Received: 22 November 2021 | Revised: 24 March 2022 | Accepted: 20 April 2022

DOI: 10.1002/hbm.25899

RESEARCH ARTICLE

WILEY

Higher-resolution quantification of white matter hypointensities by large-scale transfer learning from 2D images on the JPSC-AD cohort

Benjamin Thyreau[1] | Yasuko Tatewaki[2,3] | Liying Chen[1] | Yuji Takano[1,4] | Naoki Hirabayashi[5] | Yoshihiko Furuta[5] | Jun Hata[5] | Shigeyuki Nakaji[6] | Tetsuya Maeda[7] | Moeko Noguchi-Shinohara[8] | Masaru Mimura[9] | Kenji Nakashima[10] | Takaaki Mori[11] | Minoru Takebayashi[12] | Toshiharu Ninomiya[5] | Yasuyuki Taki[1,2,3] | Japan Prospective Studies Collaboration for Aging and Dementia (JPSC-AD) Study Group[†]

[1]Smart-Aging Research Center, Institute of Development, Aging, and Cancer, Tohoku University, Sendai, Japan

[2]Department of Aging Research and Geriatric Medicine, Institute of Development, Aging, and Cancer, Tohoku University, Sendai, Japan

[3]Department of Geriatric Medicine and Neuroimaging, Tohoku University Hospital, Sendai, Japan

[4]Department of Psychological Sciences, University of Human Environments, Matsuyama, Japan

[5]Department of Epidemiology and Public Health, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

[6]Department of Social Medicine, Graduate School of Medicine, Hirosaki University, Hirosaki, Japan

[7]Division of Neurology and Gerontology, Department of Internal Medicine, School of Medicine, Iwate Medical University, Iwate, Japan

[8]Department of Neurology and Neurobiology of Aging, Kanazawa University Graduate School of Medical Sciences, Kanazawa University, Kanazawa, Japan

[9]Keio University School of Medicine, Tokyo, Japan

[10]National Hospital Organization, Matsue Medical Center, Shimane, Japan

Abstract

White matter lesions (WML) commonly occur in older brains and are quantifiable on MRI, often used as a biomarker in Aging research. Although algorithms are regularly proposed that identify these lesions from T2-fluid-attenuated inversion recovery (FLAIR) sequences, none so far can estimate lesions directly from T1-weighted images with acceptable accuracy. Since 3D T1 is a polyvalent and higher-resolution sequence, it could be beneficial to obtain the distribution of WML directly from it. However a serious difficulty, both for algorithms and human, can be found in the ambiguities of brain signal intensity in T1 images. This manuscript shows that a cross-domain ConvNet (Convolutional Neural Network) approach can help solve this problem. Still, this is non-trivial, as it would appear to require a large and varied dataset (for robustness) labelled at the same high resolution (for spatial accuracy). Instead, our model was taught from two-dimensional FLAIR images with a loss function designed to handle the super-resolution need. And crucially, we leveraged a very large training set for this task, the recently assembled, multi-sites Japan Prospective Studies Collaboration for Aging and Dementia (JPSC-AD) cohort. We describe the two-step procedure that we followed to handle such a large number of imperfectly labeled samples. A large-scale accuracy evaluation conducted against FreeSurfer 7, and a further visual expert rating revealed that WML segmentation from our ConvNet was consistently better. Finally, we made a directly usable software program based on that trained ConvNet model, available at https://github.com/bthyreau/deep-T1-WMH.

[†] The members of study group are listed in Acknowledgments section.

[11]Department of Neuropsychiatry, Ehime University Graduate School of Medicine, Ehime University, Ehime, Japan

[12]Faculty of Life Sciences, Department of Neuropsychiatry, Kumamoto University, Kumamoto, Japan

Correspondence
Benjamin Thyreau, Smart-Aging Research Center, Institute of Development, Aging and Cancer, Tohoku University, Sendai, Japan.
Email: benjamin.thyreau@gmail.com

Funding information
Suntory Holdings Limited (Osaka, Japan); Japan Agency for Medical Research and Development, Grant/Award Number: JP21dk0207053

1 | INTRODUCTION

White matter lesions (WML) frequently occur in the brain later in life (Prins & Scheltens, 2015) and can be characterized in terms of aspect and spatial distribution. On T2-fluid-attenuated inversion recovery (FLAIR) MRI scans, they appear as punctuated or even patchy white matter hyperintensities (WMH) that are spatially distributed around the ventricles or extend deep into the white matter in some cases. Such lesions are also visible, albeit less prominently, as hypointense patches on T1 images.

Historically, MRI has been used to quantitatively evaluate WML for clinical and research purposes using visual rating scales, with the Fazekas scale (Fazekas et al., 1987) being one of the most employed. However, in practice, WMH rating scales are of limited use beyond a purely descriptive purpose. More recently, quantitative information has become available thanks to computerized image processing (Frey et al., 2019). These two metrics can partially concur (Cedres et al., 2020; Koikkalainen et al., 2019).

The etiology of WML can vary (Wardlaw et al., 2019) and is still not well understood (Alber et al., 2019; Debette & Markus, 2010). Large occurrences of WML in the brain have reportedly been associated with negative brain traits, such as cognitive decline, dementia, stroke, or intracranial hypertension (Atwi et al., 2018; Fazekas et al., 1993; Habes et al., 2018; Lampe et al., 2019; Ngai et al., 2007; Ni et al., 2021; Pantoni, 2010; Pozorski et al., 2019; Sarica et al., 2019; Tubi et al., 2020). However, despite these broad correlation trend reports, no definitive conclusion has been made or accepted (Rhodius-Meester et al., 2017; Vangberg et al., 2019). A large meta-analysis of the predictive power of several factors for dementia found either mixed or no evidence for an effect of WMH (Ansart et al., 2021).

In order to disentangle the multiple putative causative factors associated with WML, more research should be undertaken involving larger, wider cohorts and automated image analysis which is capable of efficiently and reproducibly extracting the number of characteristics of WML from images. Consequently, WMH segmentation remains a popular technical problem among the image analysis research community. A frequent approach relies on atlases and histogram clustering methods (Schirmer et al., 2019). Tools such as LST (Schmidt et al., 2012), BIANCA (Griffanti et al., 2016; Sundaresan et al., 2019), and UBO (Jiang et al., 2018) were ranked favorably in a recent comparison (Heinen et al., 2019). Moreover, deep learning methods have attracted increasing attention over the past few years. Unfortunately, (Balakrishnan et al., 2021) noted that only eight of 37 authors made their method available at all, which may explain their lack of adoption by the wider clinical community.

In this manuscript, we address the problem of WMH identification from a multi-modality (T1-weighted imaging) and large-scale learning perspective. We rely on Convolutional Neural Networks (ConvNets) for image learning. Our source material comprises a very large, aggregated dataset of 7694 brain images acquired in the context of the JPSC-AD project (Ninomiya et al., 2020), a multisite, population-based prospective cohort initiated in 2016 and has thus far recruited over 10,000 participants. To facilitate the collaboration of hospitals, the MR-imaging component of the project was designed with flexibility in mind, with only T1-weighted images being the common requirement.

1.1 | Challenges

The following challenges should be considered when performing WMH identification in the most effective way.

1.1.1 | Population heterogeneity

As much as possible, association studies should be founded on large samples, with enough heterogeneity to cover a wide array of brain health statuses. We believe that a large prospective research cohort is more suitable than a small, specific subset of patients. On the other hand, a larger dataset also requires a greater human annotation effort, or at least, a larger evaluation effort.

1.1.2 | Lesions load and localization

Apparent lesions can have different etiologies depending on their shape and location, and therefore different predictive clinical importance (Lampe et al., 2019). Periventricular (PVWM) lesions and deep-white matter lesions (DWM), which have varying effects on cognition and arterial pressure, may be entirely different entities (Griffanti et al., 2018). Small, irregular patches may be a sign of ischemia, whereas larger, smoother periventricular lesions may simply be a benign symptom of aging. Small punctate lesions may eventually develop into pathologies, but their role still needs to be clarified (Prins & Scheltens, 2015). Local features of the vasculature may also have an impact (Debette et al., 2019). Overall, to date, the etiology of WMH is still largely unknown, and existing classification scales partly reflect this (van Straaten et al., 2006).

In the computational domain, consequently, an identification algorithm should not rely solely on blindly aggregated metrics such as total voxel overlap as its unique objective, or it will risk missing this distinction and would become biased toward large patches, ignoring the low voxel contributions of punctate lesions and leading to sub-optimal clinical predictive power.

1.1.3 | Sequence and resolution

FLAIR is one of the optimal sequences for observing WMH, but 3D FLAIR sequences with high SNR are not yet widespread in hospital settings. Two-dimensional FLAIR sequences are easier to obtain but suffer from poor resolution along the slice axis, which not only reduces accuracy in that dimension but also introduces a source of mismatch for longitudinal comparisons.

Alternatively, lesion-induced WMHs are visible on T1 weighted images in most cases, where they appear as hypointense regions (and the acronym *WMH* will refer to both contrast signals from now on). It is common to use the T1-weighted contrast for structural, high-resolution images of the brain, for example, to allow accurate estimates of ventricle enlargement or hippocampal volume or atrophy (Nogovitsyn et al., 2019; Thyreau et al., 2018). Therefore, it is useful to obtain WMH delineation directly from regular 3D T1 images.

However, in a major difference from FLAIR, lesions-induced WMH on T1 images appear with the same signal intensity as gray-matter structures and would require considerably more attention from an investigator to confidently identify them by visual inspection, particularly for infra-cortical lesions. This, however, can be accomplished using moderns algorithms.

1.1.4 | Robustness and consistency

The model should be able to robustly identify lesions in all patients, with a low failure or rejection rate. In technical terms of machine learning, the model should be exposed to a dataset covering large patterns of the population. When using small samples, it is tempting to emphasize pixel-perfect metric scores that could easily be over-sensitive to the particular sample or to a particular chosen threshold. Some other types of loss exist, such as perimeter loss (el Jurdi et al., 2021), but those still assume that the target, or its perimeter, is well-defined. Therefore, while such metrics may be a perfect way to evaluate a model's ability to learn accurately, they may not necessarily be able to adequately gauge the success of the task itself.

