Limited One-time Sampling Irregularity Map (LOTS-IM): Automatic Unsupervised Quantitative Assessment of White Matter Hyperintensities in Structural Brain Magnetic Resonance Images

Muhammad Febrian Rachmadi[a,b], Maria del C. Valdés-Hernández[b], Hongwei Li[c], Ricardo Guerrero[d], Rozanna Meijboom[b], Stewart Wiseman[b], Adam Waldman[b], Jianguo Zhang[c], Daniel Rueckert[d], Taku Komura[a]

[a]*School of Informatics, University of Edinburgh, Edinburgh, UK*
[b]*Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK*
[c]*Computing, School of Science and Engineering, University of Dundee, Dundee, UK*
[d]*Department of Computing, Imperial College London, London, UK*

Abstract

We present a complete study of limited one-time sampling irregularity map (LOTS-IM), a fully automatic unsupervised approach to extract brain tissue irregularities in magnetic resonance images (MRI), including its application and evaluation for quantitative assessment of white matter hyperintensities (WMH) of presumed vascular origin and assessing multiple sclerosis (MS) lesion progression. LOTS-IM is unique compared to similar other methods because it yields irregularity map (IM) which represents WMH as irregularity values, not probability values, and retains the original MRI's texture information. We tested and compared the usage of IM for WMH segmentation on T2-FLAIR MRI with various methods, including the well established unsupervised WMH segmentation Lesion Growth Algorithm from the public toolbox Lesion Segmentation Toolbox (LST-LGA), conventional supervised machine learning schemes and *state-of-the-art* supervised deep neural networks. In our experiments, LOTS-IM outperformed unsupervised method LST-LGA, both in performance and processing speed, thanks to the limited one-time sampling scheme and its implementation on GPU. Our method also outperformed supervised conventional machine learning algorithms (*i.e.*, support vector machine (SVM) and random forest (RF)) and deep neural networks algorithms (*i.e.*, deep Boltzmann machine (DBM) and convolutional encoder network (CEN)), while yielding comparable results to the convolutional neural network schemes that rank top of the algorithms developed up to date for this purpose (*i.e.*, UResNet and UNet). The high sensitivity of IM on depicting signal change deems suitable for assessing MS progression, although care must be taken with signal changes not reflective of a true pathology.

Keywords: white matter hyperintensities (WMH), unsupervised WMH segmentation, dementia, irregularity map, penumbra of WMH, characterisation of WMH

1. Introduction

White matter hyperintensities (WMH), which are commonly assessed in T2-Fluid Attenuation Inversion Recovery (FLAIR) brain MRI, have been identified as predictor of stroke (Rensma et al., 2018) and associated with cognitive decline (Pohjasvaara et al., 2000; del C. Valdés Hernández et al., 2013) and progression of dementia (Wardlaw et al., 2013). Because of their importance, there have been many studies on the development of approaches/methods for detecting and assessing WMH automatically. Most commonly, WMH voxels are identified and "segmented" from the other "normal" brain tissues. This can be done either with the help of manually generated labels of WMH (supervised learning) or without the help of these manual labels (unsupervised learning).

Since the widespread use of deep neural network algorithms in computer vision, these methods have become the *state-of-the-art* for WMH detection and segmentation. Deep neural networks such as DeepMedic (Kamnitsas et al., 2017), UNet

(Ronneberger et al., 2015) and UResNet (Guerrero et al., 2018) have outperformed conventional machine learning algorithms such as support vector machine (SVM) and random forest (RF) (Ithapu et al., 2014) on automatically segmenting WMH. However, as supervised methods, they are highly dependent on manual labels produced by experts (*i.e.*, physicians) for training. This dependency to expert's opinion limits their applicability due to the expensiveness of manually generating WMH labels and the limited number of them. Furthermore, the quality of manual labels itself depends on and varies according to the expert's skill. These intra/inter-observer inconsistencies usually are quantified and reported, but they do not solve the problem. On the other hand, the more recent unsupervised deep neural networks methods, which are based on generative adversarial networks (GAN) (Goodfellow et al., 2014), like anomaly GAN (AnoGAN) (Schlegl et al., 2017) and adversarial auto-encoder (AAE) (Chen and Konukoglu, 2018) need large number of both healthy and unhealthy data for adversarial training processes, which are usually not available or easily accessible.

Conventional unsupervised machine learning algorithms such as Lesion Growth Algorithm from Lesion Segmentation Tool toolbox (LST-LGA) (Schmidt et al., 2012) and Lesion-TOADS (Shiee et al., 2010), do not have the aforementioned dependencies to do WMH segmentation. Because of that, these methods have been tested in many studies and become the standard references to the other WMH segmentation methods. Unfortunately, their performance are usually worse than supervised methods (Ithapu et al., 2014; Rachmadi et al., 2017a).

Novel unsupervised methods named irregularity age map (IAM) (Rachmadi et al., 2017b) and its faster version one-time sampling IAM (OTS-IAM) (Rachmadi et al., 2018c) have been recently proposed and reported to work better than the *state-of-the-art* unsupervised WMH segmentation method LST-LGA. IAM and OTS-IAM not only outperform the LST-LGA but also SVM, RF, and some other deep neural network algorithms such as deep Boltzmann machine (DBM) (Salakhutdinov and Larochelle, 2010) and convolutional encoder network (CEN) (Brosch et al., 2016). Furthermore, IAM and OTS-IAM are unique as they produces irregularity map (IM) which has the advantages over deep neural networks' probability map of capturing and retaining changes of the original T2-FLAIR intensities, which cannot be done by deep neural network algorithms (see Figure 2 and Figure 4). The gradual changes of hyperintensities along the border of WMH, usually referred as the "penumbra" (Maillard et al., 2011), have been a subject in many studies in recent years, which debate criteria to correctly identify the WMH borders (Hernández et al., 2010; Jeerakathil et al., 2004; Firbank et al., 2003). Moreover, the penumbra of WMH itself is especially important for the study of WMH progression (Kapeller et al., 2003; Bendfeldt et al., 2009; Callisaya et al., 2013). Unfortunately, not even the deep neural network algorithms are sensitive enough to assess the degrees of WMH severity automatically, until the recently proposed methods of IAM and OTS-IAM. It is also worth to mention that the progression of WMH can be easily simulated using IM and has been proposed in our previous study (Rachmadi et al., 2018a).

While IAM and OTS-IAM have been tested in previous studies and produced exceptional results for unsupervised WMH segmentation, they had one main limitation: their lengthy computing time. The most recent OTS-IAM takes 13 minutes (on GPU) to 174 minutes (on CPU) to complete 1 volumetric MRI volume of 256×256×35 in average. The aforementioned computation times are not ideal especially if thousands of MRI are to be processed.

In this study, we proposed a new version of IAM namely limited one-time sample irregularity map (LOTS-IM) which is much faster than IAM and OTS-IAM. The fastest configuration of LOTS-IM completes the computation of one volume in 25 seconds without having too much quality degradation. This study also completes the development of irregularity map methods by evaluating all parameters involved on different scenarios, not yet done in previous studies, and analyses its applicability in a clinical study.

