scientific reports

Check for updates

OPEN

SHIVA-CMB: a deep-learning-based robust cerebral microbleed segmentation tool trained on multi-source T2*GRE- and susceptibility-weighted MRI

Ami Tsuchida[1,2], Martin Goubet[3], Philippe Boutinaud[4], Iana Astafeva[1,2], Victor Nozais[4], Pierre-Yves Hervé[4], Thomas Tourdias[5,6], Stéphanie Debette[2] & Marc Joliot[1]✉

Cerebral microbleeds (CMB) represent a feature of cerebral small vessel disease (cSVD), a prominent vascular contributor to age-related cognitive decline, dementia, and stroke. They are visible as spherical hypointense signals on T2*- or susceptibility-weighted magnetic resonance imaging (MRI) sequences. An increasing number of automated CMB detection methods being proposed are based on supervised deep learning (DL). Yet, the lack of open sharing of pre-trained models hampers the practical application and evaluation of these methods beyond specific data sources used in each study. Here, we present the SHIVA-CMB detector, a 3D Unet-based tool trained on 450 scans taken from seven acquisitions in six different cohort studies that included both T2*- and susceptibility-weighted MRI. In a held-out test set of 96 scans, it achieved the sensitivity, precision, and F1 (or Dice similarity coefficient) score of 0.67, 0.82, and 0.74, with less than one false positive detection per image (FPavg = 0.6) and per CMB (FPcmb = 0.15). It achieved a similar level of performance in a separate, evaluation-only dataset with acquisitions never seen during the training (0.67, 0.91, 0.77, 0.5, 0.07 for the sensitivity, precision, F1 score, FPavg, and FPcmb). Further demonstrating its generalizability, it showed a high correlation (Pearson's $R = 0.89$, $p < 0.0001$) with a visual count by expert raters in another independent set of 1992 T2*-weighted scans from a large, multi-center cohort study. Importantly, we publicly share both the pipeline (https://github.com/pboutinaud/SHiVAi/) and pre-tr ained models (https://github.com/pboutinaud/SHIVA-CMB/) to the research community to promote the active application and evaluation of our tool. We believe this effort will help accelerate research on the pathophysiology and functional consequences of CMB by enabling rapid characterization of CMB in large-scale studies.

Cerebral microbleeds (CMB) result from hemosiderin breakdown products left after microscopic hemorrhages[1]. They are one of the cardinal features of cerebral small vessel disease (cSVD), together with other markers such as white matter hyperintensities of presumed vascular origin (WMH), perivascular spaces (PVS), and lacunes[2]. CMB can be encountered in cSVD related to vascular risk factors with a typical central, deep location (basal ganglia, thalamus, brain stem, and cerebellum) and can be even more prominent in cSVD related to cerebral amyloid angiopathy (CAA), predominantly in lobar regions[3]. They are associated with a higher risk of both hemorrhagic and ischemic stroke[4], cognitive decline, and dementia[3]. CMB has also been recognized as one of the main MRI abnormalities that can emerge in association with the use of anti-amyloid beta (Aβ) treatment for Alzheimer's disease[5], coined amyloid-related imaging abnormalities or ARIA[6]. More specifically, it is one feature of ARIA-associated hemorrhage, or ARIA-H. With the recent approval of anti-Aβ antibodies by the US Food and Drug Administration, there is an increasing need for monitoring the presence and emergence of CMB to aid clinical decision-making for such treatments[7]. While their visual detection is reasonably reliable[8], this task

[1]GIN, IMN-UMR5293, CEA, CNRS, Université de Bordeaux, Bordeaux, France. [2]BPH-U1219, INSERM, Université de Bordeaux, Bordeaux, France. [3]CHU de Clermont-Ferrand, Clermont-Ferrand, France. [4]Fealinx, Lyon, France. [5]Neuroimagerie diagnostique et thérapeutique, CHU de Bordeaux, Bordeaux, France. [6]Neurocentre Magendie-U1219, INSERM, Université de Bordeaux, Bordeaux, France. ✉email: marc.joliot@u-bordeaux.fr

can be very time-consuming and dependent on the reader's expertise and would thus greatly benefit from an advanced automatic detection tool.

Historically, CMB were often detected on two-dimensional T2*-weighted gradient echo sequences (2D T2*GRE), in which CMB appear as a small area of signal loss. More modern sequences are typically based on higher-resolution 3D acquisition and include vendor-specific processing to enhance the susceptibility effects, thus increasing the sensitivity to CMB. The term susceptibility-weighted imaging (SWI) is often used broadly to denote these high-spatial-resolution sequences with enhanced susceptibility contrasts, usually by combining the phase information with the magnitude. Although technically it is a specific processing method implemented on Siemens and other scanner vendors, colloquially it is applied more generally to include similar susceptibility-weighted sequences, such as susceptibility-weighted angiography (SWAN) by GE Healthcare and SWI with phase enhancement (SWIp) by Philips[9]. Since the imaging appearance of T2*GRE sequences and the scanner-reconstructed susceptibility-enhanced derivatives of these sequences depend strongly on the acquisition parameters and techniques, the resulting diversity of input images poses a particular challenge for developing an automated method for CMB detection. While the past decade has seen an increasing number of studies proposing automated or semi-automated CMB detection methods, a recent comprehensive review indicates that the majority base the development and evaluation of the method on a single dataset or two, with unknown generalizability in independent samples[10].

Another significant roadblock to the application of automated CMB detection methods is the scarcity of publicly available tools that a third-party user can easily deploy. With the increasing number of studies incorporating supervised deep-learning (DL)-based methods[10–14], the performance of the methods largely depends on the quality of the training datasets. Without the open sharing of the training datasets or the pre-trained model, the robustness of the methods across diverse acquisition types is difficult to test, even if the authors publicly shared code repositories for their model architecture, which in itself is still relatively rare. It hinders the comparison of methods across studies and datasets, and limits their application in both research and clinical settings.

In the present work, we describe a fully automated DL-based CMB detection tool that we call "SHIVA-CMB" detector. It has a 3D Unet-based architecture[15] similar to our previously described detectors for other cSVD markers[16,17], and takes the entire 3D volume of T2*GRE or SWI-like input image to generate a predicted map of CMB. It has been trained on diverse datasets from 6 different studies and seven acquisitions that include 2D and 3D T2*GRE acquisitions with or without susceptibility effects enhancement from different scanner manufacturers. We demonstrate the robustness of our tool on two independent datasets not seen during the training. Crucially, we make the pre-trained model described in this work (https://github.com/pboutinaud/SHIVA_CMB/) as well as the entire pipeline of our tool openly available (https://github.com/pboutinaud/SHiVAi/) to allow the application of our tool on new datasets without a need for retraining or preprocessing.

Methods

Participants and MRI Data description

Datasets from six independent studies with varying population characteristics were used to develop the SHIVA-CMB model, and two additional datasets were used for evaluation only (Table 1). Each dataset comes from cohort studies for which local ethical approval had already been described elsewhere, except for the SHIVA study used in the evaluation, whose details are given in the cohort description below.