1.2 | Objectives

The goals of this manuscripts are as follows:

- We explore the use of transfer-learning from 2D-FLAIR to 3D-T1, introducing the use of a spatially sparse loss to handle super-resolution needs;
- We leverage a new, very-large cohort, the JPSC-AD cohort. We design a two-step learning scheme to better manage the mass of data with limited human guidance;
- For validation, we run a large-sample automated quantitative comparison against FreeSurfer (Fischl, 2012), and a relatively large (200 images) qualitative visual evaluation by experts.
- We embedded the model trained on the full dataset into an open-source tool available for download

2 | MATERIALS AND METHODS

2.1 | Dataset

The MRI brain images were obtained from the Japan Prospective Studies Collaboration for Aging and Dementia (JPSC-AD), a project which collected data from over 10,000 participants across eight sites. The strategy followed to elucidate risk factors and the etiology of dementia is described in (Ninomiya et al., 2020). The mean age of the dataset was 74.4, starting from 65, 42% was male, and 8.5% of the dataset were diagnosed with dementia. The participants all underwent an MRI exam with at least one T1-weighted acquisition. Other sequences were collected and made available at the discretion of the participating hospital. This level of flexibility was designed to encourage hospital participation. The scanners used in these hospitals were diverse, including 1.5 T MRI (four Philips, one Hitachi, and one General-Electric) and 3 T MRI (one GE and one Siemens) machines.

This study collected all the FLAIR acquisitions available giving a total of 7699 images from six sites. The 2D FLAIR images comprised 18–24 slices depending on the scanning site, with slice thickness ranging from 5.9 to 7.8 mm. The axial resolution ranged from 0.47 to 0.57 mm. One site performed a systematic Fazekas rating of all its FLAIR images, but this information was not directly used for the purpose of the main ConvNet development. Regarding structural imaging, 10,017 sagittal T1-weighted images were acquired (3D MPRAGE or SPGR sequences). The voxel dimension was 1.2 mm along the sagittal axis and varied from 0.93 to 1.02 mm in the other dimensions.

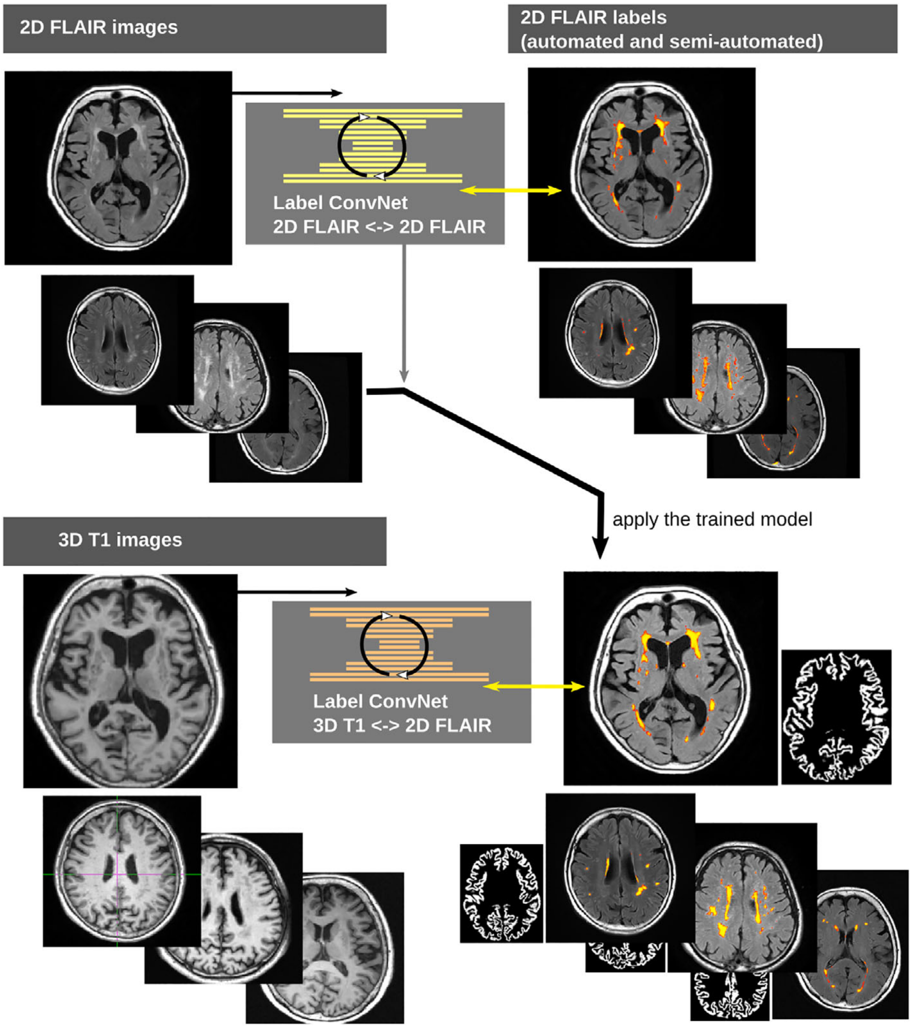

FIGURE 1 The two-step learning process. The 2D FLAIR model (top) is trained to output a consistent FLAIR mask, initially based on the LST algorithm but with some further manual corrections. The T1 model (bottom) aims to learn a T1 WMH segmentation from those 2D-FLAIR masks, learning across modality and resolution. It also learns a cortical ribbon as a secondary joint task. Both steps, including relevant preprocessing, use the same, single, training dataset.

2.2 | Preprocessing

The T1 images were processed on FreeSurfer 7.1. From these, 9985 could complete a full image processing pipeline, while the remaining images were excluded due to data handling mistakes or software failures, including non-convergence within 24 h. However, we did not attempt to visually confirm all individual results. T1 images which had no corresponding FLAIR image were of no further use to the work described here. The T1 images were registered and resampled into an MNI space clipped immediately below the temporal lobes, measuring 184 × 202 × 72 voxels of size 1 × 1 × 1.5 mm, which defines our working space for all T1-related ConvNet inputs and outputs in the following sections.

A rigid registration matrix, from FLAIR to T1, was estimated using NiftyReg Aladin (Modat et al., 2010). All of the 7699 registrations were visually quality-checked based on a single middle slice, after which only five further subjects were excluded.

2.3 | ConvNet model training

The overall idea was to use the lesion information obtained from cleanly segmented 2D-FLAIR images as targets to train a 3D-T1 ConvNet. We achieved this by developing a two-step process. The first step was to generate accurate labels in the 2D-FLAIR space, and the second was to train the 3DT1 model from these sparse 2D-FLAIR labels (Figure 1).

2.4 | FLAIR lesion ConvNet

As a first step, the LST (Schmidt et al., 2012) FLAIR segmentation software was applied to all 2D FLAIR images. Despite LST being considered accurate and a strong performer in comparative studies (Heinen et al., 2019; Ribaldi et al., 2021; Vanderbecq et al., 2020), a small but significant fraction of the results inevitably contained some inaccurate or unsatisfactory segmentation (Figure S2). While we could simply review all images and discard these failed cases, this would risk losing some interesting corner-cases, which would run contrary to our goal of robustness. Instead, we aimed to identify such cases and manually fix a subset of them. To achieve this, we rely on one ability of ConvNet, that is, they integrate knowledge across multiple samples and learn the main trends of the data before the particular deviations (Arpit et al., 2017). Therefore, we first trained a “FLAIR” ConvNet model to mimic the LST segmenter

TABLE 1 Topology of the ConvNets used through this study.

T1-WMH net	T1-ROI net	FLAIR net (WMH or ROI)
Input: a T1 image in the MNI workspace Conv3(1, 24), InstNorm, ReLU, Conv3(24, 64), ReLU (block0) MaxPool Conv3(64, 64), InstNorm, ReLU, Conv3(64, 64), ReLU (block1) MaxPool Conv3(64, 64), InstNorm, ReLU, Conv3(64, 64), ReLU (block2) MaxPool Conv3(64, 64), InstNorm, ReLU, Conv3(64, 64), ReLU (block3) MaxPool Conv3(64, 64), InstNorm, ReLU, Conv3(64, 64), ReLU (block4) MaxPool Conv3(64, 128), InstNorm, ReLU, Conv3(128, 64), ReLU Unpool Conv3(64, 64), InstNorm, ReLU, Conv3(64, 64), ReLU Unpool Conv3(64, 64), InstNorm, ReLU, Sum block3, Conv3(64, 64), ReLU Unpool Conv3(64, 64), InstNorm, ReLU, Sum block2, Conv3(64, 64), ReLU Unpool Conv3(64, 64), InstNorm, ReLU, Sum block1, Conv3(64, 64), ReLU Conv3(64, 64), InstNorm, ReLU, Sum block0, Conv3(64, 64), ReLU Conv3(64, 24), ReLU), Conv1(24, 2), Sigmoid output: WMH and Cortex masks	Input: a T1 image in the MNI workspace Conv3(1, 12, 3) ReLU, (block0) MaxPool, Conv3(12, 16), ReLU, Conv1(16, 16), ReLU (block1) MaxPool, Conv3(16, 16), ReLU, Conv1(16, 16), ReLU (block2) MaxPool, Conv3(16, 16), ReLU, Conv1(16, 16), ReLU (block3) MaxPool, Conv3(16, 16), ReLU, Conv1(16, 16), ReLU (block4) MaxPool, Conv3(16, 16), ReLU, Conv1(16, 16), InstNorm, ReLU, Unpool, Conv3(16, 16), InstNorm, ReLU, Sum block4, Conv3(16, 16), ReLU Unpool, Conv3(16, 16), InstNorm, ReLU, Sum block3, Conv3(16, 16), ReLU Unpool, Conv3(16, 16), InstNorm, ReLU, Sum block2, Conv3(16, 16), ReLU Unpool, Conv3(16, 16), InstNorm, ReLU, Sum block1 Conv3(16, 12), ReLU Unpool, Conv3(12, 12), ReLU, Sum block0, Conv1(12, 12), ReLU, Conv3(12, 8), ReLU, Conv1(8, 4), Softmask output: Four label maps	Input: a FLAIR image in its native space Conv3(1, 12), InstNorm, ReLU, Conv1(12, 12), ReLU (block0) MaxPool, Conv3(12, 16), ReLU, Conv3(16, 16), ReLU (block1) MaxPool, Conv3(16, 16), ReLU, Conv3(16, 16), ReLU (block2) MaxPool, Conv3(16, 16), ReLU, Conv3(16, 16), ReLU (block3) MaxPool, Conv3(16, 16), ReLU, Conv3(16, 16), ReLU (block4) MaxPool, Conv3(16, 16), ReLU, Conv1(16, 16), InstNorm, ReLU Unpool, Conv3d(16, 16, 3), InstNorm, ReLU, Sum block4 Conv3d(16, 16, 3), ReLU Unpool, Conv3d(16, 16, 3), InstNorm, ReLU, Sum block3, Conv3d(16, 16, 3), ReLU Unpool, Conv3d(16, 16, 3), InstNorm, ReLU, Sum block2, Conv3d(16, 16, 3), ReLU Unpool, Conv3d(16, 16, 3), InstNorm, ReLU, Sum block1, Conv3d(16, 12, 3), ReLU Unpool, Conv3d(12, 12, 3), InstNorm, ReLU, Sum block0, Conv1(12, 12), ReLU, Conv3(12, 8), ReLU, Conv1(8, 3), Sigmoid output: WMH or ROIs, depending on the loaded parameters
		Note: Starting from block 2, Unpool halve dimension in axial planes only

Note: All networks are U-shaped with skip-connections. Their main difference is the number of convolution kernels, which were hand-designed following a trade-off between training speed, model capacity, and resolution. Conv3 and Conv1 refers to the convolution kernel voxel size (3 × 3 × 3 or 1 × 1 × 1). MaxPool operators halve the dimensions and returns indices which are used by Unpool operators.

on a large training set and stopped the training when the loss function appeared to reach a slower convergence regime. Since the dataset is large, and the ConvNets do not easily overfit, we reasoned that we would find many unusual or mis-segmented samples among those that exhibited a large loss error.