2. Irregularity Map Method

The "irregularity age map" (IAM) for WMH assessment on brain MRI was originally proposed in (Rachmadi et al., 2017b), and it is based on a computer graphics method (Bellini et al., 2016) developed to detect aged/weathered regions in texture images. The term "age value" and "age map" were originally used by Bellini et al. (2016) for the 2D array of values between 0 and 1 denoting the weathered regions considered texture irregularities in natural images. In this study, we refer, instead, to "irregularity value" and "irregularity map" as the

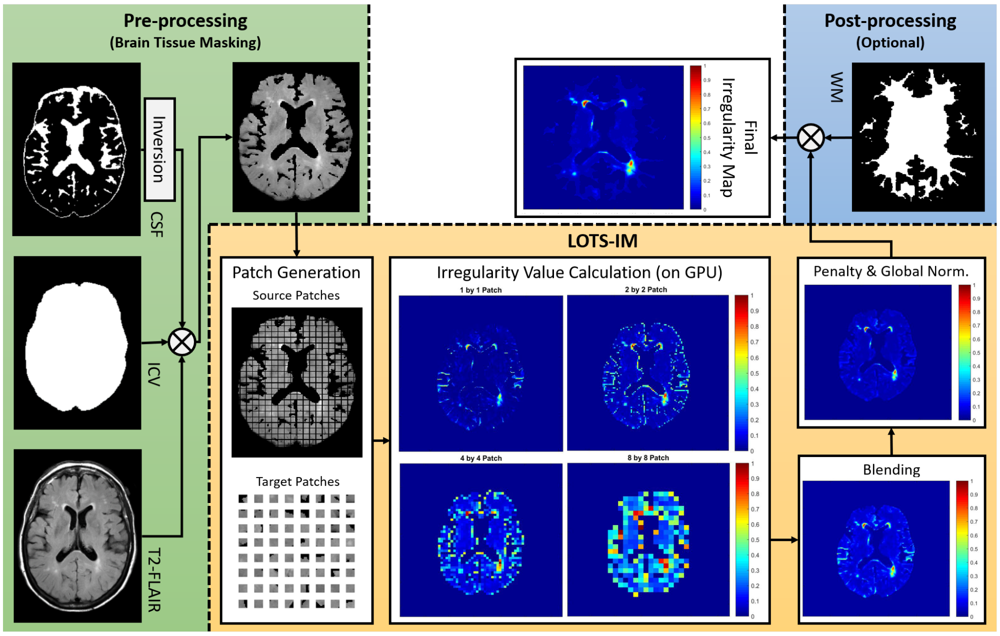

Figure 1: Flow of the proposed LOTS-IM. **1) Pre-processing**: brain tissue-only T2-FLAIR MRI 2D slices are generated from the original T2-FLAIR MRI and its corresponding brain masks (*i.e.*, intracranial volume (ICV) and cerebrospinal fluid combined with pial regions (CSF)). **2) LOTS-IM**: the brain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

concept of detecting "aged/weathered" textural regions no longer applies. In the irregularity map (IM), the closer the value to 1, the more probable a pixel/voxel belongs to a neighbourhood that has different texture from that considered "normal".

IM is calculated from structural brain MRI by applying the following steps: 1) preparation of the regions of interest where the algorithm will work (e.g. brain tissue), 2) patch generation, 3) irregularity value calculation and 4) final irregularity map generation. These four steps are visualised in Figure 1 and described in the rest of this section.

2.1. *Brain tissue masking*

For brain MRI scans, the brain tissue mask is necessary to exclude non-brain tissues which can represent "irregularities" *per se* (*e.g.* skull, cerebrospinal fluid, veins and meninges). In other words, we would like to compare brain tissue patches among themselves, not with skull or other parts of non-brain tissues. For this purpose we use two binary masks: intracranial volume (ICV) and cerebrospinal fluid (CSF) masks, the latter containing also pial elements like veins and meninges. In our experiments, the ICV mask was generated by using optiBET (Lutkenhoff et al., 2014). However, several tools that produce accurate output exist and can be used for this purpose (*e.g.* bricBET[1], freesurfer[2]). The CSF mask was generated by using a multispectral algorithm (Valdés Hernández et al., 2015).

The pre-processing step before computing LOTS-IM only involves the generation of these two masks as per in the original IAM and in OTS-IAM (Rachmadi et al., 2017b, 2018b). Their combination generates the brain tissue mask, by which the FLAIR volume is multiplied. This study also uses the normal appearing white matter (NAWM) mask, as per OTS-IAM (Rachmadi et al., 2018b), to exclude brain areas in the cortex that could be identified as false positives. NAWM masks were generated using the FSL-FLIRT tool (Jenkinson et al., 2002), but can also be generated using in-house tools or freesurfer, for example.

2.2. *Patch generation*

Patch generation generates two sets of patches; non-overlapping grid patches called *source patches*

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

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

and randomly-sampled patches called *target patches*, which can geometrically overlap each other (see Figure 1). In the IM computation, a source patch is used as a reference to the underlying pixel/patch while a target patch is used to represent the distribution of the image textures where a set of target patches is randomly sampled from the same image.

Source and target patches are used to calculate the irregularity value, where each of the source patches is compared with several randomly sampled target patches using a distance function (Bellini et al., 2016). This will be discussed in the next subsection. We use hierarchical subsets of four different sizes of source and target patches, which are 1×1, 2×2, 4×4 and 8×8. Thus, source and target patches are defined as 2D arrays.

2.3. Irregularity value calculation

Irregularity value calculation is the core computation of the IM where a value called *irregularity value* is computed. Let $\mathbf{s}$ be a source patch and $\mathbf{t}$ be a target patch, then the distance (d) between $\mathbf{s}$ and $\mathbf{t}$ is:

$$d = \text{average}(|\max(\mathbf{s} - \mathbf{t})|\,,\ |\text{mean}(\mathbf{s} - \mathbf{t})|) \qquad (1)$$

Based on Equation 1 above, the distance between source patch ($\mathbf{s}$) and target patch ($\mathbf{t}$) can be calculated by averaging the maximum difference and the mean difference between $\mathbf{s}$ and $\mathbf{t}$. The difference between $\mathbf{s}$ and $\mathbf{t}$ itself is calculated by subtracting their intensities. The averaging of maximum and mean differences is applied to make the distance value robust against outliers. To capture the distribution of textures in an image (*i.e.*, slice MRI), each source patch will be compared against a set of target patches (*e.g.* 2,048 target patches in (Rachmadi et al., 2018c)) in which the same number of distance values are produced.

The *irregularity value* for a source patch can be calculated by sorting all distance values and averaging the 100 largest distance values of the whole set. The rationale is simple: the mean of the 100 largest distance values produced by an irregular source patch is still comparably higher than the one produced by a normal source patch. Also, mean is chosen as we are comparing irregularities with respect to the normal-appearing white matter, and normal tissue intensities are known to be normally distributed. On the other hand, other statistical characteristic, such as percentiles, has been identified and used to discern degree of pathology (Dickie et al., 2015, 2014).

All irregularity values from all source patches are then mapped and normalised to real values between 0 to 1 to create the *irregularity map for one MRI slice (see Figure 1).*

2.4. Final irregularity map generation

The generation of the final irregularity map consists of three sub-steps, which are a) *blending four irregularity maps produced in irregularity value calculation*, b) *penalty* and c) *global normalisation*. *Blending of four irregularity maps* is performed by using the following formulation:

$$\text{IM}_{blended} = \alpha \cdot \text{IM}_1 + \beta \cdot \text{IM}_2 + \gamma \cdot \text{IM}_4 + \delta \cdot \text{IM}_8 \quad (2)$$

where $\alpha + \beta + \gamma + \delta$ is equal to 1 and IM_1, IM_2, IM_4 and IM_8 are irregularity maps from 1×1, 2×2, 4×4 and 8×8 source/target patches. Before the blending, irregularity maps resulting from different size of source/target patches are up-sampled to fit the original size of the MRI slice and then smoothed by using a Gaussian filter. The blended irregularity map is then *penalised* using the formulation below:

$$p = b \times o \qquad (3)$$

where b is the voxel from the blended irregularity map, o is voxel from the original MRI and p is the penalised voxel. In other words, the penalty involves element (*i.e.*, intensity) multiplication between the blended irregularity map and the original T2-FLAIR. Lastly, all irregularity maps from different MRI slices are normalised together to produce values between 0 to 1 for each voxel to describe "irregularity" with respect to the normal brain tissue across all slices. We name this normalisation procedure as *global normalisation*. Visualisation of irregularity value calculation, blending, penalty and global normalisation are shown in Fig. 1.