	Cohort type	*N* (scans)	Scanner	Acquisition	TR (msec)/TE (msec)/FA (°)	Resolution (mm)
Training/testing						
SABRE-MICCAI2021	Normal aging	11	3T Philips Achieva	2D T2*GRE	1288/21/18	0.45×0.45×3.0
RSS-MICCAI2021	Normal aging	34	1.5T GE	3D T2*GRE	45/31/13	0.49×0.49×0.8
ALFA-MICCAI2021	AD risk and normal aging	27	3T GE Discovery	2D T2*GRE	1300/23/15	1.0×1.0×3.0
DOU	Stroke and normal aging	20	3T Philips	SWIp	17/24/NA	0.45×0.45×2.0 1 mm slice spacing
BBS	First-time stroke	162	3T GE Discovery	SWAN	60/24.3/15	0.43×0.43×1.6
		159		2D T2*GRE	775/21.7/20	0.47×0.47×4.5
AIBL-Real	AD, at risk, and normal aging	57 (30)	3T Siemens Trio Trim	SWI	27/20/20	0.93×0.93×1.75
AIBL-Synthetic		76 (56)				
Testing only						
SHIVA	cSVD and normal aging	14	3T Siemens Prisma	SWI	24/17.1/15	0.8×0.8×3.0
				T2*GRE	872/20/20	1.0×1.0×2.5
MEMENTO	Memory clinic	1992	3T/1.5T multicentre	2D T2*GRE	650/20/20	1.0×1.0×2.5

Table 1. Summary of datasets used for the development and evaluation of SHIVA-CMB. For each dataset, the type of cohort (Alzheimer's disease, *AD*; cerebral small vessel disease, *cSVD*), the total number of scans used in the present study, scanner vendor (and model when this information is available), type of acquisition, basic sequence parameters (Repetiton time, *TR*; echo time, *TE*; flip angle, *FA*), and image resolution are summarized. For the AIBL dataset, some longitudinal acquisitions were included, and the number in the bracket indicates the number of unique participants. For all other datasets, the number of scans also represents the number of participants.

Most of the training datasets were SWI-like acquisitions on 3T scanners from three major vendors (Siemens, Philips, and GE), with the exception of the MICCAI2021 datasets, which were T2*GRE acquisitions on 1.5T or 3T scanners. Each dataset was divided into training/validation and separate, held-out evaluation test sets, as summarized in Table 2. The separation between the training/validation and the held-out test set was made after stratifying the data on the available CMB count (when only visual CMB count or coordinate information was available) or voxel count (when the ground truth CMB label was available) information per dataset, depending on the initial label availability, to roughly balance the frequency of CMB in each set. For each dataset, approximately 80–85% of the data were used for the training/validation, and the remaining for the held-out test.

To further evaluate our tool's generalizability and practical utility, we used two other independent datasets not included in the training and thus unseen by our model. In the first dataset (SHIVA) both SWI and T2*GRE acquisitions were available from the same subjects, with manual labels of CMB in each modality. Thus, it was possible to assess the segmentation accuracy in each modality in this dataset. While no manual CMB labels were available from the second dataset (MEMENTO), we used the expert visual rating of CMB in a large number of subjects with T2*GRE acquisitions from a multicenter study to assess the transferability of our tool in a clinical setting.

SABRE (part of MICCAI2021 Task2 training dataset[18])

The Southall and Brent Revisited (SABRE) study is a prospective population cohort of residents in Southall and Brent, UK[19]. Participants were invited to participate in a brain MRI session on a 3T Philips scanner during their third clinical visit between 2014 and 2018. Ethical approval had been obtained from the National Research Ethics Service Committee, London-Fulham (14/LO/0108)[18]. All methods were performed in accordance with the relevant guidelines and regulations defined for SHIVA study (see SHIVA description). As the cohort was recruited initially to investigate metabolic and cardiovascular disease across ethnicities, participants were composed of three ethnic backgrounds. The mean age at the time of MRI acquisition was 72 years, ranging from 39 to 92 years. The present study used T2* GRE scans from 11 subjects released publicly as part of the training dataset for the Vascular Lesions Detection and Segmentation (VALDO) challenge (https://valdo.grand-challenge.org) organized as part of MICCAI 2021[18].

RSS (part of MICCAI2021 Task2 training dataset[18])

The Rotterdam Scan Study[20] is part of a larger prospective cohort of the Rotterdam Study[21] in Rotterdam, the Netherlands. Participants of the Rotterdam Study aged 45 years and over, free of dementia, were randomly selected and invited to participate in the RSS with a brain MRI session on a 1.5T GE scanner. Ethical approval had been obtained from the Ministry of Health for Research Act[18]. All methods were performed in accordance with the relevant guidelines and regulations defined for SHIVA study (see SHIVA description). The present work uses T2*GRE scans from 34 subjects released publicly as part of the MICCAI2021 VALDO challenge.

Dataset	Training set		Test set		Ground truth label generation
	Number of scans	Number of CMB (median [range])	Number of scans	Number of CMB (median [range])	
SABRE-MICCAI2021	8	86 (3 [0, 62])	3	6 (3 [0, 3])	Provided by MICCAI2021 (Sudre et al., 2022)
RSS-MICCAI2021	28	85 (0 [0, 26])	6	9 (0 [0, 8])	
ALFA-MICCAI2021	24	30 (1 [1, 3])	3	4 (1 [1, 2])	
DOU	17	59 (2 [1, 11])	3	15 (1 [1, 13])	Region growing based on coordinates provided by (Dou et al., 2016)
BBS-SWAN	136	560 (1 [0, 47])	26	79 (1 [0, 28])	Iterative and semi-automated (training) or manual (test) annotation by TT and MG
BBS-T2*GRE	133	555 (1 [0, 47])	26	79 (2 [0, 28])	Labels defined on SWAN projected to T2*GRE (training & test) then reviewed by TT et MG (test)
AIBL-Real	44 (24)*	91 (1 [1, 7])	13 (8)*	57 (2 [1, 17])	Region growing based on coordinates provided by (Momeni et al., 2021) and reviewing by TT and MG
AIBL-Synthetic	60 (48)*	634 (10 [8, 19])	16 (16)*	175 (10.5 [9, 15])	
SHIVA-SWI	-	-	14	112 (3 [1, 63])	Semi-automated annotation by TT and MG
SHIVA-T2*GRE	-	-	14	108 (3 [0, 62])	
MEMENTO	-	-	1992	1503 (0 [0, 50])	-

Table 2. Summary of total number of scans and number of ground truth CMB in training versus test set. The number of scans used for training the model (including validation) and reserved for the test set is summarized, together with the median and range of ground truth CMB labeled by expert raters in each set. For each dataset, generation of these ground truth labels are also described briefly (see text for more details). * Numbers inside the brackets indicate the number of unique subjects.

ALFA (part of MICCAI2021 Task2 training dataset[18])
The Alzheimer's and Families (ALFA) cohort is a Spanish cohort of family and relatives of Alzheimer's Disease, and thus enriched for genetic predisposition to AD, but the participants were cognitively normal and aged 45 to 74 years (mean age ± standard deviation: 55.8 ± 6.7 years)[22]. The study was conducted in Barcelona, Spain, and T2*GRE scans acquired on a 3T GE Discovery scanner from 27 subjects were available as part of the MICCAI2021 VALDO challenge. Ethical approval had been obtained from the Independent Ethics Committee Parc de Salut Mar Barcelona and registered at Clinicaltrials.gov (NCT01835717)[18]. All methods were performed in accordance with the relevant guidelines and regulations defined for SHIVA study (see SHIVA description).

DOU[11]
This is a publicly available dataset (http://www.cse.cuhk.edu.hk/~qdou/cmb-3dcnn/cmb-3dcnn.html) for 20 subjects from Dou et al. (2016) study[11]. The study authors selected the subjects for a public release from a larger dataset of 320 subjects, 126 of whom had a stroke (mean age ± standard deviation: 67.4 ± 11.3 years) and 194 subjects of normal aging (mean age ± standard deviation: 71.2 ± 5.0 years). These participants were recruited and underwent an MRI session in Hong Kong, China. The SWI acquisition was performed with a 3D spoiled gradient-echo sequence using venous blood oxygen level-dependent series on a 3T Philips scanner. All methods were performed in accordance with the relevant guidelines and regulations defined for SHIVA study (see SHIVA description).