Half of the dataset was used to train this FLAIR ConvNet, and no test set was involved. The network had a simple U-Net architecture, and the parameters are described in Table 1. The input format was always the native FLAIR image space, even though the number of slices varied between sites. Given its "convolutional" nature, the network naturally handles input size variations, without the need to introduce inflexible padding, or worse, resizing and resampling artifacts. Consequently, for practical simplicity, the mini-batch size was set to 1 during the optimization, and batch-normalizations were avoided in favor of instance-normalizations in the network. Geometrical augmentation was performed with random rotations—only around the axial axis, to avoid inter-slice resampling—and random translations.

A set of images where the current segmentation had room for improvement (mainly the false-negatives with many small lesions missed, or missed lesions close to the cortex, or false-positive due to bright artifacts), were hand-selected for further consideration. To assist at this task, and because of the significant number of images remaining (and the lack of a meaningful cut-off threshold), the samples were displayed on a dedicated data navigator, according to their estimated WMH-volume in different regions (periventricular, deep white, and infra-cortical [defined later]), and with an instant feedback feature showing each result as a mosaic of slices, to better appreciate the model success range.

In this process, many images (100 ~ 150), the most outlier ones, were quickly explored. Although a dozen images were dropped due to their segmentation being deemed unrecoverable (e.g., strong motion), 55 images, a number that mostly reflected the amount of effort required, were eventually selected and manually corrected using ITK-Snap (Yushkevich et al., 2006). The correction used the model output as a starting point. Despite this, we spent, on average 15–20 min on a single image.

The FLAIR ConvNet model was then re-trained on the curated dataset. The weight attributed to the manually corrected samples was increased by running 10 times more augmentation than the rest of the samples. This second FLAIR ConvNet model was then used to generate the reference FLAIR segmentations, which formed the target of the next step, the T1 ConvNet.

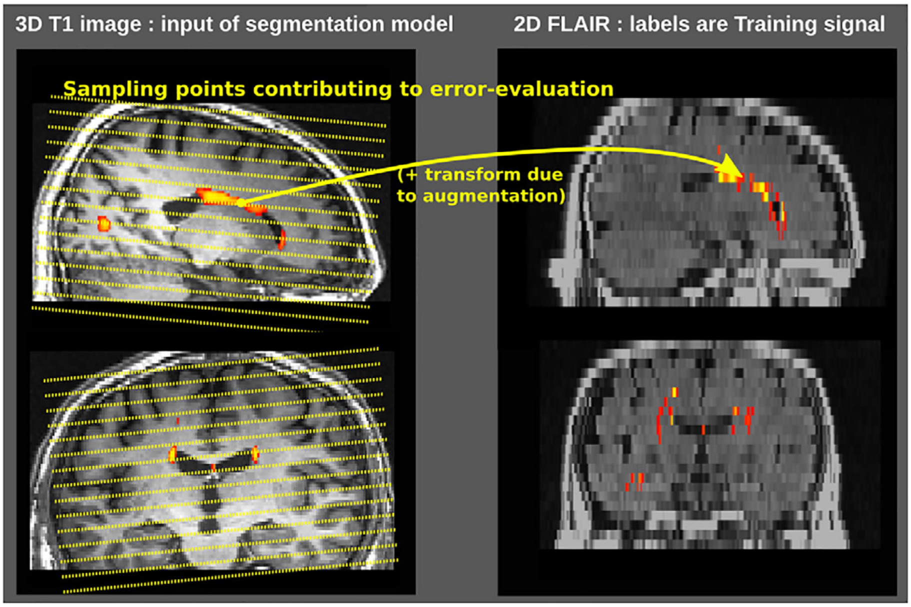
3D T1 image : input of segmentation model
2D FLAIR : labels are Training signal
Sampling points contributing to error-evaluation
(+ transform due to augmentation)

FIGURE 2 Illustration of the sampling performed during the loss computation. The training signal for each voxel of the target image, in FLAIR space, is back-projected into a corresponding voxel of the T1 output space, perhaps through an augmentation transform, then back-propagated further up through the model's convolutional layers. The T1 voxels that are not directly linkable to a FLAIR signal are not affected (in particular, they do not receive an interpolated signal).

2.5 | T1 lesions ConvNet model

The T1 ConvNet model, whose role is to segment the WMH lesions in the T1 working space, is the main component of this work. The architecture is a variant of U-Net, and the parameters are listed in Table 1. One important concern was the resolution gap between the two modalities. While modern T1WI scans are mostly isotropic (~1 mm) 3D images, typical hospital FLAIR sequences produce a relatively small set of thick (~6 mm) but high-resolution (~0.5 mm) 2D slices. Consequently, beyond the obvious contrast difference between the two modalities, the lesion-related information also differs in terms of its spatial distribution. In particular, the T1WI image would contain more detailed information in the axial direction than the 2D FLAIR could provide. For cross-modality learning, while we could simply align the two modalities with an interpolation algorithm, we observed that the final results produced by this method were unsatisfactory. Instead, we used the ConvNet ability to accumulate knowledge across multiple samples in order to supply the missing details of each individual image.

2.6 | Loss function

To achieve this, we used a spatially sparser loss (Figure 2). The primary output of our network—the WMH mask in T1 space—is dynamically sampled at points where the FLAIR provides the actual training signal, that is, the loss calculation simply ignores all voxels of T1 space that do not map closest to a FLAIR voxel center. This loss is sparse in the sense that each sample contributes only a few slices' worth of training information, less than the higher T1 processing resolution model would optimally need. However, by virtue of being exposed to a large number of subjects, themselves augmented with random affine transforms, the model generalizes naturally from the sparse training information to the full processing resolution.

Second, a T1 lesion model must learn to precisely distinguish subcortical lesions from the cortex itself. Therefore, as an auxiliary helper task, we explicitly trained the model to produce a cortical ribbon mask as a secondary output. This was performed entirely in the T1 input space. From previous experience, this is a relatively easy task in terms of required ConvNet model capacity and should not, therefore, hinder the primary outcome. Besides providing a supplemental, explicit training signal, this helped us to understand the behavior of the model during its optimization.

We used a Euclidean distance in both parts. We favor this simple metric as it can easily be interpreted in terms of an independent voxel probability. The final loss function is thus the weighted sum of the sparse FLAIR lesion signal and the cortical segmentation loss:

$$L = \Sigma_{(FLAIR)}(S(\mathrm{out}_1) - FLAIR\_wmh)^2 + 0.001 \\ \times \Sigma_{(T1)}(\mathrm{out}_2 - T1\_ribbon)^2,$$

where the out_i are the model outputs in T1 space; S: T1 → FLAIR is the voxel sampling function; FLAIR_wmh and T1_ribbon are, respectively, the target lesion and ribbon masks, in their own FLAIR and T1 spaces. The FLAIR_wmh masks are the one obtained from the previous step (the FLAIR-ConvNet step), while the T1_ribbon masks are simply FreeSurfer cortical segmentations.

The training set was made from 4096 subjects (and included the subjects used to train the FLAIR ConvNet). In order to augment the T1 dataset, and in addition to the random affine transformations, we randomly altered the image histogram to simulate the variation of contrast of different scanners and sequences.

2.7 | Regions of interest

A separate ConvNet model was created to quickly identify the main regions of interest for WMH, based on prior knowledge. This region

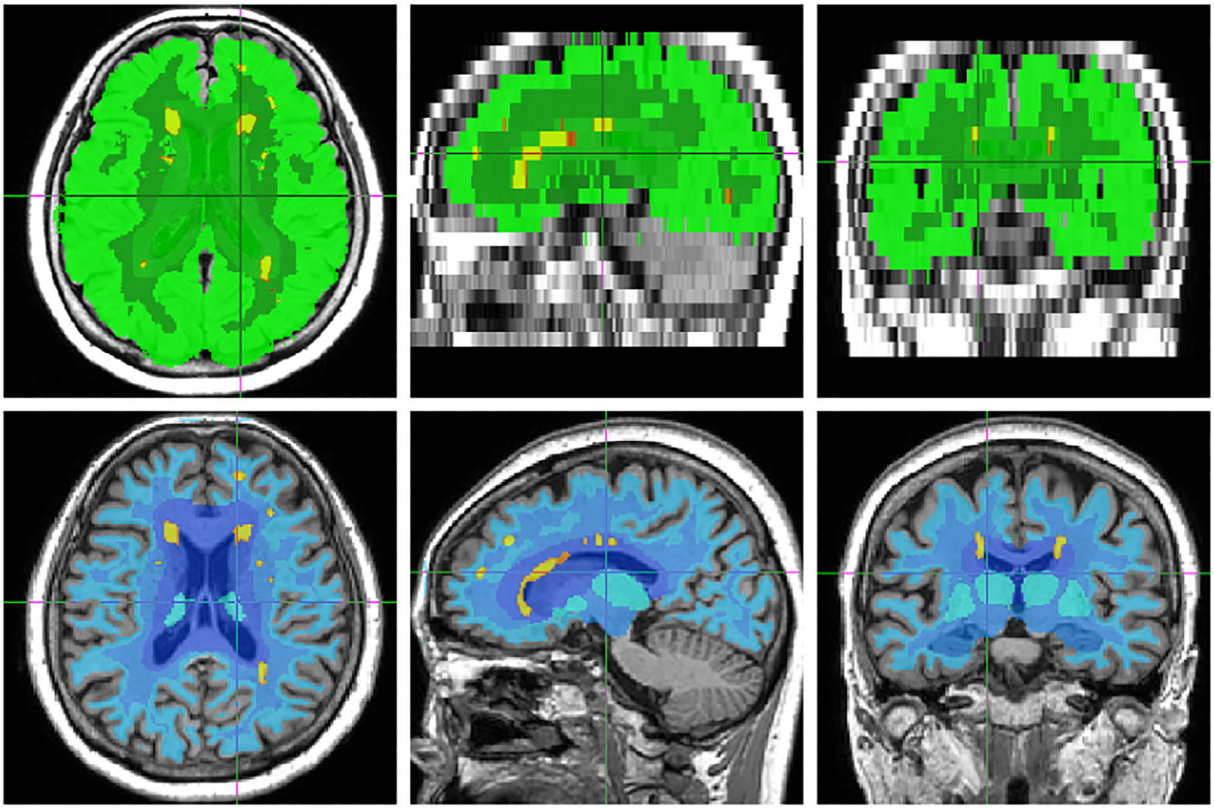

FIGURE 3 Two separate ConvNets were tasked with generating relevant regions of interest (ROIs) represented as color shades for 2D FLAIR (top, greens) and for T1 images (bottom, blues). The FLAIR model, used for temporary usage, was only approximate because of the poor resolution and lack of clearly visible structures. The final model, on T1, was intended to classify lesions (identified and depicted as yellow patches). Still, the hard borders of the ROIs make the current aggregation method perfectible.

set served two purposes. First, to organize the dataset in terms of lesion load, to ease data exploration, and to spot outliers or other unusual cases of interest. Second, to create the final, clinically relevant summary of lesion distribution. Since these two purposes applied, respectively, to the FLAIR and the T1 modality, which have different constraints, two separate models were actually trained (Figure 3), because we would not benefit from having a single multi-modal model (Zopes et al., 2021) here.