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

3. Limited one-time sampling irregularity map (LOTS-IM)

While the original IAM has been reported to work well for WMH segmentation, its computation takes considerable time because it performs one target patch sampling for each source patch, selecting different target patches per source patch. For clarity, we named this scheme as *multiple-time sampling* (MTS) scheme. MTS scheme is performed in the original IAM to satisfy the condition, stated in the original study, (Bellini et al., 2016) that target patches should not be too close to the source patch (*i.e.*, location based condition). Extra time in MTS to do sampling for each source patch is unavoidable.

To accelerate the overall IAM's computation, we proposed one-time sampling (OTS) scheme for IAM where target patches are randomly sampled only once for each MRI slice, hence abandoning the location based condition of the MTS (Rachmadi et al., 2018c). In other words, age values of all source patches from one slice are computed against one (i.e. the same) set of target patches. We named this combination of OTS scheme and IAM one-time sampling IAM (OTS-IAM).

In this study, we propose to limit the number of target patches to accelerate the overall computation, named limited one-time sampling IM (LOTS-IM). The original IAM, which run on CPUs, use an undefined large random number of target patches which could range from 10% to 75% of all possible target patches, depending on the size of the brain tissue in an MRI slice.

Six numbers of target patches are sampled and evaluated for LOTS-IM's computation, which are 2048, 1024, 512, 256, 128 and 64. We also propose a more systematic way to calculate irregularity value where the 1/8 largest distance values are used instead of a fixed number of 100. Thus, the 256, 128, 64, 32, 16 and 8 largest distance values are used to calculate irregularity values for 2048, 1024, 512, 256, 128 and 64 target patches respectively.

Smaller number of target patches in the LOTS-IM scheme enables us to implement it on GPU to accelerate the computation even more. The limited number of samples in power-of-two is carefully chosen to ease GPU implementation, especially for GPU memory allocation.

4. MRI Data, Other Machine Learning Algorithms, and Experiment Setup

For evaluating our scheme, we used a set of 60 T2-Fluid Attenuation Inversion Recovery (T2-FLAIR) MRI data from 20 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) database[3]. Each subject had three brain MRI scans obtained in three consecutive years. They were selected randomly and blind to any clinical, imaging or demographic information. All T2-FLAIR MRI sequences have the same dimension of 256 × 256 × 35. Full data acquisition information can be looked at in our previous study (Rachmadi et al., 2018b). Ground truth was produced semi-automatically by an expert in medical image analysis using the region-growing algorithm in the Object Extractor tool in Analyze^{TM} software guided by the co-registered T1- and T2-weighted sequences. For more details on this dataset, please see (Rachmadi et al., 2017a) and data-share page[4] to access the dataset. The ADNI, which hosts the database from which our dataset was extracted, was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessments can be combined to assess the progression of mild cognitive impairment and early Alzheimers disease. As such, the investigators within the ADNI[5] contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report.

We applied our method to longitudinal brain MRI data from 10 treatment-free multiple sclerosis (MS) patients, participants in the Future MS study: a multicentre study of this disease (https://future-ms.org/). Images were acquired on a Verio 3T MRI scanner with a 20-channels head coil. Two experts visually assessed the images and identified the new lesions, enlarged lesions, and rated the disease progression in none, low, moderate or high. We compared the LOTS-IM output with the expert assessments and explored the applicability of this approach to studies of MS.

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

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

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

We also compared the performances of LOTS-IM in segmenting WMH with other machine learning algorithms that have been previously tested and are commonly used for WMH segmentation; namely the original IAM, one-time sampling IAM (OTS-IAM), lesion growth algorithm from the Lesion Segmentation Tool (LST-LGA), Support Vector Machine (SVM), Random Forest (RF), Deep Boltzmann Machine (DBM), Convolutional Encoder Network (CEN), patch-based 2D CNN with global spatial information (DeepMedic-GSI-2D), UResNet and patch-based UNet.

LST-LGA (Schmidt et al., 2012) is still the *state-of-the-art* unsupervised WMH segmentation method. SVM and RF are conventional machine learning algorithms that have been used for WMH segmentation in many studies. DBM, CNN, DeepMedic, and U-Net based methods are supervised deep learning algorithms applied in recent years for WMH segmentation and reported high levels of accuracy. All supervised segmentation methods used in this study were trained and tested using 5-fold cross validation. Whereas, the performance of unsupervised methods was directly evaluated on all 60 labelled MRI data.

For clarity, we do not further elaborate on the implementation of the aforementioned algorithms. All of them were implemented as per previous studies: configurations for LST-LGA, SVM, RF, DBM and CEN algorithms are described in detail in (Rachmadi et al., 2017a), whereas configurations and parameters for DeepMedic-GSI-2D, UResNet and patch-based UNet can be found in (Rachmadi et al., 2018b), (Guerrero et al., 2018) and (Li et al., 2018) respectively.

Dice similarity coefficient (DSC) (Dice, 1945), which measures similarity between ground truth and automatic segmentation results, is used here as the primary metric of evaluation. Higher DSC score means better performance, and the DSC score itself can be computed as follow:

$$DSC = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{4}$$

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

Additional metrics of positive predictive value (PPV), specificity (SPC) and true positive rate (TPR) are also calculated. Non-parametric Spearman's correlation coefficient (Myers et al., 2010) is used to compute correlation between WMH volume produced by each segmentation method and visual ratings of WMH. Visual ratings of WMH are commonly used in clinical studies to describe and analyse severity of white matter disease (Scheltens et al., 1993). Correlation between visual ratings and volume of WMH is known to be high (Hernández et al., 2013). In this study, Fazekas's (Fazekas et al., 1987) and Longstreth's visual rating scales (Longstreth et al., 1996) are used for evaluation of each automatic method, as per (Rachmadi et al., 2017a).

5. Experiments and Results

5.1. Overview

Table 1 shows the overall results of the performance of all methods evaluated. Please note that the original IAM is listed as IAM-CPU.

From Table 1, we can see that the binary WMH segmentations produced by all IAM configurations (*i.e.*, IAM, OTS-IAM and LOTS-IM methods) outperformed LST-LGA in mean DSC, PPV, SPC and TPR metrics. The of the output of each configuration, for calculating the metrics' values listed in Table 1 was achieved using the optimum threshold value for each algorithm. IAM/OTS-IAM/LOTS-IM not only outperformed LST-LGA in these metrics but also some other supervised machine learning algorithms (*i.e.*, SVM and RF). Also, some LOTS-IM implementations outperformed supervised deep learning methods of DBM and CEN in DSC metric. The best value of each evaluation metric is highlighted in bold fonts.

Visual appearance of the irregularity map from LOTS-IM and probability maps produced by other segmentation methods such as DeepMedic-GSI-2D, and UNet/UResNet can be observed and compared in Figure 4. Figure 4 (top) shows that the irregularity map produced by LOTS-IM retains more texture information than the probability map produced by other segmentation methods, and it can be used for WMH segmentation by thresholding it, as shown in Figure 4 (bottom). Visualisation of the irregularity map on large WMH can also be seen in Figure 2. Note how the penumbra of WMH is well represented in the irregularity map in Figures 2 and 4, whereas, the probability maps produced by DeepMedic-GSI-2D and UNet/UResNet lack the ability to represent/output this information.