BBS[23]
This dataset is from a sub-sample of 162 subjects in the "brain before stroke (BBS)" cohort acquired at Bordeaux University Hospital, France[23]. As stated in Coutureau et al. (2021)[23], the study was approved by ethical standards research committees on human experimentation and all patients or legal representatives provided written informed consent. All methods were performed in accordance with the relevant guidelines and regulations defined for SHIVA study (see SHIVA description). The BBS cohort consisted of 428 patients > 18 years of age with a first-ever diagnosis of minor to severe supratentorial cerebral infarct (mean age ± standard deviation: 67.5 ± 14.1 years, 63.5% male). The selection was based on the amount of CMB visual counts made by two experts on the SWAN acquired from a 3T GE scanner. Out of 361 subjects with SWAN scans with corresponding visual CMB rating, all 92 subjects with at least one CMB identified were included. Additionally, 70 subjects without any CMB identified in the initial CMB rating were randomly selected and included in the study. In addition to the SWAN acquisition, 2D multi-echo fast T2*GRE acquisitions were also available from 159 out of the 162 participants. We extracted the T2*GRE at the TE = 21.7ms in these participants to be included in the training and test dataset.

AIBL[24,25]
This dataset is derived from the Australian Imaging Biomarkers & Lifestyle (AIBL) study[26], in which 288 participants out of > 1000 in the cohort participated in the MRI component of the study and were extensively followed up, with longitudinal acquisitions of brain MRI scans for up to 7 times after the baseline scan[27]. As stated in Momeni et al. (2021)[25] approval for the study was obtained from the Austin Health Human Research Ethics Committee and St Vincent's Health Research Ethics Committee, and written informed consent was obtained. All methods were performed in accordance with the relevant guidelines and regulations defined for SHIVA study (see SHIVA description). At baseline, all participants were > 60 years of age; 53 had been classified as mild Alzheimer's disease and 57 as mild cognitive impairment. The remaining subjects were cognitively normal but were selected to include approximately 50% participants with subjective memory complaints and 50% carrying at least one ApoE ε4 allele[28]. The data used in the present work consists of 57 SWI scans from 30 subjects acquired on a 3T Siemens TRIM TRIO scanner with at least one definite CMB in each scan, as described in[25]. Momeni et al. (2021) also described and publicly released synthetic CMB data generated from scans of subjects with and without CMB. The synthetic CMB data in subjects without real CMB were available as ten different versions of synthetically-generated CMB in 313 SWI scans from 100 subjects. However, so as not to make the model overlearn from the synthetically-generated CMB dataset with a pre-specified number of CMB in each image, we only used the first version and randomly selected 76 SWI scans from 56 subjects from this batch to be used in the training and test data.

SHIVA
This is an ongoing prospective cohort study dedicated to deriving extensive circulating and imaging biomarkers of cSVD as part of the RHU-SHIVA project (https://rhu-shiva.com/). This study was approved by the French central ethics committee (Comité de Protection des personnes: 2022-A00493-40), informed consent was obtained from all the subjects and all methods were performed in accordance with the relevant guidelines and regulations. All participants were > 60 years old, and either had extensive WMH (with a Fazekas score[29] of 2 or 3) or very little of them (Fazekas score of 0–1). As of March 2024, 150 participants have been enrolled and underwent an MRI session that included a T2*GRE and an optional SWI acquisition with 3T Siemens Prisma scanners in Bordeaux or Paris, France. The present work included 14 participants, eight of whom had extensive WMH and the remaining without, with both types of acquisitions available.

MEMENTO[30]
MEMENTO is a French nationwide cohort study aimed at improving the understanding of the natural trajectory of Alzheimer's disease and related disorders[30]. This study was performed in accordance with the guidelines of the Declaration of Helsinki. The MEMENTO study protocol has been approved by the local ethics committee ("Comité de Protection des Personnes Sud-Ouest et Outre Mer III"; approval number 2010-A01394-35). All

participants provided written informed consent and all methods were performed in accordance with the relevant guidelines and regulations. A total of 2323 participants > 60 years of age (mean age ± standard deviation: 70.9 ± 8.7, 62% female) were recruited across the 26 French university-based memory clinics with access to MRI and biobank facilities. Participants had either very mild to mild cognitive impairments or isolated subjective cognitive impairments, but none had been diagnosed with dementia at baseline. The neuroimaging protocol was harmonized across the centers, with 86% of participants scanned with a 3T MRI scanner and the remaining with a 1.5T scanner, and included a 2D T2*GRE sequence[31]. The present work included 1992 participants with experts visual counting of CMB and the complete set of anatomical scans, including the T2*GRE from the baseline measurement. CMB counting was performed by a trained rater according to the guidelines in Greenberg et al. (2009)[3] and using the Microbleed Anatomical Rating Scale (MARS)[32], which was verified by an experienced neuroradiologist, as described in Kaaouana et al. (2015)[31].

Generation of CMB ground truth labels

Table 2 includes a brief summary of how the ground truth labels were generated for each dataset. Of all the datasets used to train the SHIVA-CMB model, manually-traced CMB ground truth labels were originally available only in the three datasets from MICCAI 2021 challenge datasets (SABRE, RSS, and ALFA). The study authors provided coordinate information for CMB location for two publicly available datasets (DOU and AIBL). We used the combination of a region-growing algorithm and manual modification by expert raters (TT and MG) with 3D Slicer to prepare CMB ground truth labels for these datasets. The CMB labels for BBS and SHIVA cohorts were created semi-automatically, in which the same expert raters reviewed the potential CMB clusters generated by the precursors of the SHIVA-CMB model and modified them. The CMB labels for BBS and SHIVA cohorts were created semi-automatically, in which the same expert raters reviewed the potential CMB clusters generated by the precursors of the SHIVA-CMB model and modified them. However, this process was repeated several times for BBS during the iterative learning, in which a batch of around 20 images were reviewed by the experts, modified, and fed back to the model as the training data in each iteration, and ended when all the images were reviewed and the experts were satisfied with the CMB labels. For SHIVA, the candidate CMB labels were generated as the last iteration of the overall training process: only one round of training and reviewing was necessary to create the final ground truth CMB labels on the 14 images each from SWI and T2*GRE scans. The definition of CMB in all datasets followed the STandards for ReportIng Vascular on nEuroimaging (STRIVE) guideline[2,33]. A more detailed description of the ground truth CMB label generation in each dataset can be found in the Supplemental Material, Sect. 1.

SHIVA-CMB detector

SHIVA-CMB pipeline overview

Figure 1 shows the summary of the SHIVA-CMB detection pipeline. As with our previously published SHIVA tools[16,17], the following preprocessing steps are applied to each 3D scan (source T2*GRE or SWI-like image) and CMB label image.