We aimed to follow the common practice and guidelines (Thompson et al., 2018), which separate the WMH in periventricular, deep white, or juxtacortical (infra-cortical) lesions. For the periventricular ROIs, we used a 9 mm-expanded mask of the ventricles in MNI space, a somewhat arbitrary visually defined limit. We noted that (Coupé et al., 2018) used 3 mm in MNI space but also observed that this failed to adequately cover the periventricular lesions for a large number of our elderly subjects. Nevertheless, on a Fazekas scoring experiment, the PWML (periventricular) class was the most inconsistently rated, which suggests that a simple ROI approach may never be fully satisfactory.

The ROIs were generated using an initial combination of FreeSurfer label masking, modality co-registrations, and morphometric algorithms—all of which were relatively slow. This initial outcome was then itself learned by two dedicated ConvNet, one per modality, using 4096 subjects. This not only benefited the eventual runtime speed but also allowed to recover from algorithmic failures that may occur during the ROI creation process.

One problem worth mentioning is that FreeSurfer, or other similar tools, may sometimes get confused about pathologies in the white matter, and may not explicitly label them as white matter because of their unusual intensities, even when relying on spatial priors. However, for our purpose, the ROIs should cover the whole white matter, without any holes. Furthermore, it is also known that some WML can evolve into lacunar infarcts (de Jong et al., 2002; Gouw et al., 2008). Therefore, to overcome this systematic source of error, our T1-ROI model was trained on images with simulated lacunes that were generated by randomly scattering the white matter of the input images with small dark-intensity patches.

2.8 | Evaluation

To evaluate the trained model, we first ran a large scale analysis on the training set. We compare the mask of lesions estimated from T1W images against the LST-estimated lesions from the FLAIR images, computing the DICE coefficient as a measure of overlap. Second, to give a clinical perspective, a manual expert

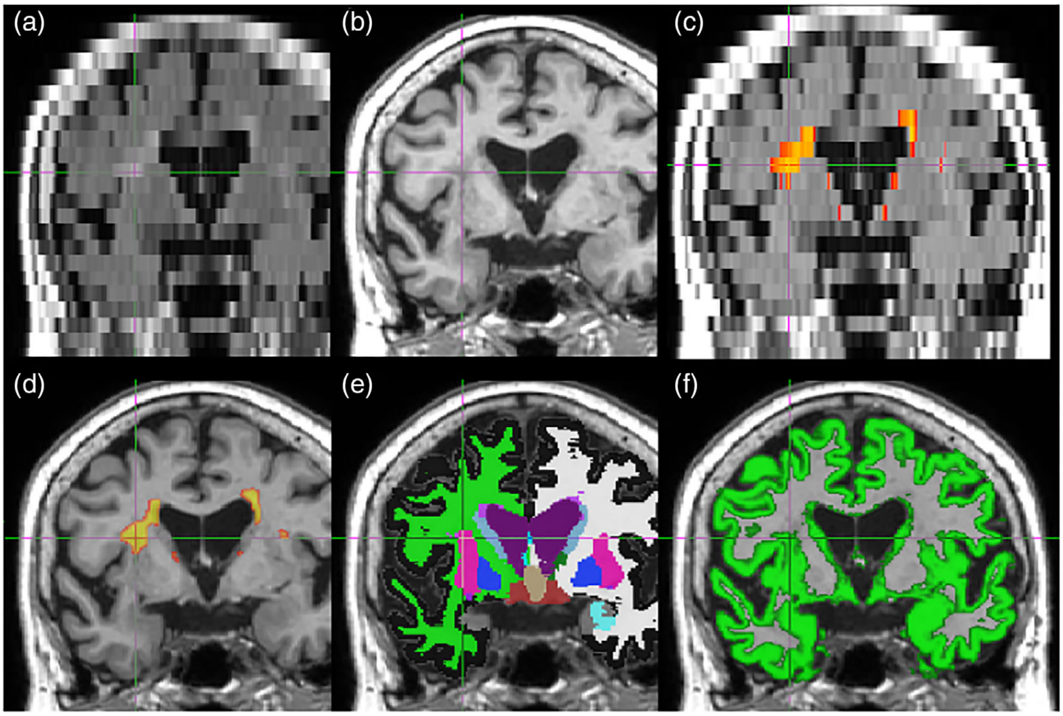

FIGURE 4 Methods overview. On this coronal slice, a periventricular lesion expands toward the insular cortex. The lesion is visible on the original 2D-FLAIR image (a), and on the higher-resolution T1 image (b) although with less contrast. The LST algorithm outcome (c) correctly identified the lesion on the FLAIR image. Our trained ConvNet successfully identified a similar area (d), although at the higher T1 resolution, despite the contrast ambiguity. However, FreeSurfer (e) misclassified part of the lesion as the putamen, while SPM (f) wrongly included it in its gray-matter mask.

rating is conducted on a subset of 200 test image, using two expert raters.

2.9 | Computational tools

Most of the spatial transformations and image processing were computed using ANTs (Avants et al., 2009), although the modality co-registration was conducted with NiftyReg. We also used the c3d tool from ITK-Snap. The MRIs were processed on FreeSurfer 7.1 on four multi-core machines, while the FLAIR images were processed using LST on MATLAB. The various analyses and image processing scripts relied on Python 3, NiBabel, and the SciPy stack. The data navigators and rating interface relied on Matplotlib and other web technologies. PyTorch (1.6 then 1.8) was used for the implementation of ConvNet, conducted on an NVidia RTX 2048 GPU using the ADAM optimizer.

3 | RESULTS

3.1 | Qualitative example

Figure 4 depicts a single slice of MRI images of a random participant under various conditions. The resolution gap between the FLAIR and T1 sequences is clearly visible on this coronal view. The contrast between the lesion and the gray matter is much lower on the T1, which explains why it can be difficult for both people and classic image processing algorithms to identify it accurately.

3.2 | Effect of the loss function

Figure 5 illustrates the effect of the sparse feedback on the resolution, using a simple contrast-transfer task. In this experiment, a ConvNet must produce a fake image of FLAIR contrast corresponding to an input T1 image. The training is conducted using pairs of high-resolution 3D T1 images and 2D FLAIR images, using either a standard voxelwise interpolating loss or a spatially sparse loss. Once trained, visual inspection reveals that while the former appears to learn spurious aliasing artifacts, the latter can preserve more spatial features of the input T1 along the axial axis.

3.3 | Systematic automated comparison

We ran a large-scale automated comparison against FreeSurfer, using the FLAIR image as a reference. The lesion ConvNet was trained on 4096 subjects, and the 3598 remaining images comprised the test set. Each T1 image in the test set was analyzed for WMH by both FreeSurfer 7.1 and our ConvNet. Next, the segmentation masks of

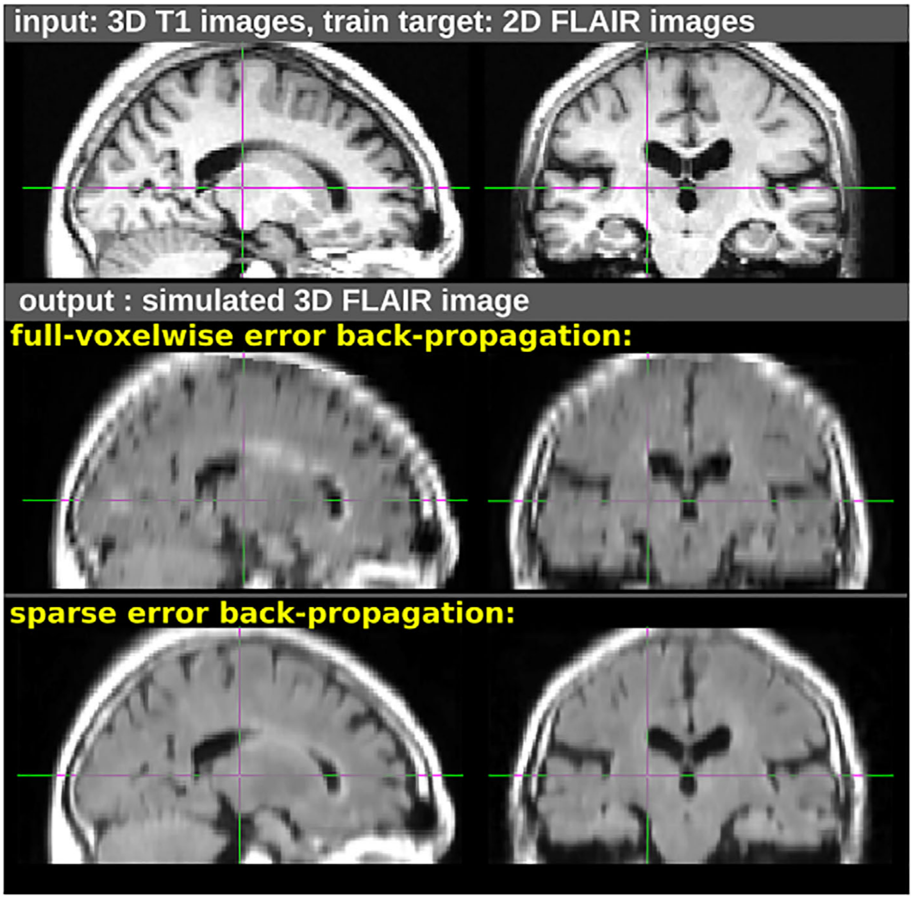
input: 3D T1 images, train target: 2D FLAIR images
output : simulated 3D FLAIR image
full-voxelwise error back-propagation:
sparse error back-propagation:

FIGURE 5 Illustration of the effect of two loss functions, used here for a cross-contrast transformation experiment. In this example, a ConvNet was trained to change an input T1-weighted image (top) into its corresponding FLAIR contrast, based on 2048 training pairs. The MRI sequences used for the target FLAIR contrast have a lower native vertical resolution. Using a standard voxelwise loss function (middle), the model over-learned the interpolation artifacts. However, using the spatially sparse feedback-signal loss (bottom), the model successfully learned the contrast transform while maintaining the resolution.

both methods were resampled into the original FLAIR resolution and a DICE metric computed against the LST-segmented FLAIR. We thresholded the probabilistic maps at 50%, although the actual threshold value did not significantly alter the overall result.