The performance of LOTS-IM and other methods on doing WMH segmentation per cutting off (threshold) values, can be appreciated in 3, where

Table 1: Algorithm's information and experiment results based on the Dice similarity coefficient (DSC), positive predictive value (PPV), specificity (SPC) and true positive rate (TPR) for each algorithm evaluated (the best value is written in **bold**). **Explanation of abbreviations**: "SPV/UNSPV" for supervised/unsupervised, "Deep Net." for deep neural networks algorithm, "Y/N" for Yes/No, "T2F/T1W" for T2-FLAIR/T1-weighted, "#MTPS" for maximum number of target patches, "#meTPS" for number of target patches used for calculating mean of age value, "TRSH" for optimum threshold and "Training/Testing" for training/testing time. Given "speed increase" is relative to IAM-CPU.

No	Method	SPV/ UNSPV	Deep Net.	Input Modality	#MTPS/ #meTPS	TRSH	DSC		PPV (Mean)	SPC (Mean)	TPR (Mean)	Training (min)	Testing (min)	Speed increase
							Mean	Std						
1	LST-LGA	UNSPV	N	T2F-T1W	-	0.134	0.3037	0.1658	0.3158	0.9946	0.3625	-	0.67	-
2	SVM	SPV	N	T2F-T1W	-	0.925	0.2630	0.1498	0.0474	0.9869	0.1259	26	1.38	-
3	RF	SPV	N	T2F-T1W	-	0.995	0.3633	0.1843	0.0482	0.9860	0.1320	37	0.68	-
4	DBM	SPV	Y	T2F	-	0.687	0.3235	0.1345	0.0642	0.9955	0.0542	1,341	0.28	-
5	CEN	SPV	Y	T2F	-	0.284	0.4308	0.1582	0.5255	0.9975	0.4815	152	0.08	-
6	Patch-UResNet	SPV	Y	T2F	-	0.200	**0.5277**	**0.1729**	0.5899	0.9970	**0.5968**	215	0.08	-
7	Patch-UNet	SPV	Y	T2F	-	0.200	0.5030	0.1487	0.6480	0.9985	0.4886	211	0.08	-
8	DeepMedic-GSI-2D	SPV	Y	T2F	-	0.801	0.5225	0.1690	0.5950	0.9985	0.5276	392	0.45	-
9	IAM (CPU)	UNSPV	N	T2F	75%/100	0.179	0.3930	0.1732	**0.7001**	**0.9993**	0.3757	-	217.18	-
10	OTS-IAM-CPU	UNSPV	N	T2F	75%/100	0.164	0.4297	0.1734	0.6994	0.9992	0.3827	-	173.50	1.26
11	LOTS-IM-2048s256m	UNSPV	N	T2F	2,048/256	0.178	0.4710	0.1816	0.6111	0.9984	0.4564	-	12.43	17.52
12	LOTS-IM-1024s128m	UNSPV	N	T2F	1,024/128	0.178	0.4721	0.1830	0.6082	0.9983	0.4607	-	3.82	56.85
13	LOTS-IM-512s64m	UNSPV	N	T2F	512/64	0.178	0.4729	0.1852	0.5918	0.9980	0.4710	-	1.87	116.14
14	LOTS-IM-256s32m	UNSPV	N	T2F	256/32	0.178	0.4711	0.1878	0.5722	0.9977	0.4865	-	0.77	282.05
15	LOTS-IM-128s16m	UNSPV	N	T2F	128/16	0.178	0.4660	0.1918	0.5357	0.9970	0.5158	-	0.45	482.62
16	LOTS-IM-64s8m	UNSPV	N	T2F	64/8	0.178	0.4539	0.2035	0.4769	0.9952	0.5589	-	0.42	517.10

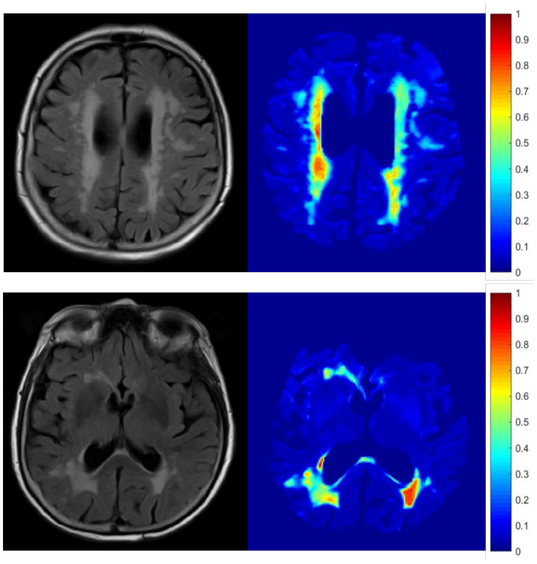

Figure 2: Large WMH visualised using irregularity map produced by the proposed method LOTS-IM. Note how the penumbra of WMH are represented by irregularity values.

the DSC performance curves for threshold value in each algorithm are graphed. LOTS-IM uses lower threshold values than the other methods to produce better WMH segmentation as irregularity map gives finer details of WMH penumbra than the other methods (see Figure 4 and additional Figure 2).

5.2. *IAM vs. OTS-IAM vs. LOTS-IM*

One-time sampling (OTS) and limited one-time sampling (LOTS) not only accelerated the computation time but also improved the overall performance, as shown in Table 1. Implementation of LOTS-IM on GPU increased the processing speed

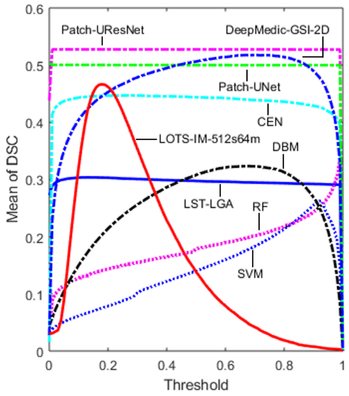

Figure 3: Mean of dice similarity coefficient (DSC) score for LST-LGA, SVM, RF, DBM, CEN, Patch-UResNet, Patch-UNet, DeepMedic-GSI-2D and LOTS-IM-512s64m in respect to all possible threshold values.

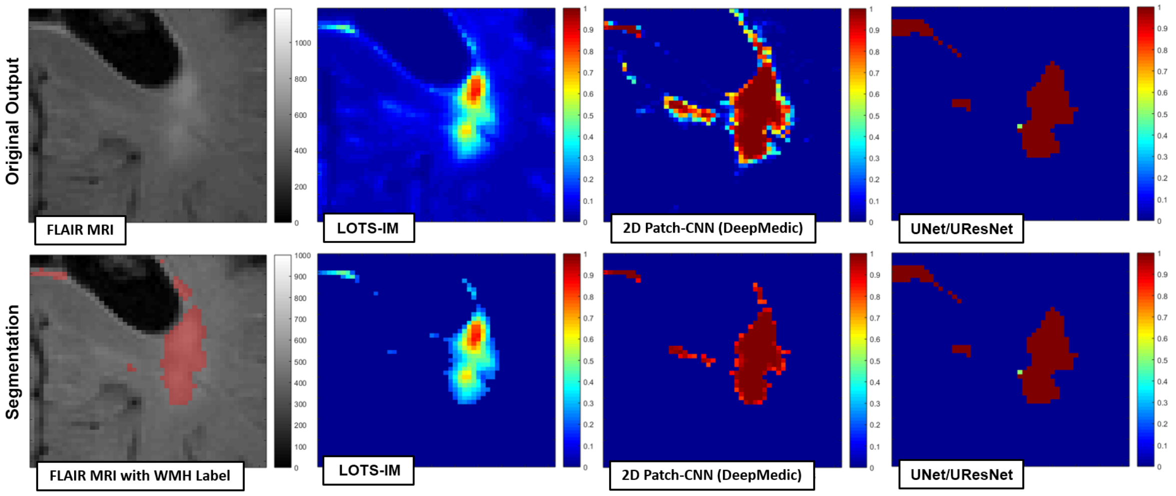

Figure 4: **Top:** Visualisation of original outputs produced by LOTS-IM (*i.e.*, irregularity map) and other machine learning methods such as CNN, UNet and UResNet (*i.e.*, probability maps). **Bottom:** Visualisation of WMH segmentation by cutting off the original values of irregularity/probability map. This figure shows that irregularity map not only nicely represents the penumbra of WMH by retaining the original textures but also is able to segment WMH by cutting off its values.

by 17 to 435 times with respect to the original IAM which was implemented on CPU. However, it is worth stressing that this increase in processing speed was not only due to the use of GPU instead of CPU, but also due to the limited number of target patch samples used in the the computation that LOTS-IM uses compared to the previous approaches. One of the GPU implementations of LOTS-IM (*i.e.*, LOTS-IM-64s8m) ran faster than LST-LGA. Note that the testing time listed in Table 1 excludes registrations and the generation of other brain masks used either in pre-processing or post-processing steps. The increase in speed achieved by the GPU implementation of LOTS-IM shows the effectiveness of the proposed method in terms of computation time and overall performance.