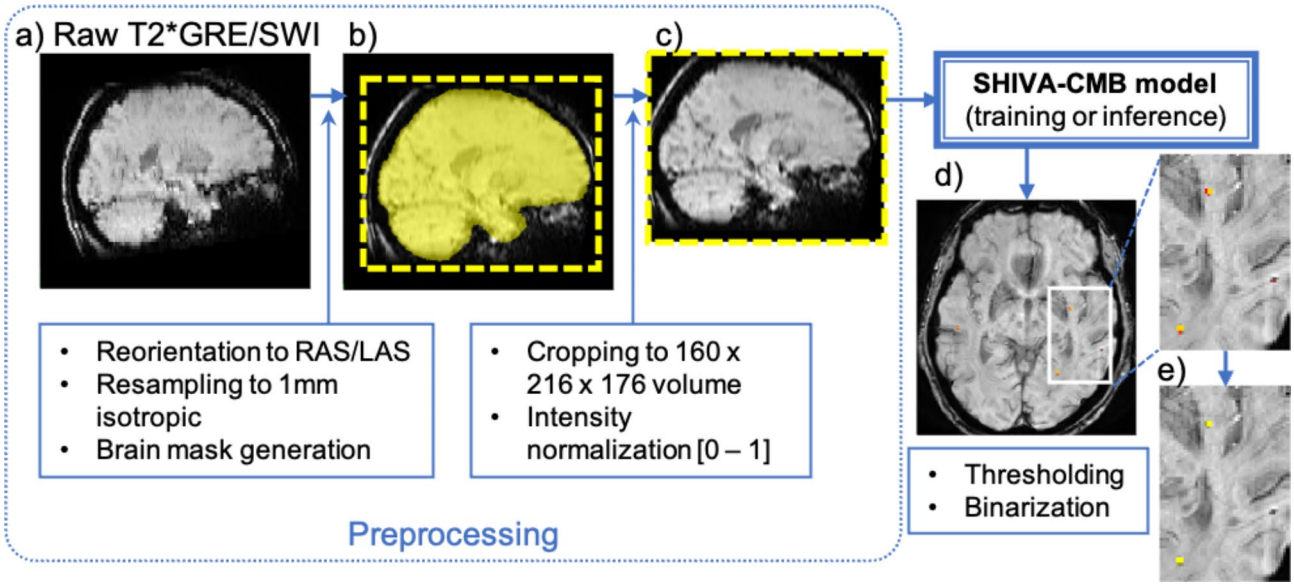

Fig. 1. SHIVA-CMB pipeline summary. (**a**) The raw input image (T2*GRE or SWI) is first reoriented, resampled to 1 mm isotropic, and the brain mask is generated to get the center of mass for cropping. (**b**) The image is cropped to 160 × 216 × 176 and intensity-normalized between values of 0 and 1. (**c**) The preprocessed image is used as the input for the SHIVA-CMB model. When training, CMB label in the same cropped space is used as the ground truth. (**d**) The raw output of the model has values between 0 and 1. (**e**) Thresholding is applied to binarize the image to produce the final segmentation of the predicted CMB.

1. Reorientation to match either LAS or RAS (left or right/anterior/superior) orientation using *fslreorient2std* tool from the FMRIB Software Library (FSL: https://fsl.fmrib.ox.ac.uk/fsl).
2. Resampling to 1 mm isotropic using *flirt* from the FSL[34], with -applyisoxfm and -noresampblur options.
3. A brain mask created based on the source T2*GRE/SWI-like scan is used to obtain the brain mass center, and a bounding box is used to crop all images to a uniform dimension of 160 × 216 × 176 voxels.
4. Voxel intensity values inside the brain mask are linearly rescaled to values between 0 and 1 by setting the 99th percentile value as the maximum and any higher intensity values as 1.

Following these preprocessing steps, the input image is then fed to the SHIVA-CMB model either for training or for inference in new data. When training the model, the ground truth CMB label image of the corresponding image was used to minimize the loss function as detailed below. The raw output of the trained model is the prediction map of CMB, valued between 0 and 1. During the semi-automated generation of CMB labels for BBS and SHIVA datasets (see Supplemental Material, Sect. 1), a low threshold (0.05) was applied when binarizing the final map so that as many potential candidate CMB clusters were reviewed by the experts and rejected if deemed false positive. For the evaluation of the final model, a chosen threshold of 0.4 was applied to binarize the predicted CMB, as described in Supplemental Material, Sect. 3.

Model architecture and implementation

The SHIVA-CMB detector is a modified Unet[35] that takes a 3D input array representing the whole brain to perform CMB segmentation on each input scan. It shares the same basic architecture as the previously published SHIVA tools[16,17], and is directly derived from the latest version of SHIVA-PVS (T1.PVS/v1, https://github.com/pboutinaud/SHIVA_PVS/) that had been trained to segment perivascular spaces from a single 3D image input (individual T1w scan) and had the following architectural features: the number of initial kernels (feature maps) = 10, the number of stages (depths) = 7, the number of 3D convolutions at each stage = 2, the multiplication factor applied at the first convolution layer of each stage = 1.8 (see Fig. 1 in Tsuchida et al., (2023) for a schematic overview of the network architecture).

For every iteration of model training, the following set of data augmentations were applied: (i) randomly flipping on the midsagittal plane, (ii) random voxel translations (up to plus or minus 10 voxels in each orthogonal axis), (iii) random rotations (plus or minus 10 degrees around all orthogonal axis) (iv) non-linear voxel intensity value transformation using a Bézier curve, similarly as in Zhou et al. (2021)[36], with two endpoints set to [0, 0] and [1, 1] and two control points within this range generated randomly.

The network was implemented in Python 3.9, using *Tensorflow* 2.10 with the included *Keras* version, *scikit-learn* (1.0.1), and *scikit-image* (0.18.3), and was trained on a computer (Ubuntu 22.04) with a Xeon ES2640, 40 cores, 256 Gb RAM, and a Tesla A100 GPU with 80Gb RAM. The training was performed iteratively as described in the section below, with each iteration of training using a 3-fold cross-validation scheme on the training/validation set stratified on the CMB load per dataset used at any given iteration. Every iteration used the Adam optimizer with the default parameters of *beta1* = 0.9, *beta2* = 0.999, *epsilon* = 1e–7. It used a cyclical learning rate with exponential decay, with the initial and maximum learning rates set to 1e-6 and 0.001, respectively. As with other SHIVA tools, the loss function for training the model was a Dice loss function computed at the voxel-level per input scan (Eq. 1):

$$Dice\ loss = 1 - \left(\frac{2 \times \sum_{voxels} (y_{true} * y_{pred}) + \epsilon}{\sum_{voxels} y_{true} + \sum_{voxels} y_{pred} + \epsilon} \right) \quad (1)$$

where y_{true} and y_{pred} represent the image arrays for the ground truth CMB label and predicted CMB, respectively, and $(y_{true} * y_{pred})$ represents the intersection between the two images. The smoothing constant of 1e-6 was used to prevent the division by 0. To generate CMB prediction in the test set data, the output from each of the 3 folds valued between 0 and 1 was averaged to create the CMB prediction map, also valued between 0 and 1.

The final trained models described in the present study are publicly available as H5 files that can be used to detect CMB in any new T2*GRE or SWI-like scans as part of the SHIVA-CMB detector package (https://github.com/pboutinaud/SHIVA_CMB). A NVIDIA GPU with at least 9Go of RAM is required to apply the SHIVA-CMB detector.

Incremental and iterative training

Since MICCAI2021 challenge datasets were the only images with manually delineated CMB labels at the outset, we implemented incremental and iterative training of our model to generate or refine the training dataset like those described in Lutnick et al. (2019)[37]. In this process, predictions of early generations of our model were thresholded and binarized to be reviewed and manually modified when necessary by two neuroradiologists experienced in cSVD marker evaluation (MG and TT), then fed back as new training data. The threshold was set to a very low value of 0.05 to produce as many predicted CMB clusters as possible since it is easier to review and reject false positive clusters than to find unmarked false negative CMB. At each iteration of training, the model inherited the initial weights from the previous model to speed up the training.

The details of the successive steps taken to train the current model are provided in the Supplemental Material. Briefly, we initially tried using all MICCAI2021 challenge data (all T2*GRE) to train the first generation of our model and predict CMB in SWIp acquisitions of DOU dataset, from which the coordinate information of the CMB was available. However, likely due to specific acquisition parameters in each of MICCAI2021 datasets, only a model trained with RSS was able to produce a reasonable prediction on DOU dataset. Subsequently, a model trained on the combined RSS and DOU were used to generate predictions of CMB on BBS SWAN datasets, which underwent several iterations of generated CMB reviewing by neuroradiologists, adding the

reviewed CMB label to the training data to generate improved predictions of CMB. This process was repeated for real and synthetic CMB data on AIBL SWI images. Finally, in order to make the model robust to prediction on T2*GRE, the remaining datasets from MICCAI2021 (SABRE and ALFA), as well as T2*GRE scans from the same participants in BBS were added to the training dataset of the final model.