Figure 6 shows the DICE distribution for each method with the samples ordered by lesion volume. We note that the overall DICE metrics are low, particularly for small lesion loads, but this does not preclude us from using the metric for large-scale observation purposes. This is due to the transfer-learning across modalities, which reveals the difficulty of precisely locating the lesions and agreeing on their boundaries in true anatomical space. Our ConvNet segmentation was consistently better than FreeSurfer 7 ($paired\text{-}t_{\_[3598]} = 116.0$, $p = 0$), although 1.63% of the test set showed a better DICE against LST for FreeSurfer. Naturally, part of ConvNet's success can be attributed to its more explicitly chosen goal of segmenting a T1 in a way that resembles a FLAIR LST delineation.

A selection of extreme points are highlighted to illustrate the variety of patterns that can occur. Low DICE values may be attributed to the failure to segment FLAIR itself due to motion or intensity threshold (e.g., top row). Other causes of low DICE were due to the plain lack of lesions, where pixel-perfect segmentation is rarely achieved. The best DICE scores were obtained when the lesion load was heavy, although those cases are not necessarily the most relevant from a clinical perspective, since it is often already too late for the subject to positively alter the course of WMH evolution.

3.4 | Human visual rating

While certainly useful, the lesion mask obtained from using LST on FLAIR also has limitations, even when disregarding putative cases of algorithmic errors. To pursue the evaluation further, a random subset of the dataset was selected for a visual investigation.

In order to submit a relatively large sample to visual judgment, we designed a web rating user interface that enabled a three-way comparison between pictures of the three methods and allowed the expert to pick the best and worst proposal, in a double-blinded setup.

We found that such a system of comparison was more suitable than an absolute-scale rating requirement, which proved hard to keep consistent across a large number of subjects, let alone across raters. This also proved more reasonable in terms of the effort required by the human experts, than an alternative like producing manual segmentations.

A set of 200 subjects were randomly selected from the test set, and pictures of each segmentation (ConvNet, FS7, and original LST) were generated, thresholding the masks when appropriate (LST and ConvNet). The model outcome was reviewed by two certificated radiologists (Y. Tat and L. C) over their original FLAIR slices. This had the following advantages: the expert raters were already used to inspecting FLAIR image rather than 3D T1; images; the FLAIR images had fewer axial slices and therefore needed less visual effort to review a subject; and some interpolation artifacts could have otherwise

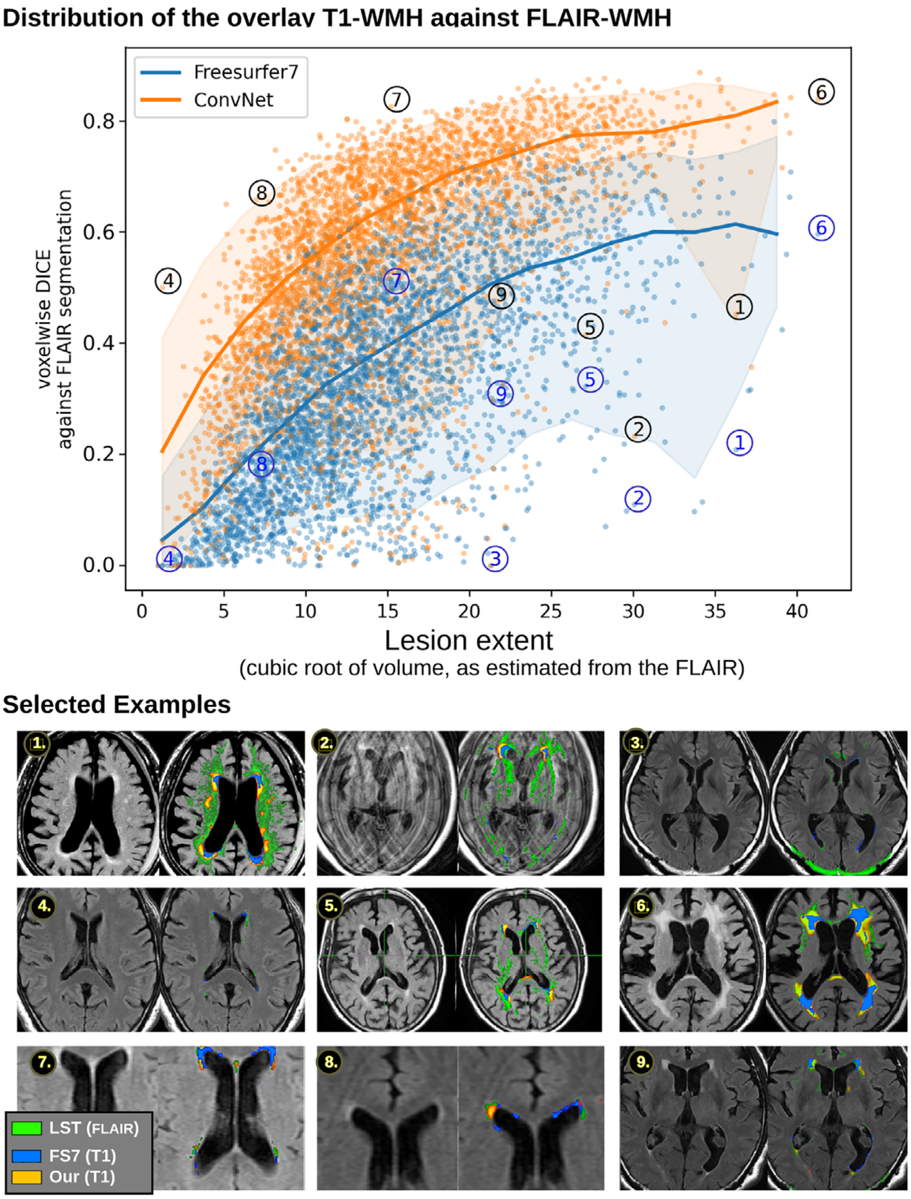

FIGURE 6 Agreement between the T1-based segmentation methods and the FLAIR segmentation method (top). The graph shows the distributions of the overlap as a DICE metric calculated in the resampled FLAIR space and ordered by lesion volume. Colors encode the two methods, and the filled areas contain 90% of the subjects at each volume bin. The mean/SD DICE is 0.33/0.18 (FS7) and 0.55/0.19 (our). Selected MRI cases (bottom), corresponding to numbers on the plot, are selected to illustrate different occurring patterns of overlap.

revealed which image comes from the FLAIR space (i.e., LST method), thereby partially nullifying the double-blind setup.

The two human raters independently rated the same 200 subjects. Figure 7 depicts the number of times each method was ranked the best. Some differences were observed among the two raters ($\chi^2 = 7.53$, $p = 0.023$). When raters agreed on a positive rating (152 out of 200 subjects), 91% were in favor of our ConvNet. Furthermore, when selecting LST as the best segmentation, their ratings were unanimous for only 18% of the images, revealing the role of subjectivity in that choice. Finally, the raters largely agreed that the FreeSurfer segmentations were generally less accurate, at least in our setup. These method choices appeared to be irrespective of the lesion volume (i.e., the rater choices were insignificant as predictors of the volume, $F_{-[3196]} = 0.85$, $p = 0.46$), even though one rater reported that the rating task was easier and faster for images with larger lesion loads.

4 | DISCUSSION

We endeavored to train a ConvNet to extract WMH from the T1 images. Obtaining a ground truth was not envisionable beyond a small sample due to the large annotation effort required (and the lack of such public datasets, with a compatible license). Instead, as a starting point, we used cross-modality learning to benefit from the algorithms that already existed for FLAIR images. A ConvNet was trained to learn the dataset, and some of the most difficult cases, selected by an informed-guess strategy, were hand-fixed and returned back to the training loop. Next, those cleaned masks of FLAIR space were used as targets for the training of our main T1W ConvNet across contrasts, using a loss function intended to overcome the resolution difference. We evaluated the results with an automated region-overlap comparison against other methods, then with expert rating.

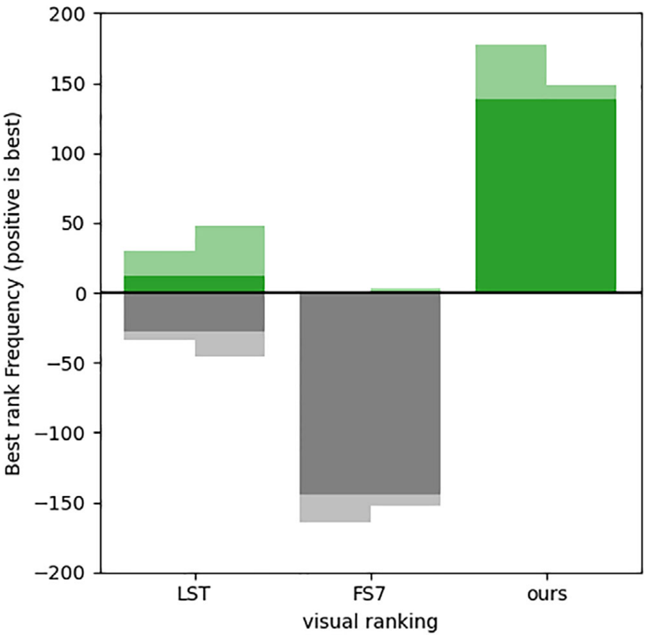

FIGURE 7 Results of the human expert evaluation. The bar chart shows the number of times the method result was considered as the best (or, for negative, worst) by the two raters (sub-columns). The darkest areas represent segmentations where the two raters' judgments agreed. The results may not sum exactly to 200 as ties were allowed, albeit discouraged.

The lack of ground truth was not only a technical matter, however, as we observed disagreement among the experts related to qualitative description of the lesions. In fact, on a Fazekas rating effort of 200 subjects from a single site, two experts from different hospitals obtained an agreement (same rank on the 4-level qualitative scale) of only 62.6% for deep-white lesions, and as low as 38% for the periventricular lesions. Although reportedly not always necessary (Boutet et al., 2016), our results suggest the need for some preliminary coordination before using the scale. In the design of our visual evaluations, the two raters largely agreed across an equally sized set of 200 subjects.

