is conditioned to $\boldsymbol{z}_{post} \sim \mathcal{N}\left(\boldsymbol{\mu}_{post}, \boldsymbol{\sigma}_{post}\right) = Q(X_{post}, Y_{post})$ sampled from the Posterior Net. As per the original paper [5], the probabilistic segmentation loss $\mathcal{L}_{seg}^{prob}$ is defined by Eq. 5 with $\beta = 1$. Note that the segmentation loss of $\mathcal{L}_{seg}$ can be either WCE (Eq. 2) or FL (Eq. 3).

$$\mathcal{L}_{seg}^{prob} = \mathcal{L}_{seg}(P_i(p_i | X_{prior}, \boldsymbol{z}_{post})) + \beta \cdot \mathcal{D}_{KL}\left(Q \parallel P\right) \tag{5}$$

In the testing/inference phase, each segmentation prediction is conditioned to $\boldsymbol{z}_{prior} \sim \mathcal{N}\left(\boldsymbol{\mu}_{prior}, \boldsymbol{\sigma}_{prior}\right) = P(X_{prior})$ sampled from the Prior Net. To get the final segmentation, we sampled 30 different $\boldsymbol{z}_{prior}$ from Prior Net to produce 30 different segmentation predictions for each patient and averaged all of them.

3.4 Adversarial Loss ($\mathcal{L}_{adv}$)

In this study, we modified the original adversarial loss [3] by adding a segmentation loss ($\mathcal{L}_{seg}$) for optimizing the generator to segment the DEM. Similar to the original paper [3], here the generator tries to minimize Eq. 6 while the discriminator tries to maximize it.

$$\mathbb{E}_{y \sim Y_{GAN}}\left[\log(D(y))\right] + \mathbb{E}_{x \sim X_{GAN}}\left[\log(1 - D(G(x))) + \mathcal{L}_{seg}(G(x))\right] \tag{6}$$

In the Eq. 6, G is the generator, D is the discriminator, $x \sim X_{GAN}$ is the set of input images, $y \sim Y_{GAN}$ is the combination of true DEM and true images

Table 2. Performance of U-Net and Probabilistic U-Net in Dice similarity coefficient (DSC) and volume point estimation (VPE). Note that higher DSC value is better (↑), lower MSE value is better (↓), and closer to 0 is better for Error (→ 0). The best result for each column is shown in bold and the second best is underlined. WCE stands for weighted cross entropy and while FL stands for focal loss.

Model	**Cost**	**DSC ↑**				**VPE**	
	Function	**Shrink**	**Grow**	**Stable**	**Average**	**Error → 0**	**MSE ↓**
U-Net	WCE	0.1794 (0.072)	0.1970 (0.097)	0.6413 (0.159)	0.3393 (0.078)	-2.7127 (10.31)	112.87 (247.44)
	FL	0.1757 (0.077)	**0.2073** **(0.104)**	0.6483 (0.156)	0.3438 (0.076)	-2.7002 (10.08)	108.17 (256.61)
Prob. U-Net (t0 & DEM as inputs to Posterior Net)	WCE	0.1491 (0.061)	0.1524 (0.090)	0.6220 (0.171)	0.3079 (0.086)	-2.5095 (9.84)	102.44 (234.61)
	FL	0.1673 (0.074)	0.1858 (0.089)	0.6147 (0.184)	0.3226 (0.090)	-2.0297 (9.27)	89.56 (220.73)
Prob. U-Net (t1 & DEM as inputs to Posterior Net)	WCE	0.1964 (0.071)	0.2040 (0.091)	**0.6564** **(0.162)**	0.3522 (0.080)	**-0.2953** **(8.33)**	69.05 (224.94)
	FL	**0.2092** **(0.082)**	0.2056 (0.092)	0.6507 (0.160)	**0.3552** **(0.080)**	-0.6650 (8.02)	**64.33** **(220.39)**

(i.e., T2-FLAIR for t0 and t1), $G(x)$ is the predicted DEM, $\mathbb{E}_y \sim Y_{GAN}$ is the expected value over Y_{GAN}, and $\mathbb{E}_x$ is the expected value over X_{GAN}. If G is U-Net then $X_{GAN} = X_{prior}$. Whereas, if G is probabilistic U-Net then $X_{GAN} = (X_{prior}, X_{post}, Y_{post})$. As in the previous section, X_{prior}, X_{post}, Y_{post} correspond to the T2-FLAIR for t0, t1, and true DEM respectively.