5.3. *Evaluation of Speed vs. Quality in LOTS-IM*

The biggest achievement of this work is the increase in processing speed achieved by the implementation of LOTS-IM on GPU, compared to the original IAM and OTS-IAM. The first iteration of IAM can only be run on CPU because it uses multiple-time sampling (MTS). OTS-IAM samples patches only once, but still uses a high number of target patches to compute the age map. In this study, we show that using a limited number of target patches leads not only to faster computation but also small to none quality degradation.

The relation between speed and quality of the output (mean DSC) produced by IAM, OTS-IAM and all configurations of LOTS-IM is illustrated in Figure 5. Note that Figure 5 is extracted from Table 1. Also, it is worth mentioning that more target patches used in LOTS-IM produced better PPV and SPC evaluation metrics than LOTS-IM using less target patches. This case is then reversed in TPR metric where using less target patches is better than using more target patches.

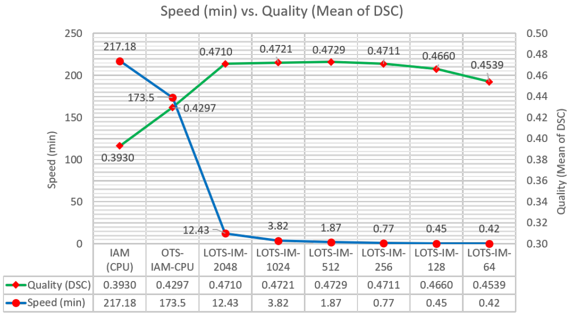

Figure 5: Speed (min) versus quality (mean of DSC) of different settings of LOTS-IM (extracted from Table 1). By implementing LOTS-IM on GPU and limiting number of target patch samples, computation time and result's quality are successfully improved and retained.

5.4. *Analysis of IM's Blending Weights*

As previously discussed, the only parameters that LOTS-IM has are four weights used to blend four

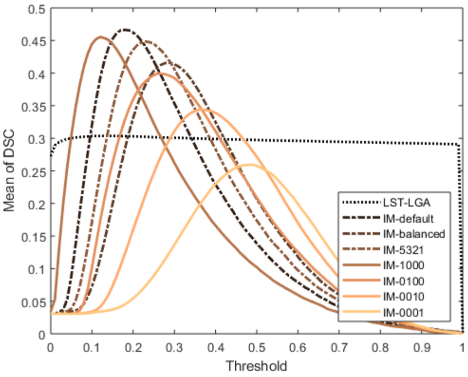

Figure 6: Curves of mean dice similarity coefficient (DSC) produced by using different settings of weights for blending different age maps. LOTS-IM used in this experiment is LOTS-IM-512s64m, and all weights are listed in Table 2.

irregularity maps hierarchically produced to generate the final irregularity map (see Equation 2 in Section 2.4). In this experiment, different sets of blending weights in LOTS-IM's computation were evaluated. This experiment aimed to identify which irregularity map produced by four different sizes of source/target patch influences more the performance of the algorithm and the quality of WMH segmentation.

We tested the 7 different sets of blending weights listed in Table 2. The first 3 sets blend all four irregularity maps while the other 4 only use one of the irregularity maps. The effect of different sets of blending weights is illustrated in Figure 6.

From Figure 6 and Table 2, we can see that blending irregularity values from different irregularity maps produced better WMH segmentation results than the others. Also, the irregularity map produced by 1 × 1 source/target patch influences the WMH segmentation result more than the others. Thus, the skewed blending weights of 0.65, 0.20, 0.10 and 0.05 produced the best DSC score. As this combination produced the best DSC score in this experiment, we made this set the default set for computing LOTS-IM. This set of blending weights has been used in previous IAM publications throughout the development of this approach (Rachmadi et al., 2017b).

As combining irregularity maps of different patch sizes is the best option, it is necessary to consider not only the intensity of the individual pixels but also the local distribution of the pixel intensities. Also, combining irregularity maps produced by different sizes of non-overlapping sources is similar to using information of an irregularity map produced by overlapping source patches, which means using more texture information. However, this experiment shows that the individual pixel intensity is the strongest feature for irregularity detection.

Table 2: Mean and standard deviation of DSC produced by using different settings of weights for blending different age maps. Plots corresponding to settings listed in this table can be seen in Figure 6. LOTS-IM tested in this experiment is LOTS-IM-512s64m.

Name	**Blending Weights**				**TRSH**	**DSC**	
	1x1	**2x2**	**4x4**	**8x8**		**mean**	**std**
LST-LGA	-	-	-	-	0.134	0.2936	0.1658
IM-default	0.65	0.20	0.10	0.05	0.179	0.4664	0.1820
IM-balanced	0.25	0.25	0.25	0.25	0.287	0.4158	0.1754
IM-5321	0.40	0.30	0.20	0.10	0.228	0.4486	0.1776
IM-1000	1	0	0	0	0.128	0.4555	0.1774
IM-0100	0	1	0	0	0.267	0.3995	0.1646
IM-0010	0	0	1	0	0.376	0.3439	0.1627
IM-0001	0	0	0	1	0.495	0.2594	0.1289

5.5. WMH Burden Scalability Test

In this experiment, all methods were tested and evaluated to see their performances on segmenting WMH in images with different burden of WMH. The DSC metric is still used, but the dataset is grouped into three different groups according to each patient's WMH burden. The groups are listed in Table 3 while the results can be seen in Figure 7 and Table 4. Please note that LOTS-IM is represented by LOTS-IM-512s64m as the best performer amongst the LOTS-IM methods (see Table 1).

Table 3: Three groups of MRI data based on WMH volume.

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

From Figure 7, it can be appreciated that LOTS-IM-512s64m performed better than LST-LGA in all groups. LOTS-IM-512s64m also performed better than the conventional supervised machine learning algorithms (i.e. SVM and RF) in 'Small' and 'Medium' groups. Whereas, LOTS-IM-512s64m's performance was at the level, if not better, than the supervised deep neural networks algorithms DBM and CEN. However, LOTS-IM-512s64m still could not beat the *state-of-the-art* supervised deep neural networks in any group.

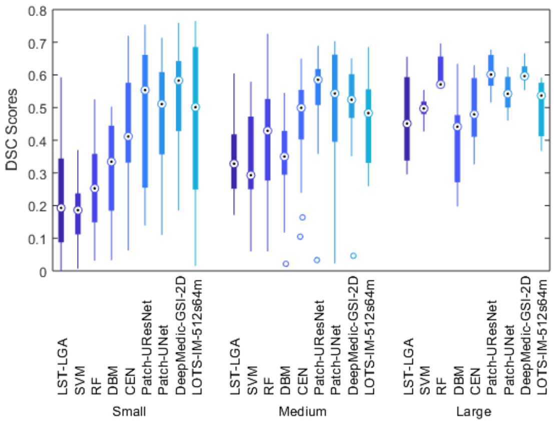

Figure 7: Distributions of dice similarity coefficient (DSC) scores for all methods tested in this study in respect to WMH burden of each patient (see Table 3).