Performance evaluation

Performance evaluation metrics to evaluate data with ground truth manual labels

We used the standard spatial similarity metrics at the lesion cluster level to quantify the similarity of the ground truth and predicted lesion labels. Specifically, we count the number of true positive (TP), false negative (FN), and false positive (FP) clusters, with each individual lesion cluster defined as a 3D connected component using a voxel connectivity of 26. The TP cluster was defined as the predicted CMB cluster that overlaps with the ground truth CMB cluster at least by one voxel, while the predicted cluster without any overlap was counted as the FP. The total numbers of TP, FN, and FP were used to compute the following performance metrics.

- $Sensitivity\ \ (or\ true\ positive\ rate):\ \frac{TP}{TP+FN}$
- $Precision\ \ (or\ positive\ predictive\ value):\ \frac{TP}{TP+FP}$
- *F1 score (or Dice coefficient)*: $\left(\frac{2\times TP}{(2\times TP)+FN+FP}\right)$

As with our previous work, we compute these metrics on a *per input scan* (i.e., single 3D input image volume) basis. However, this method would weigh each input image equally regardless of the amount of CMB present in a given image. Since some of the images may contain no or very few CMB, we define the edge cases where the denominators of sensitivity (TP + FN, or the number of ground truth CMB) or precision (TP + FP, or the number of detected CMB) are zeros as follows:

- If $TP+FN=0$ (no ground truth CMB to be detected)

  Sensitivity = 0 if FP > 0, otherwise 1.

- If $TP+FP=0$ (no detected CMB).

  Precision = 0 if FN > 0, otherwise 1.

While the definition of these edge cases allows numerical computations of metrics in images with no true or detected CMB, it results in a large fluctuation of metrics with a small number of misdetections (FN or FP) in images with no or only a few CMB (e.g., a single FP detection in an image with no CMB would result in the F1 score of 0 instead of 1 in the case of no FP). Alternatively, these metrics can be computed *across input scans* for any given dataset by counting TP, FN, and FP clusters across multiple scans. This effectively avoids the edge cases and also has the effect of weighing every CMB cluster in the dataset equally. This way of computing the metric *across scans* is implicitly adopted in most prior work on CMB detection algorithms that are 2D or 3D patch-based, where only a single value per metric per dataset is reported. Nonetheless, given that the real use case in a clinical setting would involve the evaluation of a single input image from a patient, we report both the average of metrics computed *per scan* and the summary metric computed *across scans* of a given dataset.

In addition, because the high number of false positives is a common problem in CMB detection task, we report the rate of FP per scan (*FPavg*) and per CMB (*FPcmb*) across scans, as recommended by Ferlin et al., (2023) in their review[10]. They are calculated as follows:

- $FPavg\ =\ \frac{FPtotal}{Nimg}$
- $FPcmb\ =\ \frac{FPtotal}{Ncmb}$

where *FPtotal* is the total number of FP clusters, *Nimg* is the number of scans, and *Ncmb* is the total number of ground truth CMB in a given test dataset.

Evaluation of performance on held-out test dataset

We first evaluate the performance of the SHIVA-CMB tool on the held-out test set of all the datasets used to train our model, using the expert-reviewed CMB labels as the ground truth. Collectively, they represent 7 different acquisitions from 6 different scanners. We evaluated the performance across different threshold values (from 0.1 to 1.0, with a step size of 0.1) and cluster-size filters (from 1 to 10 voxels, with a step size of 1) to assess the best threshold and also check the effects of filtering out clusters based on their size, as described in the Supplemental Material. Based on the evaluation, we adopted a threshold of 0.4 for comparing scores across different test datasets. We also filter out clusters ≤ 2 voxels (equivalent to 2 mm^3).

Evaluation of performance on the out-of-sample dataset with ground truth labels

We then evaluate the performance of SHIVA-CMB on the SWI and T2* GRE scans from 14 subjects in the SHIVA cohort, representing a dataset completely unseen by the model during the training. Hence, it provides a better indication of the generalizability of our tool in real-world clinical applications for the two major input acquisition types.

Evaluation of performance on the out-of-sample dataset with visual rating

Lastly, we evaluate our tool in a larger, multi-center study MEMENTO dataset. No manually traced CMB labels can be used for this dataset to compute the spatial similarity with the predicted CMB, but a visual rating with the MARS scale is available. Therefore, we evaluate the performance by comparing the predicted CMB count against the visual count of expert raters using Pearson's correlation. We also evaluate how well the patient classification based on the predicted CMB count aligns with that based on the expert visual count into four groups, highlighting two clinically relevant cutoffs, in the context of ARIA-H monitoring and stroke treatment[38–40]: those without CMB (CMB count = 0), those with a few (1–3), moderate (4), and severe (10 or more) CMB burden.

Results

Performance on held-out test-set data

SHIVA-CMB was trained on seven different 2D or 3D T2*GRE-based acquisitions taken from six cohort-based studies, of which two were population-based (SABRE, RSS) and the remaining targeting high-risk individuals (ALFA and AIBL for Alzheimer's) and/or including clinical population (AIBL for Alzheimer's, DOU and BBS for stroke). We present the performance evaluation after binarizing the CMB prediction map at the chosen threshold of 0.4, as described in the Supplemental Material. We use the standard spatial similarity metrics (sensitivity, precision, and F1) that compare the predicted CMB labels with the ground truth labels reviewed by expert raters. These metrics can be calculated either *per scan* and averaged across scans, or by pooling the true positive (TP), false negative(FN), and false positive (FP) clusters *across scans* on a given dataset to get a single value. We present both in Table 3, which summarizes each performance metric in each of the seven acquisition types separately and across all the held-out test sets (top row, 'All'). Table 3 additionally presents the amount of FP, either per scan (FPavg) or per CMB cluster in a given dataset (FPcmb).

Overall, our model shows a good balance between sensitivity (mean score *per scan* of 0.68 and 0.67 across all 96 test set scans) and precision (mean score *per scan* of 0.72 and 0.82 *across scans*), with a global mean F1 of 0.68 and across-scan F1 of 0.74. The performance metrics vary across different datasets and scan types, with generally better performance in SWI-like scans (DOU, BBS-SWAN, AIBL) compared to T2*GRE scans without susceptibility enhancement (MICCAI 2021 datasets and BBS-T2*GRE), except the RSS data, which shows a high mean F1 score *per scan* and *across-scan* F1 score of 0.96 and 0.82, respectively, on T2*GRE scans. Notably, the FP rates are very low across the datasets, with <1 FP per image for all but AIBL with real CMB dataset, with the average of 2.3 FP per image, and <0.6 FP per CMB for each dataset and across the datasets. The qualitative examination of successfully detected CMB clusters in each dataset, as presented in Fig. 2, highlights the remarkable robustness of the model to detect CMB across diverse input scans with very different tissue contrasts. Notably, the detector did not segment potential sources of CMB mimics, such as iron deposits in the basal ganglia (Fig. 2d) and hemorrhagic lesions (hematoma in Fig. 2f). Figure 3 presents examples of the failed detection (either falsely detected FP or missed FN clusters) in each dataset. It reveals that in some cases, these failures may not be, in fact, failures but rather likely indicate imperfect ground truth labels, representing true CMB missed or non-CMB mis-segmented by human raters. For example, the FP clusters in Fig. 3b and d appear to be true CMB missed by the expert raters, while clusters labeled as CMB in (a) and (h) appear questionable.