In this study, we also evaluate three different combinations of Y_{GAN} to investigate which produces the best result. The tested combinations are 1) only the true DEM (DEM GAN), 2) true DEM and T2-FLAIR normalised values at t0 (t0-DEM GAN), and 3) true DEM, T2-FLAIR normalised values at t0, and T2-FLAIR normalised values at t1 (t0-t1-DEM GAN). In these experiments, we used spectral normalization [7] for the discriminator network and trained it 5 times for each epoch.

4 Results

4.1 U-Net vs. Probabilistic U-Net

Table 2 shows the performances of U-Net and Probabilistic U-Net for predicting the spatial progression of WMHs (shown in Dice Similarity Coefficient (DSC)) and in volume point estimation (VPE). Following the original paper that proposed the Probabilistic U-Net [5], we first used T2-FLAIR at t0 and true DEM as inputs to Posterior Net. However, this approach was outperformed by the U-Net model. By, consequently, changing the input of Posterior Net to be T2-FLAIR at t1 and true DEM, the model using Probabilistic U-Net outperformed U-Net in our experiments. These show that the input data for the Posterior Net in the Probabilistic U-Net should differ from the input data for the other modules of this probabilistic architecture (i.e., U-Net and Prior Net). Table 2 also shows that the FL cost function produced better prediction results than the WCE in both DSC and VPE for all experimental settings.

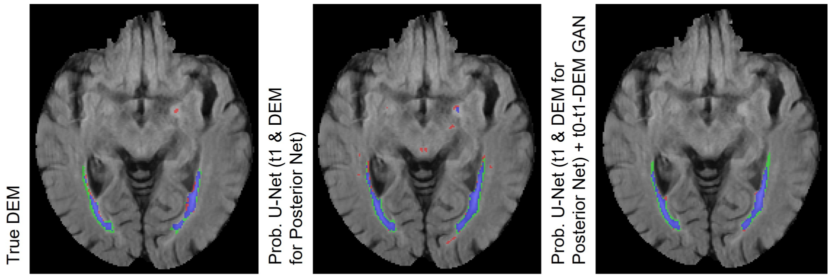
True DEM
Prob. U-Net (t1 & DEM for Posterior Net)
Prob. U-Net (t1 & DEM for Posterior Net) + t0-t1-DEM GAN

Fig. 5. Comparison of the true DEM (left) and predicted DEMs produced by using Probabilistic U-Net without adversarial training (middle) and Probabilistic U-Net with adversarial training with T2-FLAIR at t0 and true DEM (right).

Table 3. Performance of deep learning models trained with adversarial training for predicting the progression of WMHs in Dice similarity coefficient (DSC) and volume point estimation (VPE). Higher DSC value is better (↑), lower MSE value is better (↓), and closer to 0 is better for Error (→ 0). The best result for each column is shown in bold and the second best is underlined.

Model	**DSC ↑**				**VPE**	
	Shrink	**Grow**	**Stable**	**Average**	**Error → 0**	**MSE ↓**
Prob. U-Net **(t1 & DEM for Posterior Net)**	**0.2092** **(0.082)**	0.2056 (0.092)	0.6507 (0.160)	**0.3552** **(0.080)**	-0.6650 (8.02)	**64.33** **(220.39)**
Prob. U-Net (t1 & DEM for Posterior Net) + DEM GAN	0.1739 (0.083)	0.2083 (0.103)	0.6374 (0.172)	0.3399 (0.090)	2.0216 (9.32)	90.34 (180.32)
Prob. U-Net (t1 & DEM for Posterior Net) + t0-DEM GAN	0.1911 (0.093)	0.2184 (0.103)	**0.6530** **(0.163)**	0.3541 (0.089)	**0.3155** **(8.90)**	78.83 (156.17)
Prob. U-Net (t1 & DEM for Posterior Net) + t1-DEM GAN	0.1737 (0.083)	**0.2367** **(0.100)**	0.6427 (0.169)	0.3511 (0.086)	-3.4385 (8.97)	91.70 (205.29)
Prob. U-Net (t1 & DEM for Posterior Net) + t0-t1-DEM GAN	0.1701 (0.083)	0.2282 (0.102)	0.6425 (0.167)	0.3469 (0.083)	-3.3115 (8.83)	88.36 (220.39)
U-Net	0.1757 (0.077)	0.2073 (0.104)	0.6483 (0.156)	0.3438 (0.076)	-2.7002 (10.08)	108.17 (256.61)
U-Net **+ t0-DEM GAN**	0.1849 (0.091)	0.2134 (0.099)	0.6468 (0.159)	0.3484 (0.079)	-1.1187 (9.58)	92.44 (191.44)