4.1 | The use of T1 images

In this study, we targeted T1-weighted images. WMH lesions had been associated, with various levels of confidence, with multiple brain pathologies. Therefore, it is useful to access this information using more than the FLAIR modality alone, particularly in the context of cohort research, which often acquires a T1 image for structure localization and measurement purpose. In fact, the JPSC-AD large-cohort project is such a case of a project that did not enforce a FLAIR acquisition at all the participating sites. As another example, (Tanaka et al., 2021) studied a large cohort of various ages, but only acquired T1 images rather than FLAIR.

Additionally, (Wei, Poirion, et al., 2019; Wei, Tran, et al., 2019) compared the predictive power of both T1 and FLAIR lesions—respectively segmented with FreeSurfer 6 and SPM-LST—and concluded a "general equivalence between these two," based on a similar association with abnormal β-amyloid and tau, although other researchers (Hotz et al., 2021) assessed that WMH lesions labeled by FreeSurfer v6.0.1 are no substitute for manual effort. Still, a T1 image provides potentially more spatial precision, which could prove useful for accurate longitudinal monitoring of early smaller lesions, particularly with tailored methods (Sudre et al., 2017).

4.2 | Cross-domain learning

Initially, we tried to work in T1 space directly, by using the masks generated by FreeSurfer as a training signal. Unfortunately, this method turned out to be too unreliable, despite the ability of ConvNet to be robust against occasional mislabeling in the training set (Thyreau & Taki, 2020). We observed that the model initially learned to segment a majority of the WMH, and then, instead of refining its learning, appeared to unlearn whole classes of lesions as more training iterations progressed and more inconsistent samples were presented. We had to cease pursuing this initially attractive approach of copying FreeSurfer T1-processed images, and instead developed the new cross-contrast learning model described here.

We proposed a transfer-learning approach where the set of 2D FLAIR images would provide sufficient information to train a T1 classifier. A spatially sparse loss mostly solved the resolution issue. The reason for this was because even though each image contains little information individually, the whole training set was large. Such a spatially sparse feedback loss is suitable for any similar cross-modality learning setup where a super-resolution approach of the target is required, due to images having widely different resolutions, by themselves or by the corresponding labels (e.g., to reduce the annotation effort). Learning from sparser annotations is often a worthwhile goal in the field of medical image segmentation.

4.3 | ConvNets for the WMH lesions

The ConvNet research work has often focused on the small sample regime and architecture improvement (Fartaria et al., 2016; Ghafoorian et al., 2017; Guerrero et al., 2018; La Rosa et al., 2020; Li et al., 2018; Li et al., 2021; Liang et al., 2021; Moeskops et al., 2018; Orbes-Arteaga et al., 2018; Valverde et al., 2017). Recently (Kuijf et al., 2019) organized a competition, using 60 (3D or thin-sliced 2D) FLAIR images as training set, and 110 manually labeled subjects for evaluation. Here again, the focus was squarely on the ability of the method to generalize from the provided sample. Notably, ConvNets ranked favorably, although the organizers pointed that that three-dimensional ConvNets achieved lower-ranking results, which they attributed to the 2D nature of FLAIR images. They also noted that the moderate recall of the individual observers is mainly caused by either not segmenting or missing small WMHs (Kuijf et al., 2019). We hope that the current work improved on that situation.

Finally, we note that ConvNets provide attractive runtime properties. They offer robust results with a constant-time inference. Our current implementation requires around 20 s to fully process a T1 image, using four threads of a recent CPU and can consume up to 4Gb peak memory. However, the network topology of our model was partly defined by a practical concern: the maximum size that could allow a single sample to fit in the GPU memory for a forward and a backward pass, at the useful input resolution of 1 × 1 × 1.5 mm, and with the added cortical-ribbon segmentation requirement. This is because we wanted to focus on the dataset learning behavior, without worrying about additional failures attributable to a more limited model capacity. A logical future work would aim to reduce the model size, using one of the many model reduction techniques proposed in the field.

4.4 | Limitations of the final model

To validate our model, we ran an automated evaluation based on the similarity with FLAIR masks, and compared it to the FreeSurfer 7 model. There are some limitations in this approach. The DICE coefficient was conducted in FLAIR space, considered as a rough ground-truth, and therefore, would not capture the fine lesion details along the slice axis. And as mentioned above, the FreeSurfer 7 output was not always of satisfying quality. Importantly, we observed that the DICE value was generally higher when the lesion volume was bigger. This phenomenon is in part because, in the DICE computation, the relative impact of each missed voxel is stronger for small segmentations, but also because the numerous smaller lesions are less consensual over different methods, or even different human ratings. To highlight this limitation, we reported the DICE coefficients as a function of the lesion load.

The visual evaluation showed segmentation in FLAIR space even when created from T1 images. Therefore, the accuracy was not judged explicitly in the axial direction. The segmentation quality would probably be evaluated even more favorably when seen through coronal or sagittal slices; however, we aimed for a clinically realistic comparison. Overall, the limited axial resolution in many FLAIR sequences is generally not perceived as a major hindrance for the descriptive purposes of clinical reviews. It becomes a difficulty only when dealing with quantification. Some successful quantitative algorithms of FLAIR images actually rely on an additional structural image to compensate for the loss of localization power, and they are generally only used for research.

Finally, we note that the lesion delineation is necessarily sensitive to the source image contrast and the threshold chosen, as often with MRI image quantitative analyses. Our ConvNet has been trained on the variety of scanner of the JPSC-AD cohort, and should therefore generalize to most T1 images (see also Appendix S1 for an application over the OASIS 3 dataset [(LaMontagne et al., 2019]); However, should high accuracy be required, the lesion volume should not be compared between different contrasts without care, as the actual absolute volume value may impacted by the acquisition sequence.

5 | CONCLUSION

In this manuscript, we described the training procedure of a ConvNet whose goal was to segment WMHs from T1 images, and which constitutes the core component of a corresponding end-user software, which is available for download. The training of the T1 model had been achieved by transferring some information that is more reliably found in FLAIR images, enhancing it using a limited amount of human guidance, and making adjustments to the loss function. The large JPSC-AD dataset used as training material was key to the robustness. This study provides another good example of the power of ConvNet from the perspective of working across domains to solve problems more optimally.

AUTHOR CONTRIBUTIONS

Benjamin Thyreau and Yasuko Tatewaki designed research; Benjamin Thyreau developed the method; Yasuko Tatewaki, Liying Chen analyzed the images; Yasuko Tatewaki, Liying Chen, Naoki Hirabayashi, Yoshihiko Furuta and Jun Hata performed clinical evaluation; Yuji Takano, Naoki Hirabayashi, Yoshihiko Furuta, Jun Hata, Shigeyuki Nakaji, Tetsuya Maeda, Moeko Noguchi-Shinohara, Masaru Mimura, Kenji Nakashima, Takaaki Mori, Minoru Takebayashi collected data; Yasuyuki Taki and Toshiharu Ninomiya coordinated the project; Benjamin Thyreau, Yasuko Tatewaki and Liying Chen wrote the paper with inputs from all authors.

ACKNOWLEDGMENTS

We would like to thank all researchers and investigators in the JPSC-AD Study Group for their diligent work and contribution. We would like to additionally thank participants for volunteering their time to take part in the JPSC-AD study. Participating institutes and principal collaborators for the JPSC-AD Study group: Kyushu University ([Epidemiology and Public Health] Toshiharu Ninomiya, Jun Hata, Mao Shibata, and Takanori Honda, [Neuropsychiatry] Tomoyuki Ohara, [Pathology and Imaging Science] Masato Akiyama); Hirosaki University (Shigeyuki Nakaji, Koichi Murashita, Tatsuya Mikami, Songee Jung, and Mina Misawa); Iwate Medical University (Tetsuya Maeda, Naoki Ishizuka, and Hiroshi Akasaka); Shonan Keiiku Hospital (Yasuo Terayama); Japanese Red Cross Morioka Hospital (Hisashi Yonezawa); Morioka Tsunagi Onsen Hospital (Junko Takahashi); Kanazawa University (Moeko Noguchi-Shinohara, Masahito Yamada, Kazuo Iwasa, Sohshi Yuki-Nozaki, and Kenjiro Ono); Keio University School of Medicine (Masaru Mimura, Shogyoku Bun, Hidehito Niimura, Ryo Shikimoto, and Hisashi Kida); Matsue Medical Center (Kenji Nakashima, Yasuyo Fukada, and Hisanori Kowa); Kawasaki Medical School (Kenji Wada); Tottori Prefectural Kousei Hospital (Masafumi Kishi); Ehime University (Takaaki Mori, Yuta Yoshino, Hideaki Shimizu, Ayumi Tachibana, and Shu-ichi Ueno); Kumamoto University (Minoru Takebayashi, Tomohisa Ishikawa, Ryuji Fukuhara, and Asuka Koyama); Osaka University Medical School (Mamoru Hashimoto and Manabu Ikeda); National Cerebral and Cardiovascular Center) (Yoshihiro Kokubo); Nakamura-Gakuen University (Kazuhiro Uchida, and Midori Esaki); Tohoku University (Yasuyuki Taki, Yuji Takano, Yasuko

Tatewaki, and Benjamin Thyreau);University of the Ryukyu (Koji Yonemoto); Osaka City University Graduate School of Medicine (Hisako Yoshida); University of Tokyo (Kaori Muto, Yusuke Inoue); RIKEN Center for Integrative Medical Sciences (Yukihide Momozawa and Chikashi Terao); Hisayama Research Institute for Lifestyle Diseases (Michiaki Kubo and Yutaka Kiyohara).

FUNDING INFORMATION

The JPSC-AD study was supported by the Japan Agency for Medical Research and Development (JP21dk0207053) and Suntory Holdings Limited (Osaka, Japan). The funders had no role in the design of the study, the collection, analysis, and interpretation of data, or the writing of the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The trained model available for download at https://github.com/bthyreau/deep-T1-WMH. The datasets generated and/or analyzed during the current study are not publicly available due to restrictions included in the informed consent of research participants. However, data are available from the authors upon reasonable request and with the permission of the Japan Agency for Medical Research and Development.

ETHICS STATEMENT

JPSC-AD was first approved by the Kyushu University Institutional Review Board and then by the ethics committee of each research institute. All participants provided written informed consent for baseline and follow-up surveys.