Performance on out-of-sample test-set from SHIVA cohort

Although the diverse training data in the model makes it less likely to overlearn any specific input features, it is critical to evaluate the detection performance in a completely new, unseen dataset to gauge its generalizability. Table 4 summarizes the performance evaluation of our tool in the SHIVA cohort, in which both SWI and T2*GRE acquisitions were available from 14 participants who had either no or extensive MRI features of cSVD. It indicates that the performance in this unseen cohort is very close to the overall performance in the in-sample

			Metric per scan (mean (SD))			Metric across scans				
Datasets	Nimg	Ncmb	sensitivity	precision	F1	sensitivity	precision	F1	FPavg	FPcmb
All	96	396	0.68 (0.37)	0.72 (0.39)	0.68 (0.36)	0.67	0.82	0.74	0.6	0.15
SABRE	3	6	0.56 (0.51)	0.67 (0.58)	0.60 (0.53)	0.33	1	0.5	0	0
RSS	6	9	0.96 (0.10)	0.97 (0.07)	0.96 (0.09)	0.78	0.88	0.82	0.17	0.11
ALFA	3	4	0.50 (0.50)	0.67 (0.58)	0.56 (0.51)	0.5	1	0.67	0	0
DOU	3	15	0.64 (0.56)	0.60 (0.53)	0.62 (0.54)	0.87	0.81	0.84	1	0.2
BBS-SWAN	26	76	0.75 (0.35)	0.78 (0.35)	0.75 (0.34)	0.58	0.77	0.66	0.5	0.17
BBS-T2*GRE	26	76	0.48 (0.43)	0.58 (0.46)	0.51 (0.43)	0.41	0.84	0.55	0.23	0.08
AIBL-Real	13	57	0.75 (0.37)	0.52 (0.33)	0.58 (0.33)	0.77	0.59	0.67	2.31	0.53
AIBL-Synth	16	153	0.80 (0.11)	0.96 (0.07)	0.86 (0.07)	0.79	0.96	0.87	0.31	0.03

Table 3. Summary of SHIVA-CMB cluster-level performance metrics on in-sample test-set scans at threshold = 0.4, with clusters $<2\text{mm}^3$ filtered out. Performance metrics computed *per scan* (sensitivity, precision, F1 score) or *across scans* of a given dataset (sensitivity, precision, F1 score, FPavg and FPcmb) at threshold = 0.4 are summarized. Mean and standard deviation (SD) are shown for performance metrics computed per scan. *Nimg*: the number of scans in a given dataset, *Ncmb*: the total number of the ground truth CMB clusters that are $>2\text{mm}^3$.

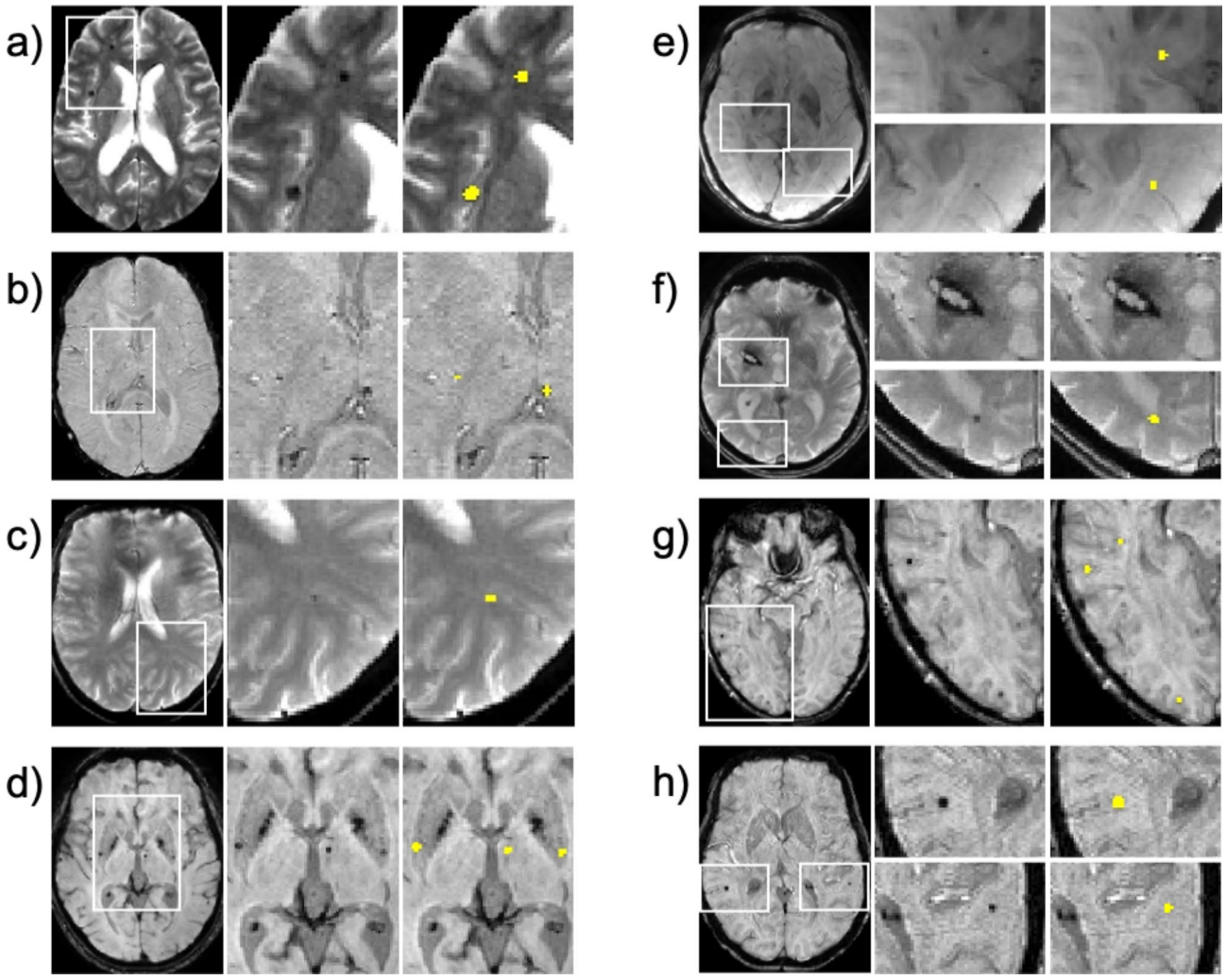

Fig. 2. Examples of CMB detections in test-set scans from each dataset used for model training. The heterogeneity of the input scans from different datasets is evident. For each example, the left-most panel shows a representative axial slice with at least one CMB cluster identified by human raters, and the middle panel shows a magnification of the region(s) indicated by the white box on the left panel. The right-most panel indicates the true positive clusters of the SHIVA-CMB (i.e. predicted CMB clusters in yellow that overlapped with ground truth labels) on the same magnified region, (**a**) T2*GRE from SABRE (part of MICCAI 2021 VALDO challenge), (**b**) T2*GRE from RSS (part of MICCAI 2021 VALDO challenge), (**c**) T2*GRE from ALFA (part of MICCAI 2021 VALDO challenge), (**d**) SWip from DOU, (**e**) SWAN from BBS, (**f**) T2*GRE from BBS, (**g**) SWI with real CMB from ABL, (**h**) SWI with synthetic CMB from AIBL. For (**b**), note that a non-CMB lesion (hematoma, top box) is not detected by SHIVA-CMB.

test set, with the mean F1 *per scan* of 0.62 and 0.87, and *across-scan* F1 scores of 0.76 and 0.78 for SWI and T2*GRE inputs, respectively. Slightly better precision in this dataset (mean precision *per scan* of 0.72 and 0.90 or 0.92 and 0.9 *across-scan* for SWI and T2*GRE inputs, respectively) is reflected in the even lower FP rates compared to those of the in-sample test sets (FPavg of 0.43 and 0.57 and FPcmb of 0.06 and 0.07 for SWI and T2*GRE). Figure 4 presents the examples of SWI (Fig. 4a,d) and T2*GRE (Fig. 4b,c,e,f) inputs and both successful and failed detections in each type of inputs from this cohort.