4.2 Probabilistic U-Net with Adversarial Training

We investigated whether applying adversarial training with different input images can improve the performance of Probability U-Net. We evaluated these experiments using DSC, VPE, and the newly proposed VIE evaluation.

Table 3 shows that adversarial training with T2-FLAIR at t0 and true DEM slightly improved the prediction produced by Probabilistic U-Net in VPE (Error) and DSC (Stable). Fig. 5 also shows that the predicted DEM produced by adversarial training more closely followed the true DEM by removing the small false positive clusters in the prediction results. These experiments show that, while Probabilistic U-Net without adversarial training consistently produced some of the best prediction results in terms of DSC, the Probabilistic U-Net with adversarial training predicted more realistic DEM, closer to the true DEM, and

Table 4. Performance of deep learning models for predicting the future volume of WMHs evaluated in the newly proposed Volume Interval Estimation (VIE). The best result for each column is shown in bold and the second best is shown in underline. Symbol (↑) indicates that higher values are better while symbol (→ 0) indicates that values closer to 0 are better. *Abbreviations*: "CP" stands for "Correct Prediction", "CPinEVI" stands for "Correct Prediction in Estimated Volume Interval", "(CP + WP)inEVI" stands for "Correct Prediction + Wrong Prediction but still in EVI", "VPE" stands for Volume Point Estimation, "MaxVE" stands for Maximum Volume Estimation, and "MinVE" stands for Minimum Volume Estimation.

Model	**CP ↑**	**CPinEVI ↑**	**(CP+WP) inEVI ↑**	**Distance to VPE (in *ml*)**	
				MaxVE → 0	**MinVE → 0**
Prob. U-Net (t1 & DEM for Posterior Net)	**73.03**%	44.74%	51.32%	4.0862 (3.241)	-5.5700 (3.918)
Prob. U-Net (t1 & DEM for for Posterior Net) + DEM GAN	63.16%	30.26%	39.47%	2.5377 (3.0779)	-5.5978 (4.5046)
Prob. U-Net (t1 & DEM for for Posterior Net) + t0-DEM GAN	69.74%	39.47%	50.00%	2.6563 (3.0834)	-6.7103 (5.4319)
Prob. U-Net (t1 & DEM for for Posterior Net) + t1-DEM GAN	68.42%	44.74%	57.24%	2.8499 (2.7111)	-7.9550 (5.3201)
Prob. U-Net (t1 & DEM for for Posterior Net) + t0-t1-DEM GAN	**73.03**%	**48.68**%	57.89%	2.9383 (3.0793)	-7.6224 (5.5935)
U-Net	61.84%	36.84%	48.68%	2.9911 (3.3676)	-6.1355 (4.5706)
U-Net + t0-DEM GAN	72.37%	46.71%	**59.87**%	4.5915 (6.7208)	-6.2326 (4.7695)

with better VPE values. Additionally, U-Net with adversarial training produced better prediction results than the original U-Net without adversarial training.

Table 4 shows the performances of the deep learning models evaluated using VIE. The percentage of patients with correctly predicted DEM (i.e., subjects with shrinking and growing WMHs correctly predicted as having shrinking and growing WMHs respectively) is given by the metric called "CP" (Correctly Predicted). We also calculated the percentage of patients having their true future volumes of WMHs (Tt1V) correctly estimated and located between MinVE and MaxVE, and expresses it under a metric named "CPinEVI" (Correctly predicted in Estimated Volume Interval (EVI)). Based on the VIE categorization (Fig. 4 and Table 1), "CPinEVI" covers categories A, B+, and B-. Lastly, "(CP+WP)inEVI" shows the percentage of correctly and wrongly predicted patients with their Tt1V still located between MinVE and MaxVE. Based on Fig. 4 and Table 1, "(CP+WP)inEVI" covers categories A, B+, B-, D+, and D-.