ORCID

Benjamin Thyreau https://orcid.org/0000-0001-6748-8192

REFERENCES

Alber, J., Alladi, S., Bae, H. J., Barton, D. A., Beckett, L. A., Bell, J. M., Berman, S. E., Biessels, G. J., Black, S. E., Bos, I., Bowman, G. L., Brai, E., Brickman, A. M., Callahan, B. L., Corriveau, R. A., Fossati, S., Gottesman, R. F., Gustafson, D. R., Hachinski, V., … Hainsworth, A. H. (2019). White matter hyperintensities in vascular contributions to cognitive impairment and dementia (VCID): Knowledge gaps and opportunities. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *5*, 107–117. https://doi.org/10.1016/j.trci.2019.02.001

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronné, R., Faouzi, J., Koval, I., Louis, M., Thibeau-Sutre, E., Wen, J., Wild, A., Burgos, N., Dormont, D., Colliot, O., & Durrleman, S. (2021). Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Medical Image Analysis*, *67*, 101848. https://doi.org/10.1016/j.media.2020.101848

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., Lacoste-Julien, S., 2017. A closer look at memorization in deep networks. ICML'17 Proc. 34th Int. Conf. Mach. Learn. 70, 233–242.

Atwi, S., Metcalfe, A. W. S., Robertson, A. D., Rezmovitz, J., Anderson, N. D., & MacIntosh, B. J. (2018). Attention-related brain activation is altered in older adults with white matter Hyperintensities using multi-Echo fMRI. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00748

Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS). *The Insight Journal*, *2*, 1–35.

Balakrishnan, R., Valdés Hernández, M. C., & Farrall, A. J. (2021). Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data: A systematic review. *Computerized Medical Imaging and Graphics*, *88*, 101867. https://doi.org/10.1016/J.COMPMEDIMAG.2021.101867

Boutet, C., Rouffiange-Leclair, L., Schneider, F., Camdessanché, J.-P., Antoine, J.-C., & Barral, F.-G. (2016). Visual assessment of age-related white matter hyperintensities using FLAIR images at 3 T: inter- and intra-rater agreement. *Neurodegenerative Diseases*, *16*, 279–283. https://doi.org/10.1159/000441420

Cedres, N., Ferreira, D., Machado, A., Shams, S., Sacuiu, S., Waern, M., Wahlund, L. O., Zettergren, A., Kern, S., Skoog, I., & Westman, E. (2020). Predicting Fazekas scores from automatic segmentations of white matter signal abnormalities. *Aging (Albany. NY)*, *12*, 894–901. https://doi.org/10.18632/aging.102662

Coupé, P., Tourdias, T., Linck, P., Romero, J. E., & Manjón, J. V. (2018). *LesionBrain: An online tool for white matter lesion segmentation*. In Patch-based techniques in medical imaging (pp. 95–103). https://doi.org/10.1007/978-3-030-00500-9_11

de Jong, G., Kessels, F., & Lodder, J. (2002). Two types of lacunar infarcts: Further arguments from a study on prognosis. *Stroke*, *33*, 2072–2076. https://doi.org/10.1161/01.STR.0000022807.06923.A3

Debette, S., & Markus, H. S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ*, *341*, c3666. https://doi.org/10.1136/bmj.c3666

Debette, S., Schilling, S., Duperron, M. G., Larsson, S. C., & Markus, H. S. (2019). Clinical significance of magnetic resonance imaging markers of vascular brain injury: A systematic review and meta-analysis. *JAMA Neurology*, *76*, 81–94. https://doi.org/10.1001/jamaneurol.2018.3122

el Jurdi, R., Petitjean, C., Honeine, P., Cheplygina, V., & Abdallah, F. (2021). High-level prior-based loss functions for medical image segmentation: A survey. In *A surprisingly effective perimeter-based loss for medical image segmentation*. Medical Imaging with Deep Learning.

Fartaria, M. J., Bonnier, G., Roche, A., Kober, T., Meuli, R., Rotzinger, D., Frackowiak, R., Schluep, M., Du Pasquier, R., Thiran, J.-P., Krueger, G., Bach Cuadra, M., & Granziera, C. (2016). Automated detection of white matter and cortical lesions in early stages of multiple sclerosis. *Journal of Magnetic Resonance Imaging*, *43*, 1445–1454. https://doi.org/10.1002/jmri.25095

Fazekas, F., Chawluk, J., Alavi, A., Hurtig, H., & Zimmerman, R. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Roentgenology*, *149*, 351–356. https://doi.org/10.2214/ajr.149.2.351

Fazekas, F., Kleinert, R., Offenbacher, H., Schmidt, R., Kleinert, G., Payer, F., Radner, H., & Lechner, H. (1993). Pathologic correlates of incidental mri white matter signal hyperintensities. *Neurology*, *43*, 1683–1689. https://doi.org/10.1212/wnl.43.9.1683

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Frey, B. M., Petersen, M., Mayer, C., Schulz, M., Cheng, B., & Thomalla, G. (2019). Characterization of white matter hyperintensities in large-scale MRI-studies. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00238

Ghafoorian, M., Karssemeijer, N., Heskes, T., Van Uden, I. W. M., Sanchez, C. I., Litjens, G., De Leeuw, F. E., Van Ginneken, B., Marchiori, E., & Platel, B. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, *7*, 1–12. https://doi.org/10.1038/s41598-017-05300-5

Gouw, A. A., van der Flier, W. M., Fazekas, F., van Straaten, E. C. W., Pantoni, L., Poggesi, A., Inzitari, D., Erkinjuntti, T., Wahlund, L. O., Waldemar, G., Schmidt, R., Scheltens, P., & Barkhof, F. (2008). Progression of white matter Hyperintensities and incidence of new lacunes over a 3-year period. *Stroke*, *39*, 1414–1420. https://doi.org/10.1161/STROKEAHA.107.498535

Griffanti, L., Jenkinson, M., Suri, S., Zsoldos, E., Mahmood, A., Filippini, N., Sexton, C. E., Topiwala, A., Allan, C., Kivimäki, M., Singh-Manoux, A., Ebmeier, K. P., Mackay, C. E., & Zamboni, G. (2018). Classification and characterization of periventricular and deep white matter hyperintensities on MRI: A study in older adults. *NeuroImage*, *170*, 174–181. https://doi.org/10.1016/j.neuroimage.2017.03.024

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M., & Jenkinson, M. (2016). BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, *141*, 191–205. https://doi.org/10.1016/j.neuroimage.2016.07.018

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M. C., Dickie, D. A., Wardlaw, J., & Rueckert, D. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, *17*, 918–934. https://doi.org/10.1016/j.nicl.2017.12.022

Habes, M., Sotiras, A., Erus, G., Toledo, J. B., Janowitz, D., Wolk, D. A., Shou, H., Bryan, N. R., Doshi, J., Völzke, H., Schminke, U., Hoffmann, W., Resnick, S. M., Grabe, H. J., & Davatzikos, C. (2018). White matter lesions spatial heterogeneity, links to risk factors, cognition, genetics, and atrophy. *Neurology*, *91*, E964–E975. https://doi.org/10.1212/WNL.0000000000006116

Heinen, R., Steenwijk, M. D., Barkhof, F., Biesbroek, J. M., van der Flier, W. M., Kuijf, H. J., Prins, N. D., Vrenken, H., Biessels, G. J., de Bresser, J., van den Berg, E., Biessels, G. J., Boomsma, J. M. F., Exalto, L. G., Ferro, D. A., Frijns, C. J. M., Groeneveld, O. N., Heinen, R., van Kalsbeek, N. M., … Vriens, E. (2019). Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset. *Scientific Reports*, *9*, 1–12. https://doi.org/10.1038/s41598-019-52966-0

Hotz, I., Deschwanden, P. F., Liem, F., Mérillat, S., Malagurski, B., Kollias, S., & Jäncke, L. (2021). Performance of three freely available methods for extracting white matter hyperintensities: FreeSurfer , UBO Detector, and BIANCA. *Human Brain Mapping*, *43*(5), 1481–1500. https://doi.org/10.1002/hbm.25739

Jiang, J., Liu, T., Zhu, W., Koncz, R., Liu, H., Lee, T., Sachdev, P. S., & Wen, W. (2018). UBO detector: A cluster-based, fully automated pipeline for extracting white matter hyperintensities. *NeuroImage*, *174*, 539–549. https://doi.org/10.1016/J.NEUROIMAGE.2018.03.050

Koikkalainen, J. R., Rhodius-Meester, H. F. M., Frederiksen, K. S., Bruun, M., Hasselbalch, S. G., Baroni, M., Mecocci, P., Vanninen, R., Remes, A., Soininen, H., van Gils, M., van der Flier, W. M., Scheltens, P., Barkhof, F., Erkinjuntti, T., & Lötjönen, J. M. P. (2019). Automatically computed rating scales from MRI for patients with cognitive disorders. *European Radiology*, *29*, 4937–4947. https://doi.org/10.1007/s00330-019-06067-1

Kuijf, H. J., Casamitjana, A., Collins, D. L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Biesbroek, J. M., Luna, M., Mahmood, Q., Mckinley, R., Mehrtash, A., Ourselin, S., Park, B. Y., Park, H., … Cardoso, M. J. (2019). Standardized assessment of automatic segmentation of white matter Hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, *38*, 2556–2568. https://doi.org/10.1109/TMI.2019.2905770

La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P.-J., Galbusera, R., Barakovic, M., Thiran, J.-P., Granziera, C., & Cuadra, M. B. (2020). Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: A deep learning method based on FLAIR and MP2RAGE. *NeuroImage Clinical*, *27*, 102335. https://doi.org/10.1016/j.nicl.2020.102335

LaMontagne, P. J., Benzinger, T. L. S., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., Raichle, M. E., Cruchaga, C., & Marcus, D. (2019). OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *MedRxiv*, 19014902. https://doi.org/10.1101/2019.12.13.19014902

Lampe, L., Kharabian-Masouleh, S., Kynast, J., Arelin, K., Steele, C. J., Löffler, M., Witte, A. V., Schroeter, M. L., Villringer, A., & Bazin, P. L. (2019). Lesion location matters: The relationships between white matter hyperintensities on cognition in the healthy elderly. *Journal of Cerebral Blood Flow and Metabolism*, *39*, 36–43. https://doi.org/10.1177/0271678X17740501

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W. S., & Menze, B. (2018). Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage*, *183*, 650–665. https://doi.org/10.1016/j.neuroimage.2018.07.005

Li, X., Zhao, Y., Jiang, J., Cheng, J., Zhu, W., Wu, Z., Jing, J., Zhang, Z., Wen, W., Sachdev, P. S., Wang, Y., Liu, T., & Li, Z. (2021). White matter hyperintensities segmentation using an ensemble of neural networks. *Human Brain Mapping*, *43*, 929–939. https://doi.org/10.1002/hbm.25695

Liang, L., Zhou, P., Lu, W., Guo, X., Ye, C., Lv, H., Wang, T., & Ma, T. (2021). An anatomical knowledge-based MRI deep learning pipeline for white matter hyperintensity quantification associated with cognitive impairment. *Computerized Medical Imaging and Graphics*, *89*, 101873. https://doi.org/10.1016/J.COMPMEDIMAG.2021.101873

Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., & Ourselin, S. (2010). Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, *98*, 278–284. https://doi.org/10.1016/j.cmpb.2009.09.002

Moeskops, P., de Bresser, J., Kuijf, H. J., Mendrik, A. M., Biessels, G. J., Pluim, J. P. W., & Išgum, I. (2018). Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. *NeuroImage Clinical*, *17*, 251–262. https://doi.org/10.1016/J.NICL.2017.10.007

Ngai, S., Tang, Y. M., Du, L., & Stuckey, S. (2007). Hyperintensity of the precentral gyral subcortical white matter and hypointensity of the precentral gyrus on fluid-attenuated inversion recovery: Variation with age and implications for the diagnosis of amyotrophic lateral sclerosis. *American Journal of Neuroradiology*, *28*, 250–254.