Comparison with CMB visual rating in MEMENTO cohort

Lastly, we evaluate the performance of the SHIVA-CMB detector on a large, multi-center study with T2*GRE scans acquired from almost 2000 participants in 26 university-based memory clinics across France, with MRI systems from different vendors, models, and field strengths. Figure 4c and f shows two examples of MEMENTO data and possible CMB clusters detected (and possibly missed) by our detector. Although it is not possible to assess the segmentation accuracy in the same manner as other datasets due to lack of manually traced CMB labels, we leveraged the visual rating of CMB by experts using the MARS scale in this dataset to assess how well our tool can emulate the quantification by human experts. Figure 5a shows the correlation between the CMB count in the whole brain detected by experts and the SHIVA-CMB in the 1992 T2*GRE scans of the MEMENTO

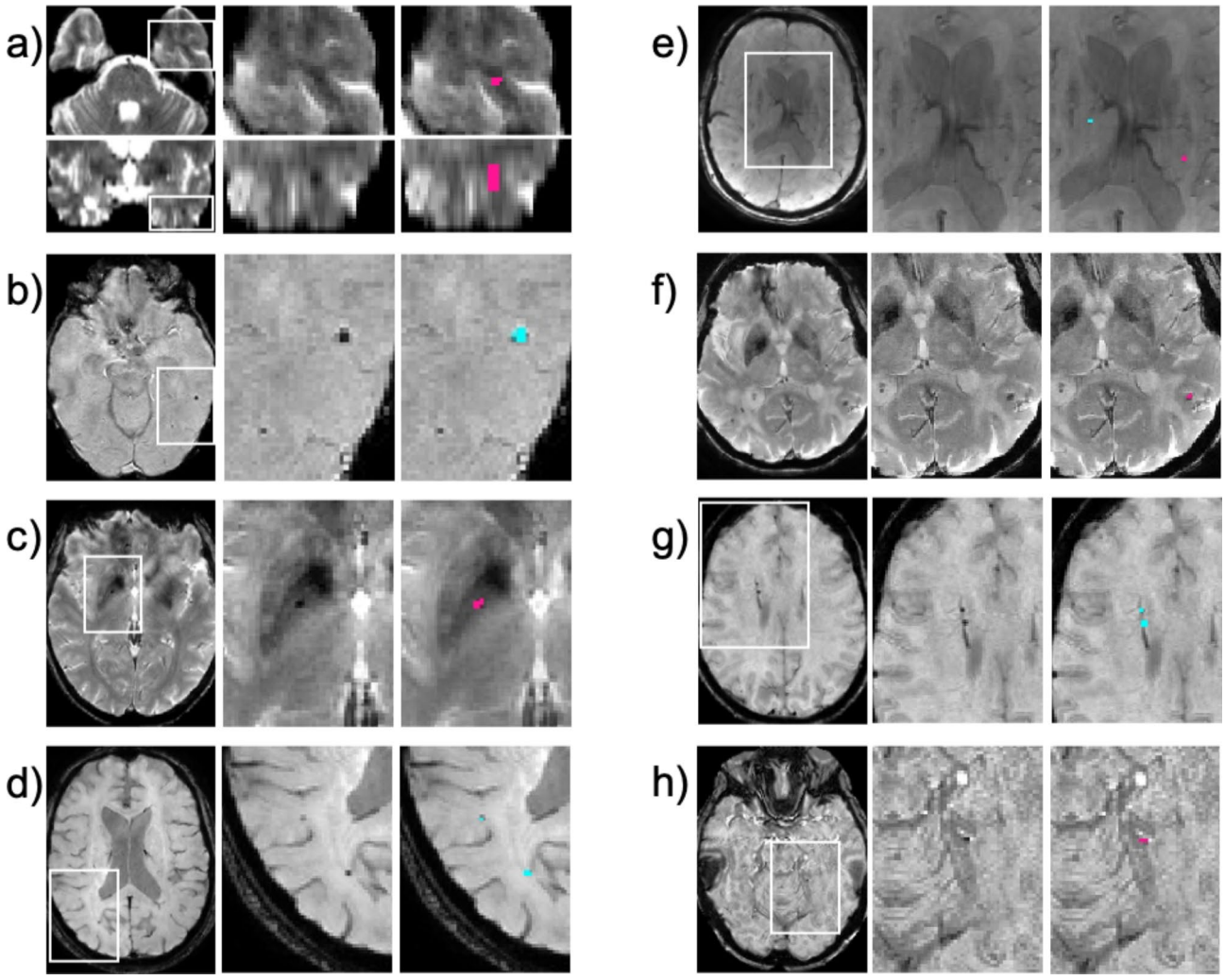

Fig. 3. Examples of detection failures (false negative, in magenta) or false detections (false positives, in cyan) in test-set scans from each dataset used for model training. A similar set of examples as Fig. 2 are shown, but showing false negative or false positive clusters of SHIVA-CMB. For each example, the left- most panel shows a representative axial slice with or without CMB identified by human raters, and the middle panel shows a magnification of the region(s) indicated by the white box on the left panel. The right-most panel shows the false negative (magenta) or false positive (cyan) clusters on the same magnified region. (**a**) T2*GRE from SABRE (part of MICCAI 2021 VALDO challenge), (**b**) T2*GRE from RSS (part of MICCAI 2021 VALDO challenge), (**c**) T2*GRE from ALFA (part of MICCAI 2021 VALDO challenge), (**d**) SWIp from DOU, (**e**) SWAN from BBS, (**f**) T2*GRE from BBS, (**g**) SWI with real CMB from AIBL, (h) SWI with synthetic CMB from AIBL.

			Metric per scan (mean (SD))			Metric across scans				
Datasets	Nimg	Ncmb	sensitivity	precision	f1	sensitivity	precision	f1	FPavg	FPcmb
All	28	210	0.73 (0.31)	0.81 (0.30)	0.74 (0.28)	0.67	0.91	0.77	0.5	0.07
SHIVA-SWI	14	103	0.58 (0.32)	0.72 (0.37)	0.62 (0.30)	0.65	0.92	0.76	0.43	0.06
SHIVA-T2*GRE	14	107	0.88 (0.22)	0.90 (0.18)	0.87 (0.19)	0.69	0.9	0.78	0.57	0.07

Table 4. Summary of SHIVA-CMB performance metrics on evaluation-only SHIVA dataset at threshold = 0.4. Performance metrics computed per scan (sensitivity, precision, f1 score) or across scans of a given dataset (sensitivity, precision, f1 score, FPavg and FPcmb) at threshold = 0.4 are summarized. Mean and standard deviation (SD) are shown for performance metrics computed per scan.