Both "CPinEVI" and "(CP+WP)inEVI" are important for better interpretation and higher confidence in our predictive model. Metric "CPinEVI" is important not only in evaluation but also in real-word testing/inference. A predictive model with higher rate of "CPinEVI" in testing means that there is a high probability that the Tt1V lies between the predicted/estimated MinVE and MaxVE produced by the predictive model. On the other hand, "(CP+WP)inEVI" captures difficult cases where the future volume of WMHs is wrongly predicted by the predictive model but the Tt1V still lies between the predicted/estimated MinVE and MaxVE. These cases happen mostly when the WMHs volume change from t0 to t1 is very small. For example, a patient with WMHs volume of 5 *ml* at t0 and 5.5 *ml* at t1 (i.e., growing WMHs) is wrongly predicted by the model to have future WMHs volume of 4.5 *ml* (i.e., shrinkage in the total WMHs volume at t1) while having predicted MinVE and MaxVE of 4 *ml* and 6 *ml* respectively.

The results in Table 4, show that Probability U-Net with adversarial training using T2-FLAIR for t0, t1, and true DEM produced the best results in all metrics of VIE. While the rate of CP is the same with the Probabilistic U-Net without adversarial training, Probabilistic U-Net with adversarial training using T2-FLAIR for t0, t1, and true DEM produced better results than other probabilistic models in “CPinEVI” and “(CP+WP)inEVI” (48.68% and 57.89% respectively). It is worth to mention that the best result for “(CP+WP)inEVI” was produced by the U-Net with adversarial training using T2-FLAIR for t0 and true DEM (i.e., 59.87% respectively). However, as shown in Table 3, it did not outperform any Probabilistic U-Net settings in DSC and/or VPE.

Lastly, one can argue that higher rates of “CPinEVI” and “(CP+WP)inEVI” can be produced by expanding the VIE itself (i.e., smaller value of MinVE and larger value of MaxVE). However, as shown in Table 4, the predicted values of MinVE and MaxVE from different predictive models are relatively close to the predicted VPE in all settings (calculated by performing MinVE - VPE and MaxVE - VPE for the whole dataset).

5 Conclusion and Discussion

In this study, we propose the use of a probabilistic deep learning model (i.e., Probability U-Net) for capturing/modelling spatial uncertainty in the estimation of WMHs from brain MRI scans. The adversarial loss successfully improved the prediction results, ensuring the predicted DEM closely follows the global context of the true DEM by removing small clusters of false positives. Furthermore, we also propose a procedure to evaluate the predictive model called Volume Interval Estimation (VIE) for better evaluation, interpretation, and higher confidence in our predictive model. While the probability model with adversarial training produced some of the best results, VIE proved to be effective for interpreting and evaluating the predicted results. It is also worth to mention that there are still many useful evaluation metrics that can be derived from the VIE. Future works include incorporating VIE into the predictive model as a regularization term in the cost function. Preliminary results show an improvement in the prediction of WMHs evolution. Furthermore, to reduce aleatoric uncertainty, information from other MRI sequences (e.g. T1-weighted) and modalities (e.g. diffusion-weighted images) could be advantageous. Given the presence of WMHs in scans of older adults and dementia patients, re-training and testing the proposed schemes in a wider sample would be also beneficial.

Acknowledgements Funds from JSPS (Kakenhi Grant-in-Aid for Research Activity Start-up, Project No. 20K23356) (MFR); Row Fogo Charitable Trust

(Grant No. BRO-D.FID3668413) (MCVH); Wellcome Trust (patient recruitment, scanning, primary study Ref No. WT088134/Z/09/A); Fondation Leducq (Perivascular Spaces Transatlantic Network of Excellence); EU Horizon 2020 (SVDs@Target); and the MRC UK Dementia Research Institute at the University of Edinburgh (Wardlaw programme) are gratefully acknowledged. This research was also supported by the program for Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) from the Japan Agency for Medical Research and Development AMED (JP21dm0207001).