To make this observation clearer, Table 4 lists the mean and standard deviation values that correspond to the box-plot shown in Figure 7. From both Figure 7 and Table 4 it can be observed that the standard deviation of LOTS-IM's performances in 'Small' WMH burden is still high compared to the other methods evaluated. However, LOTS-IM's performance is more stable in 'Medium' and 'Large' WMH burdens.

Table 4: Mean and standard deviation values of dice similarity coefficient (DSC) score's distribution for all methods tested in this study in respect to WMH burden of each patient (see Table 3). Note that LOTS-IM-512s64m is listed as LIM-512s64m in this table.

Method	TRSH	DSC - Small		DSC - Medium		DSC - Large	
		Mean	Std	Mean	Std	Mean	Std
LST-LGA	0.138	0.2335	0.1785	0.3524	0.1208	0.4645	0.1399
SVM	0.925	0.1792	0.0958	0.3360	0.1284	0.4966	0.0377
RF	0.995	0.2512	0.1298	0.4150	0.1662	0.6055	0.0559
DBM	0.687	0.3127	0.1432	0.3442	0.1350	0.4014	0.1474
CEN	0.284	0.4359	0.1802	0.4474	0.1485	0.4896	0.1122
Patch-UResNet	0.200	0.5007	0.2064	0.5403	0.1432	0.6064	0.0579
Patch-UNet	0.200	0.4872	0.1596	0.5079	0.1697	0.5447	0.0574
2D Patch-CNN	0.801	0.5230	0.1722	0.5118	0.1340	0.6053	0.0341
LIM-512s64m	0.179	0.4682	0.2278	0.4660	0.1331	0.4940	0.0932

5.6. Analysis of LOTS-IM's Random Sampling Scheme

To automatically detect FLAIR's irregular textures (*i.e.*, WMH) without any expert supervision, LOTS-IM works on the assumption that normal brain tissue is predominant compared with the extent of abnormalities. Due to this assumption, random sampling is used in the computation of LOTS-IM to choose the target patches. However, it raises an important question on the stability of LOTS-IM's performance to produce the same level of results for one exact MRI data, especially using different number of target patches.

In the first experiment, we randomly chose one MRI data out of the 60 MRI data that we have and ran LOTS-IM 10 times using different number of target patches. Each result was then compared to the ground truth and listed in Table 5. From this experiment, we can see that each setting produced low standard deviation values which indicates that the results are closely distributed around the corresponding mean values. However, there is an indication that higher deviations are produced when using lower number of target patches.

Table 5: Distribution metrics (mean and standard deviation) based on DSC for each LOTS-IM's settings. Each LOTS-IM setting is tested on a random MRI data 10 times.

No	Method	TRSH	DSC	
			mean	std
1	LOTS-IM-2048s128m	0.178	0.5681	0.0041
2	LOTS-IM-1024s128m	0.178	0.5901	0.0018
3	LOTS-IM-512s64m	0.178	0.5922	0.0033
4	LOTS-IM-256s32m	0.178	0.5925	0.0075
5	LOTS-IM-128s32m	0.178	0.5848	0.0092
6	LOTS-IM-64s16m	0.178	0.5852	0.0141

In the second experiment, we chose three random MRI data from each group of WMH burden (*i.e.*, 'Small', 'Medium' and 'Large') based on Table 3, ran LOTS-IM-512s64m 10 times, and compared the results with the ground truth. The results are listed in Table 6. Similar to the first experiment, low standard deviation values were produced for each subject, regardless of the WMH burden.

Table 6: Distribution metrics (mean and standard deviation) based on DSC for subject with different WMH burden. Each subject is tested 10 times using LOTS-IM-512s64m.

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

The two experiments done for this analysis indicates that LOTS-IM produces stable result of WMH segmentation in multiple test instances re-

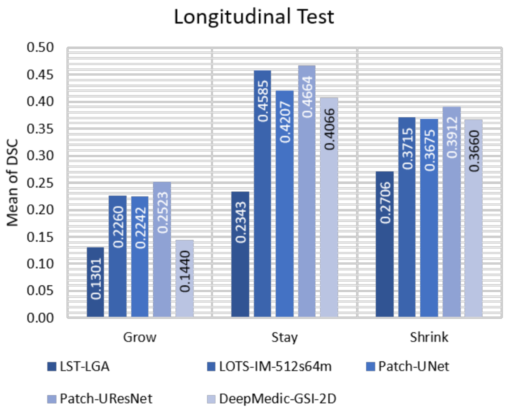

Figure 8: Quality of spatial agreement (Mean of DSC) of the produced results in longitudinal test. Longitudinal test is done to see the performance of tested methods in longitudinal dataset of MRI (see Table 7 for full report).

gardless of WMH burden while employing a simple random sampling scheme. However, of course, more sophisticated sampling method could be applied to make sure patches of normal brain tissue are more likely to be sampled.

5.7. *Longitudinal Test*

In this experiment, we evaluated spatial agreement between the produced results in three consecutive years. For each subject, we aligned Year-2 (Y2) and Year-3 (Y3) MRI and derived data to the Year-1 (Y1) using niftyReg through TractoR (Clayden et al., 2011), subtracted the aligned WMH labels of the baseline/previous year from the follow-up year(s) (*i.e.*, Y2-Y1, Y3-Y2, and Y3-Y1), and then labelled each voxel as 'Grow' if it has value above zero after subtraction, with 'Shrink' if it has value below zero after subtraction, and with 'Stay' if it has value of zero after subtraction and one before subtraction. This way, we can see whether the method captures the progression of WMH across time.

Figure 8 summarises the results listed in Table 7 for all methods (*i.e.*, LST-LGA, LOTS-IM-512s64m, Patch-UNet, Patch-UResNet and DeepMedic-GSI-2D). We can see that LOTS-IM-512s64m outperforms LST-LGA and competes with deep neural networks methods of Patch-UNet, Patch-UResNet and DeepMedic-GSI-2D, where LOTS-IM-512s64m is the second best performer after Patch-UResNet in this longitudinal evaluation. This, again, confirms that the LOTS-IM shows comparable performance with the *state-of-the-art* deep learning methods.

Table 7: Mean and standard deviation values produced in longitudinal tes (see Table 3). LOTS-IM-GPU-512s64m is listed as LIM-512s64m in this table.

Method	Dice Similarity Coefficient (DSC)					
	Grow		Stay		Shrink	
	Mean	Std	Mean	Std	Mean	Std
LST-LGA	0.1301	0.0350	0.2343	0.0199	0.2706	0.0058
LIM-512s64m	0.2260	0.0084	0.4585	0.0104	0.3715	0.0018
Patch-UNet	0.2242	0.0125	0.4207	0.0125	0.3675	0.0242
Patch-UResNet	**0.2523**	0.0199	**0.4664**	0.0211	**0.3912**	0.0044
2D Patch-CNN	0.1440	0.0228	0.4066	0.0298	0.3660	0.0129

5.8. *Correlation with Visual Scores*

In this experiment, we want to see how close IAM's results correlate with visual rating scores of WMH, specifically Fazekas (Fazekas et al., 1987) and Longstreth's visual scores (Longstreth et al., 1996).

The correlation was calculated by using Spearman's correlation. The correlation coefficients calculated were: 1) between the total Fazekas score (*i.e.*, the sum of periventricular and deep white matter hyperintensities) and manual/automatic WMH volumes and 2) between Longstreth's total score and manual/automatic WMH volumes. The results are listed in Table 8.