Ni, L., Zhang, B., Yang, D., Qin, R., Xu, H., Ma, J., Shao, P., & Xu, Y. (2021). Lower cerebrovascular reactivity contributed to white matter hyperintensity-related cognitive impairment: A resting-state functional MRI study. *Journal of Magnetic Resonance Imaging*, *53*, 703–711. https://doi.org/10.1002/jmri.27376

Ninomiya, T., Nakaji, S., Maeda, T., Yamada, M., Mimura, M., Nakashima, K., Mori, T., Takebayashi, M., Ohara, T., Hata, J., Kokubo, Y., Uchida, K., Taki, Y., Kumagai, S., Yonemoto, K., Yoshida, H., Muto, K., Momozawa, Y., Akiyama, M., … Kiyohara, Y. (2020). Study design and baseline characteristics of a population-based prospective cohort study of dementia in Japan: The Japan Prospective Studies Collaboration for Aging and Dementia (JPSC-AD). *Environmental Health and Preventive Medicine*, *25*, 64. https://doi.org/10.1186/s12199-020-00903-3

Nogovitsyn, N., Souza, R., Muller, M., Srajer, A., Hassel, S., Arnott, S. R., Davis, A. D., Hall, G. B., Harris, J. K., Zamyadi, M., Metzak, P. D., Ismail, Z., Bray, S. L., Lebel, C., Addington, J. M., Milev, R., Harkness, K. L., Frey, B. N., Lam, R. W., … MacQueen, G. M. (2019). Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants. *NeuroImage*, *197*, 589–597. https://doi.org/10.1016/j.neuroimage.2019.05.017

Orbes-Arteaga, M., Cardoso, M.J., Sørensen, L., Modat, M., Ourselin, S., Nielsen, M., Pai, A., 2018. *Simultaneous synthesis of FLAIR and segmentation of white matter hypointensities from T1 MRIs*. arXiv:1808.06519

Pantoni, L. (2010). Cerebral small vessel disease: From pathogenesis and clinical characteristics to therapeutic challenges. *Lancet Neurology*, *9*, 689–701. https://doi.org/10.1016/S1474-4422(10)70104-6

Pozorski, V., Oh, J. M., Okonkwo, O., Krislov, S., Barzgari, A., Theisen, F., Sojkova, J., Bendlin, B. B., Johnson, S. C., & Gallagher, C. L. (2019).

Cross-sectional and longitudinal associations between total and regional white matter hyperintensity volume and cognitive and motor function in Parkinson's disease. *NeuroImage Clinical*, *23*, 101870. https://doi.org/10.1016/j.nicl.2019.101870

Prins, N. D., & Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: An update. *Nature Reviews Neurology*, *11*, *157–165*. https://doi.org/10.1038/nrneurol.2015.10

Rhodius-Meester, H. F. M., Benedictus, M. R., Wattjes, M. P., Barkhof, F., Scheltens, P., Muller, M., & van der Flier, W. M. (2017). MRI visual ratings of brain atrophy and white matter Hyperintensities across the Spectrum of cognitive decline are differently affected by age and diagnosis. *Frontiers in Aging Neuroscience*, *9*. https://doi.org/10.3389/fnagi.2017.00117

Ribaldi, F., Altomare, D., Jovicich, J., Ferrari, C., Picco, A., Pizzini, F. B., Soricelli, A., Mega, A., Ferretti, A., Drevelegas, A., Bosch, B., Müller, B. W., Marra, C., Cavaliere, C., Bartrés-Faz, D., Nobili, F., Alessandrini, F., Barkhof, F., Gros-Dagnac, H., … Marizzoni, M. (2021). Accuracy and reproducibility of automated white matter hyperintensities segmentation with lesion segmentation tool: A European multi-site 3T study. *Magnetic Resonance Imaging*, *76*, 108–115. https://doi.org/10.1016/j.mri.2020.11.008

Sarica, A., Curcio, M., Rapisarda, L., Cerasa, A., Quattrone, A., & Bono, F. (2019). Periventricular white matter changes in idiopathic intracranial hypertension. *Annals of Clinical Translational Neurology*, *6*, 233–242. https://doi.org/10.1002/ACN3.685

Schirmer, M. D., Dalca, A. V., Sridharan, R., Giese, A.-K., Donahue, K. L., Nardin, M. J., Mocking, S. J. T., McIntosh, E. C., Frid, P., Wasselius, J., Cole, J. W., Holmegaard, L., Jern, C., Jimenez-Conde, J., Lemmens, R., Lindgren, A. G., Meschia, J. F., Roquer, J., Rundek, T., … Rost, N. S. (2019). White matter hyperintensity quantification in large-scale clinical acute ischemic stroke cohorts: The MRI-GENIE study. *NeuroImage Clinical*, *23*, 101884. https://doi.org/10.1016/j.nicl.2019.101884

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V. J., Zimmer, C., Hemmer, B., & Mühlau, M. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, *59*, 3774–3783. https://doi.org/10.1016/j.neuroimage.2011.11.032

Sudre, C. H., Cardoso, M. J., & Ourselin, S. (2017). Longitudinal segmentation of age-related white matter hyperintensities. *Medical Image Analysis*, *38*, 50–64. https://doi.org/10.1016/j.media.2017.02.007

Sundaresan, V., Zamboni, G., Le Heron, C., Rothwell, P. M., Husain, M., Battaglini, M., De Stefano, N., Jenkinson, M., & Griffanti, L. (2019). Automated lesion segmentation with BIANCA: Impact of population-level features, classification algorithm and locally adaptive thresholding. *NeuroImage*, *202*, 116056. https://doi.org/10.1016/j.neuroimage.2019.116056

Tanaka, S. C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., … Imamizu, H. (2021). A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data*, *8*, 227. https://doi.org/10.1038/s41597-021-01004-8

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., … Cohen, J. A. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurology*, *17*, 162–173. https://doi.org/10.1016/S1474-4422(17)30470-2

Thyreau, B., Sato, K., Fukuda, H., & Taki, Y. (2018). Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. *Medical Image Analysis*, *43*, 214–228. https://doi.org/10.1016/J.MEDIA.2017.11.004

Thyreau, B., & Taki, Y. (2020). Learning a cortical parcellation of the brain robust to the MRI segmentation with convolutional neural networks. *Medical Image Analysis*, *61*, 101639. https://doi.org/10.1016/j.media.2020.101639

Tubi, M. A., Feingold, F. W., Kothapalli, D., Hare, E. T., King, K. S., Thompson, P. M., & Braskie, M. N. (2020). White matter hyperintensities and their relationship to cognition: Effects of segmentation algorithm. *NeuroImage*, *206*, 116327. https://doi.org/10.1016/j.neuroimage.2019.116327

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., Rovira, À., Oliver, A., & Lladó, X. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, *155*, 159–168. https://doi.org/10.1016/J.NEUROIMAGE.2017.04.034

Van Straaten, E. C. W., Fazekas, F., Rostrup, E., Scheltens, P., Schmidt, R., Pantoni, L., Inzitari, D., Waldemar, G., Erkinjuntti, T., Mäntylä, R., Wahlund, L. O., & Barkhof, F. (2006). Impact of white matter hyperintensities scoring method on correlations with clinical data: The LADIS study. *Stroke*, *37*, 836–840. https://doi.org/10.1161/01.STR.0000202585.26325.74

Vanderbecq, Q., Xu, E., Ströer, S., Couvy-Duchesne, B., Diaz Melo, M., Dormont, D., & Colliot, O. (2020). Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. *NeuroImage Clinical*, *27*, 102357. https://doi.org/10.1016/j.nicl.2020.102357

Vangberg, T. R., Eikenes, L., & Håberg, A. K. (2019). The effect of white matter hyperintensities on regional brain volumes and white matter microstructure, a population-based study in HUNT. *NeuroImage*, *203*, 116158. https://doi.org/10.1016/j.neuroimage.2019.116158

Wardlaw, J. M., Smith, C., & Dichgans, M. (2019). Small vessel disease: Mechanisms and clinical implications. *Lancet Neurology*, *18*, 684–696. https://doi.org/10.1016/S1474-4422(19)30079-1

Wei, K., Tran, T., Chu, K., Borzage, M. T., Braskie, M. N., Harrington, M. G., & King, K. S. (2019). White matter hypointensities and hyperintensities have equivalent correlations with age and CSF β-amyloid in the nondemented elderly. *Brain and Behavior: A Cognitive Neuroscience Perspective*, *9*, e01457. https://doi.org/10.1002/brb3.1457

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Colliot, O., Stankoff, B., & Ayache, N. (2019). Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis. *Journal of Medical Imaging*, *6*, 1. https://doi.org/10.1117/1.JMI.6.1.014005

Yushkevich, P., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, *31*, 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015

Zopes, J., Platscher, M., Paganucci, S., & Federau, C. (2021). Multi-modal segmentation of 3D brain scans using neural networks. *Frontiers in Neurology*, *0*, 1167. https://doi.org/10.3389/FNEUR.2021.653375

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Thyreau, B., Tatewaki, Y., Chen, L., Takano, Y., Hirabayashi, N., Furuta, Y., Hata, J., Nakaji, S., Maeda, T., Noguchi-Shinohara, M., Mimura, M., Nakashima, K., Mori, T., Takebayashi, M., Ninomiya, T., Taki, Y., & Japan Prospective Studies Collaboration for Aging and Dementia (JPSC-AD) Study Group (2022). Higher-resolution quantification of white matter hypointensities by large-scale transfer learning from 2D images on the JPSC-AD cohort. *Human Brain Mapping*, *43*(13), 3998–4012. https://doi.org/10.1002/hbm.25899