References

1. Castorina, L.V., et al.: Metrics for quality control of results from super-resolution machine-learning algorithms – data extracted from publications in the period 2017-may 2021 [dataset] (2021). https://doi.org/doi.org/10.7488/ds/3062
2. Chappell, F.M., et al.: 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 (2017). https://doi.org/10.1186/s13063-017-1825-7
3. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27. (2014)
4. Hüllermeier, E., et al.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning **110**(3), 457–506 (2021). https://doi.org/10.1007/s10994-021-05946-3
5. Kohl, S., et al.: A probabilistic u-net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems. vol. 31. (2018)
6. Lin, T.Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324
7. Miyato, T., et al.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations. (2018)
8. Rachmadi, M.F., et al.: Automatic spatial estimation of white matter hyperintensities evolution in brain MRI using disease evolution predictor deep neural networks. Medical Image Analysis **63**, 101712 (2020). https://doi.org/10.1016/j.media.2020.101712
9. Rachmadi, M.F., et al.: Predicting the evolution of white matter hyperintensities in brain MRI using generative adversarial networks and irregularity map. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 146–154. Springer (2019). https://doi.org/10.1007/978-3-030-32248-9_17
10. Ronneberger, O., et al.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
11. Valdés Hernández, M.d.C., et al.: 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 (2015). https://doi.org/10.1002/brb3.415
12. Wardlaw, J.M., et al.: White matter hyperintensity reduction and outcomes after minor stroke. Neurology **89**(10), 1003–1010 (2017). https://doi.org/10.1212/WNL.0000000000004328

Supplementary Materials

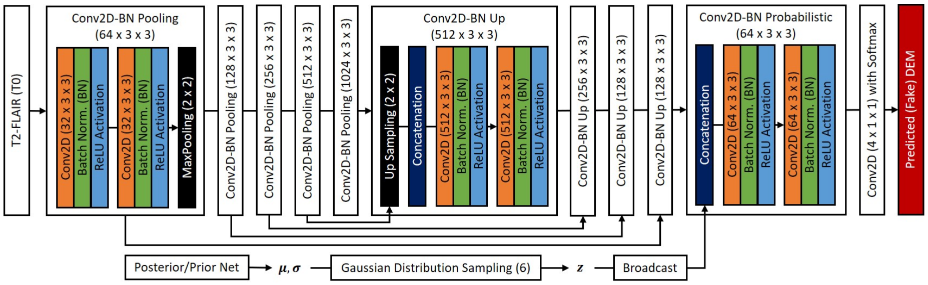

Fig. 6. Architecture of generator (i.e., U-Net) used in this study. Note that spectral normalization [7] is used in this study.

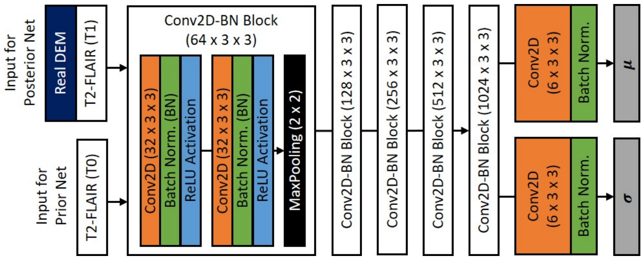

Fig. 7. Architecture of Posterior/Prior Net used in this study. Note that the networks produce mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$) that will be used to sample $\boldsymbol{z}$.

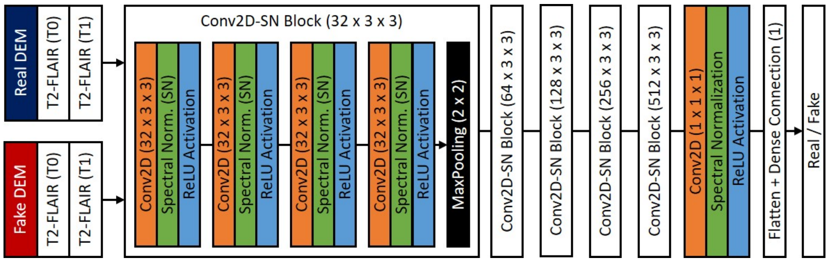

Fig. 8. Architecture of Discriminator used in this study. Note that spectral normalization [7] is used in this study.