Table 8 shows that, although not much better, all LOTS-IM methods are highly correlated with visual rating clinical scores. It is also worth to mention that while LST-LGA's results are highly correlated with visual ratings, it still has one of the lowest DSC metric (see Table 1). LOTS-IM has high values of DSC and correlation with visual scores at the same time.

Table 8: Non-parametric correlation using Spearman's correlation coefficient between WMH volume and Fazekas and Longstreth visual ratings.

Visual Rating		Fazekas (Total)		Longstreth	
	Method	Spearman's Corr.		Spearman's Corr.	
		rho	p	rho	p
1	Manual label	0.7562	1.04×1^{-12}	0.7752	1.45×10^{-12}
1	LST-LGA	0.5718	3.38×6^{-12}	0.4813	1.50×10^{-4}
2	SVM	0.4062	1.70×10^{-2}	0.3602	5.90×10^{-3}
3	RF	0.2447	6.66×10^{-2}	0.2128	1.12×10^{-1}
4	DBM	0.2436	6.79×10^{-2}	0.1659	2.17×10^{-1}
5	CEN	0.2359	7.74×10^{-2}	0.3618	5.70×10^{-3}
1	Patch-UResNet	0.3602	5.90×10^{-3}	0.5171	3.80×10^{-5}
1	Patch-UNet	0.4618	2.99×10^{-4}	0.5140	4.33×10^{-5}
6	DeepMedic-GSI-2D	0.7054	9.01×10^{-10}	0.8664	3.19×10^{-18}
7	LIM-2048s128m	0.4727	2.05×10^{-4}	0.4579	3.42×10^{-4}
8	LIM-1024s128m	0.4892	1.13×10^{-4}	0.4849	1.32×10^{-4}
9	LIM-512s64m	0.5010	7.19×10^{-5}	0.5065	5.82×10^{-4}
10	LIM-256s64m	0.5009	7.22×10^{-5}	0.5085	5.37×10^{-4}
11	LIM-128s32m	0.4505	4.38×10^{-4}	0.4946	9.22×10^{-4}
12	LIM-64s16m	0.4393	6.30×10^{-4}	0.4858	1.28×10^{-4}

5.9. Applicability to assess MS lesion progression

The evolution of interval or enlarging lesions on T2-dependent imaging are key criterion for assessing MS disease activity, which informs clinical decision-making and as a surrogate endpoint for clinical trials of therapeutic agents. We coregistered the raw (i.e. not post-processed) LOTS-IM output obtained from the baseline and follow-up FLAIR images to a mid-space and subtracted both maps. Then, we performed 3d connected component analysis to the "positive" and "negative" regions of the subtracted maps, being these regions comprised by the voxels with modular values higher than 0.18. We followed the same thresholding criterion as the one followed to extract the WMH to neglect subtle differences due to misregistrations, cortical effects, or differences in image contrast not related to the disease. We counted the "positive" spatial clusters (i.e. connected components) with IM values (i.e. in at least one voxel) higher than 0.7 and labelled those spatial clusters as "New or enlarged lesions". We summed the areas (in number of voxels) covered by all the "positive" and "negative" spatial clusters weighted by their mean IM value, separately and subtracted them to determine the overall change and rated it in none-low (less than 25% of the "positive" areas), low-moderate (between 25 and 50% of the "positive" areas) and moderate-high (more than 50% of the "positive" areas. Two expert raters, independently and blind to any quantitative analysis, visually assessed the baseline and follow-up images and identified the number of new and/or enlarged MS lesions, and rated the disease progression. Discrepancies among raters were discussed and a final result was agreed. Table 9 shows the results from both assessments.

Table 9: Visual expert vs. LOTS-IM longitudinal MS lesion assessments in 10 treatment-free MS patients

Patient	Visual expert assessment		LOTS-IM output	
	New or enlarged lesions	Dis. progres.	New or enlarged lesions	Dis. progres.
1	0	None-Low	0	None-Low
2	1	None-Low	3	Moderate-High
3	0	None-Low	2	None-Low
4	1	Moderate-High	5	Low-Moderate
5	0	None-Low	1	None-Low
6	2	Moderate-High	3	Moderate-High
7	5	Moderate-High	6	Low-Moderate
8	3	Moderate-High	1	None-Low
9	2	Moderate-High	5	Moderate-High
10	6	Moderate-High	2	Moderate-High

The agreement between LOTS-IM and the expert assessment in rating the disease progression was 80%. The criterion followed to count the new or enlarged lesions from the LOTS-IM needs to be revised though, as the automatic count does not reflect actual disease change in some cases. MS lesions which are active at one imaging time point and subsequently quiescent frequently decrease in size; summed volume measures may therefore not identify active disease when some lesions are enlarging and others are shrinking. The automatic processing of LOTS-IM output as described above, identified changes produced by CSF flow artifacts in the choroid plexus, temporal poles and junction between the septum and the callosal genu or splenium (Figure 9). These, although genuine signal changes, are not related to the disease. Instead, these are particular confounders in MS where genuine lesions frequently abut ventricular margins. In addition, the centre of T1-weighted "black holes", not hyperintense in baseline FLAIR images but hyperintense in the follow-up FLAIR, which maybe due to variation in efficacy of fluid signal suppression, was also counted as new lesion (Figure 9). Although initial results are promising, post-processing to correct these "false" positives and negatives will be necessary for applying LOTS-IM to clinical research in MS.

6. Conclusion and Future Work

Through this study, we have shown that the optimisation of the irregularity map method presented (LOTS-IM) accelerates processing time by large margin without excessive quality degradation compared with the previous iterations (IAM and OTS-IAM). LOTS-IM speeds up the overall computation time, attributable not only to implementation on GPU, but also to the use of a limited number of target patch samples. In addition, we have evaluated LOTS-IM in different scenarios, which was not done in previous studies.

Unlike other WMH segmentation methods, LOTS-IM successfully identifies and represents the WMH "penumbra" using irregularity value and irregularity map. Despite not being a WMH segmentation method *per se*, LOTS-IM can be applied for this purpose by thresholding the value of the irregularity map. Being unsupervised confers an additional value to this fully automatic method as it does not depend on expert-labelled data, and therefore is independent from any subjectivity and inconsistency from human experts,

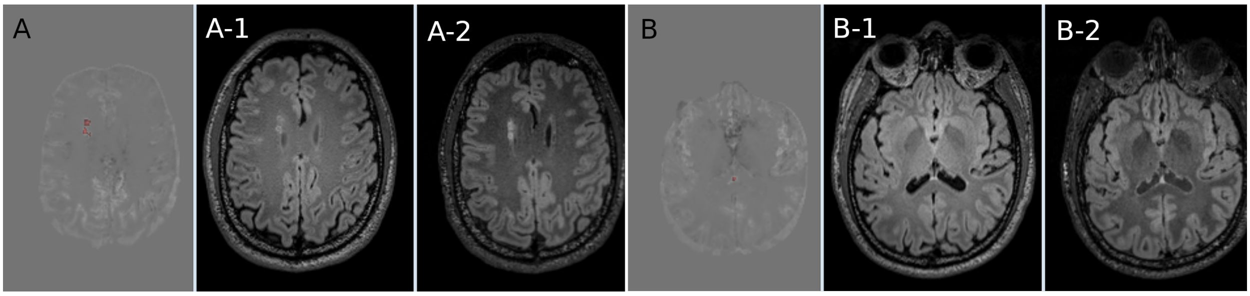

Figure 9: Two examples (i.e. A and B) of WMH change captured by LOTS-IM that do not represent actual disease progression from the neuroradiological perspective. A-1 and B-1 are the baseline FLAIR axial slices and A-2 and B-2 are the corresponding follow-up FLAIR slices. The "new" lesions captured by LOTS-IM are represented in red on the "change" LOTS-IM. In A, the two "new" lesions are void regions on the baseline FLAIR, in the centre of two neighbouring hyperintensities, which are hyperintense at follow-up. In B, the "new" lesion is in reality a CSF flow artefact in the intersection between the septum and the splenium of the corpus callosum.

which typically influence supervised machine learning algorithms. Our results show that LOTS-IM outperforms LST-LGA, the current *state-of-the-art* unsupervised method for WMH segmentation, conventional supervised machine learning algorithms (SVM and RF), and some supervised deep neural networks algorithms (DBM and CEN). Our results also show that LOTS-IM has comparable performance with the *state-of-the-art* supervised deep learning algorithms DeepMedic, UResNet, and UNet.

LOTS-IM is still influenced by the quality of brain masks (*i.e.*, CSF and NAWM) and its random sampling scheme to sample target patches. We have shown that random sampling has a small impact to the final result on WMH segmentation, but more sophisticated sampling could be used as well. Some improvements also could be done by adding or using different sets of brain tissues masks other than CSF and NAWM, such as cortical and cerebrum brain masks.

We believe that the irregularity map could provide unsupervised information for pre-training supervised deep neural networks, such as UResNet and UNet. In (Rachmadi et al., 2018a), UNet successfully learned the irregularity map produced by LOTS-IM. Progression/regression of brain abnormalities also can be achieved with LOTS-IM(Rachmadi et al., 2018a). Due to its principle, it could be applicable to segment brain lesions in CT scans or different brain pathologies, but further evaluation would be necessary. Further works could also explore its implementation on a multispectral approach that combines different MRI sequences. The implementation of LOTS-IM on GPU is publicly available on the following GitHub page[6]

Acknowledgements

The first author 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.

Data collection and sharing for this project was partially funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharma-

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

ceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

FutureMS is supported by an exemplar grant from the Stratified Medicine Scotland Innovation Centre and funding from Biogen, Inc (Cambridge, Massachusetts, https://www.biogen.com/). RM (and partially MCVH) salaries are supported by the CSO-PME grant to the Stratified Medicine Scotland Innovation Centre.

References

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

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

Brosch, T., Tang, L. Y., Yoo, Y., Li, D. K., Traboulsee, A., Tam, R., 2016. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE transactions on medical imaging 35 (5), 1229–1239.

Callisaya, M. L., Beare, R., Phan, T. G., Blizzard, L., Thrift, A. G., Chen, J., Srikanth, V. K., 2013. Brain structural change and gait decline: A longitudinal population-based study. Journal of the American Geriatrics Society 61 (7), 1074–1079.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/jgs.12331`

Chen, X., Konukoglu, E., 2018. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972.

Clayden, J., Maniega, S., Storkey, A., King, M., Bastin, M., Clark, C., 2011. Tractor: Magnetic resonance imaging and tractography with r. Journal of Statistical Software, Articles 44 (8), 1–18.
URL `https://www.jstatsoft.org/v044/i08`

del C. Valdés Hernández, M., Booth, T., Murray, C., Gow, A. J., Penke, L., Morris, Z., Maniega, S. M., Royle, N. A., Aribisala, B. S., Bastin, M. E., Starr, J. M., Deary, I. J., Wardlaw, J. M., 2013. Brain white matter damage in aging and cognitive ability in youth and older age. Neurobiology of Aging 34 (12), 2740 – 2747.
URL `http://www.sciencedirect.com/science/article/pii/S0197458013002455`

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

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

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

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., Zimmerman, R. A., 1987. Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. American journal of roentgenology 149 (2), 351–356.

Firbank, M., Minett, T., Obrien, J., 2003. Changes in dwi and mrs associated with white matter hyperintensities in elderly subjects. Neurology 61 (7), 950–954.

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.

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M., 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.

Hernández, M. d. C. V., Ferguson, K. J., Chappell, F. M., Wardlaw, J. M., 2010. New multispectral mri data fusion technique for white matter lesion segmentation: method and comparison with thresholding in flair images. European radiology 20 (7), 1684–1691.

Hernández, M. d. C. V., Morris, Z., Dickie, D. A., Royle, N. A., Maniega, S. M., Aribisala, B. S., Bastin, M. E., Deary, I. J., Wardlaw, J. M., 2013. Close correlation between quantitative and qualitative assessments of white matter lesions. Neuroepidemiology 40 (1), 13–22.

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

Jeerakathil, T., Wolf, P. A., Beiser, A., Massaro, J., Seshadri, S., Dagostino, R. B., DeCarli, C., 2004. Stroke risk profile predicts white matter hyperintensity volume: the framingham study. Stroke 35 (8), 1857–1861.

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

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

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

Li, H., Jiang, G., Wang, R., Zhang, J., Wang, Z., Zheng, W.-S., Menze, B., 2018. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. arXiv preprint arXiv:1802.05203.

Longstreth, W., Manolio, T. A., Arnold, A., Burke, G. L., Bryan, N., Jungreis, C. A., Enright, P. L., O'Leary, D., Fried, L., Group, C. H. S. C. R., et al., 1996. Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people the cardiovascular health study. Stroke 27 (8), 1274–1282.

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

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

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., Beckett, L., 2005. The alzheimer's disease neuroimaging initiative. Neuroimaging Clinics of North America 15 (4), 869–877.

Myers, J. L., Well, A., Lorch, R. F., 2010. Research design and statistical analysis. Routledge.

Pohjasvaara, T., Mäntylä, R., Salonen, O., Aronen, H. J., Ylikoski, R., Hietanen, M., Kaste, M., Erkinjuntti, T., 2000. How complex interactions of ischemic brain infarcts, white matter lesions, and atrophy relate to poststroke dementia. Archives of neurology 57 (9), 1295–1300.

Rachmadi, M. F., del C. Valdés-Hernández, M., Komura, T., 2018a. Transfer learning for task adaptation of brain lesion assessment and prediction of brain abnormalities progression/regression using irregularity age map in brain mri. In: Rekik, I., Unal, G., Adeli, E., Park, S. H. (Eds.), PRedictive Intelligence in MEdicine. Springer International Publishing, Cham, pp. 85–93.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Agan, M. L. F., Di Perri, C., Komura, T., Initiative, A. D. N., et al., 2018b. Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain mri with none or mild vascular pathology. Computerized Medical Imaging and Graphics 66, 28–43.

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

Rachmadi, M. F., Valdés-Hernández, M. d. C., Komura, T., 2017b. Voxel-based irregularity age map (iam) for brain's white matter hyperintensities in mri. In: Advanced Computer Science and Information Systems (ICACSIS), 2017 International Conference on. IEEE, pp. 321–326.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Komura, T., 2018c. Automatic irregular texture detection in brain mri without human supervision. In: Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 506–513.

Rensma, S. P., van Sloten, T. T., Launer, L. J., Stehouwer, C. D., 2018. Cerebral small vessel disease and risk of incident stroke, dementia and depression, and all-cause mortality: A systematic review and meta-analysis. Neuroscience & Biobehavioral Reviews.

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

Salakhutdinov, R., Larochelle, H., 2010. Efficient learning of deep boltzmann machines. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 693–700.

Scheltens, P., Barkhof, F., Leys, D., Pruvo, J., Nauta, J., Vermersch, P., Steinling, M., Valk, J., 1993. A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. Journal of the neurological sciences 114 (1), 7–12.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 146–157.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V. J., Zimmer, C., et al., 2012. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. Neuroimage 59 (4), 3774–3783.

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D. S., Calabresi, P. A., Pham, D. L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49 (2), 1524–1535.

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).

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R. I., T O'Brien, J., Barkhof, F., Benavente, O. R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology 12 (8), 822–838.