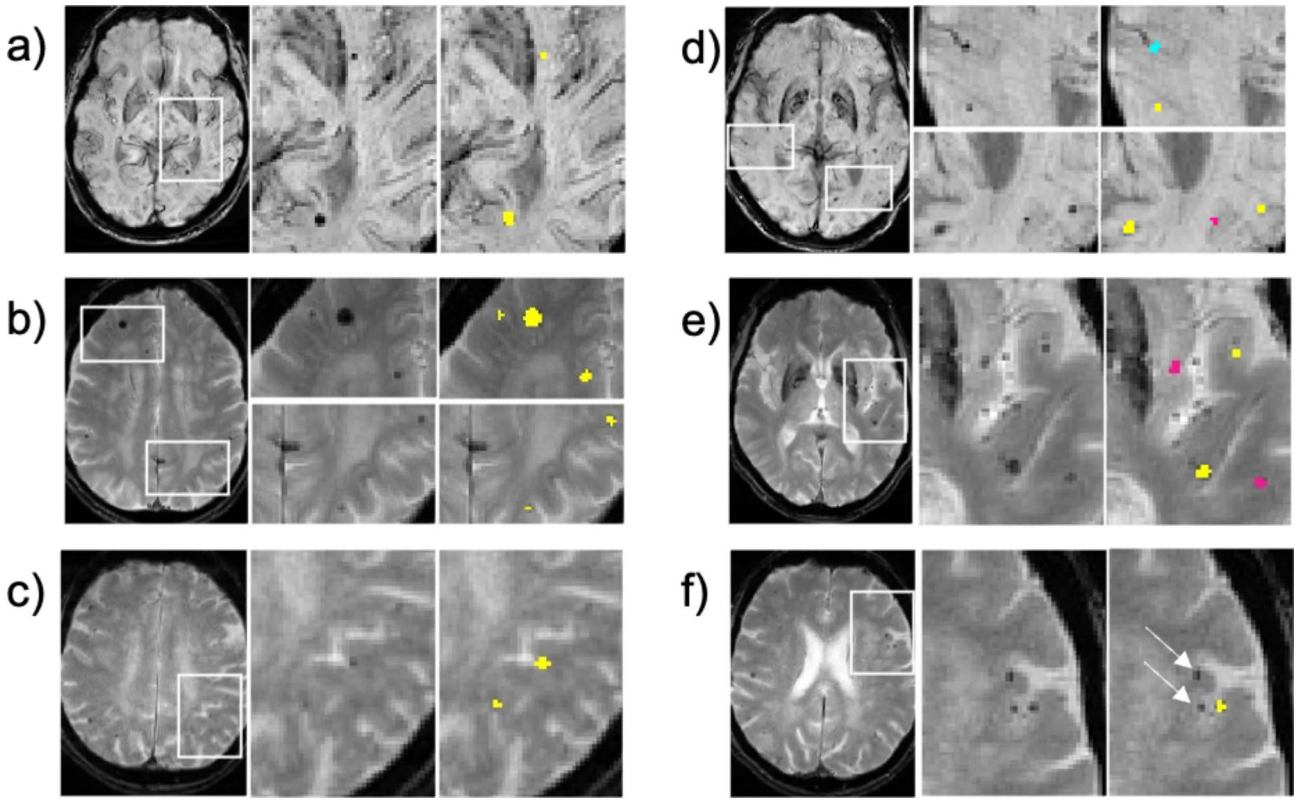

Fig. 4. Examples of CMB detections in scans from unseen dataset. Representative examples of CMB detections by the SHIVA-CMB are shown for SHIVA (**a** and **d** SWI scans, **b** and **e** T2*GRE scans) and MEMENTO (**c** and **f**) cohorts. For each example, the left-most panel shows a representative axial slice with at least one visible CMB cluster, and the middle panel shows a magnification of the region(s) indicated by the white box on the left panel. The right-most panel indicates true positive clusters (i.e. overlapping with ground truth labels: yellow) for (**a**) and (**b**), or false positive (cyan)/ negative (magenta) clusters for (**d**) and (**e**), on the same magnified region. For examples in MEMENTO shown in (**c**) and (**f**), the detected clusters cannot be classified as true or false since there are no ground truth CMB labels in this cohort. Thus, only detected clusters are shown (yellow), but possible false negative clusters are indicated by white arrows in (**f**).

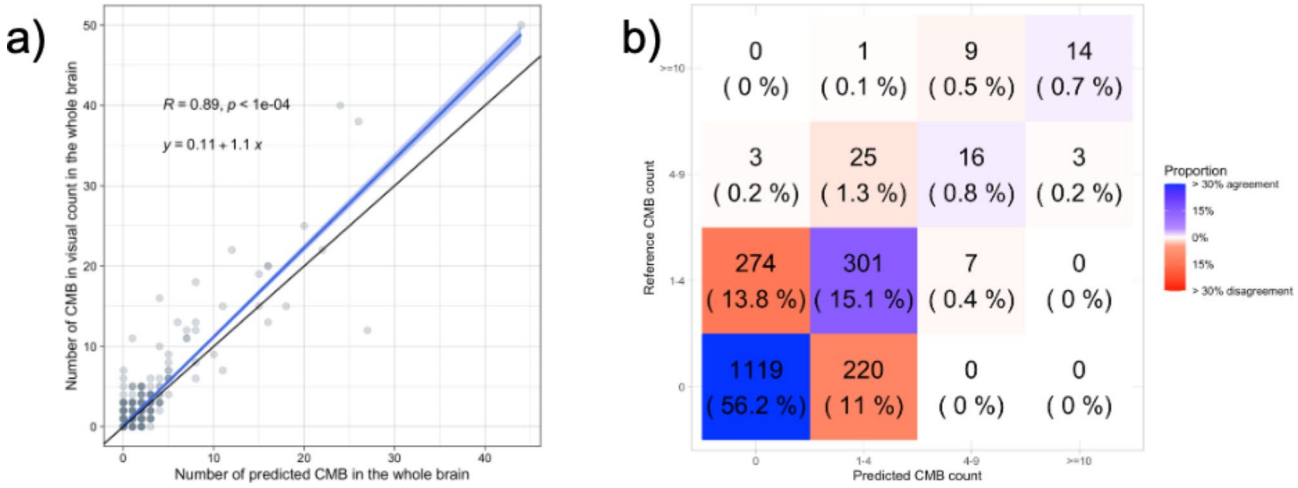

Fig. 5. Comparison of CMB count and patient classification performance in MEMENTO. (**a**) Scatterplot comparing the number of CMB in the visual count by expert human raters and the number of CMB clusters detected by the SHIVA-CMB. Pearson's correlation R and the associated p value, as well as regression equation are indicated. (**b**) Confusion matrix of patient classification into those with no CMB, a small number (1–3 CMB), moderate (4–9), and high (10 or more) burden of CMB. Blue cells show agreement between the human rater- and SHIVA-CMB-based classification, while red cells show disagreement.

participants, demonstrating the highly significant correlation (*Pearson's* $R = 0.89$, $p < 0.0001$) and a regression slope close to 1. We also converted the CMB count information into a clinically relevant classification task of patients into those (1) without any CMB, (2) with a small number (1–3), (3) moderate (4–9), and (4) high burden of CMB (10 or more) to assess the classification accuracy if the SHIVA-CMB tool was to replace an expert visual assessment. Figure 5b shows a confusion matrix of the four classes. The overall accuracy of the SHIVA-CMB in such a classification scheme was 73%. The misclassification of the remaining 27% was mostly between classifying subjects as having no CMB and having a small number of CMB, with a similar proportion of cases where the human rater detected a small number of CMB but not the SHIVA-CMB (274 cases, 13.8%) or vice versa (220, 11.0%). In no cases was there a disagreement in classifying subjects with no CMB and a high burden of CMB.

Discussion

In the current work, we presented the SHIVA-CMB detector, a 3D Unet-based model trained on the diverse T2*GRE acquisitions with or without susceptibility-enhancement provided by different scanner vendors. We aimed to develop and share an openly accessible and automated tool that can robustly detect CMB across different datasets without a need for retraining. We demonstrated that the detector showed a good balance of sensitivity and precision in a heterogeneous held-out test dataset, with the average lesion cluster-level F1 score per scan of 0.68, or across-scan score of 0.74 in the test set consisting of 96 images from seven different acquisition types. It showed a similar level of, or even nominally better, performance in a smaller but never-seen evaluation dataset, for both SWI (0.62 and 0.76 for mean F1 per scan and across-scan F1 scores, respectively) and T2*GRE (0.87 and 0.78 for mean F1 per scan and across-scan F1 scores) scans, highlighting the generalizability of our tool to the new dataset. Across all the datasets used for evaluation, our detector also showed remarkably low FP rates (less than 1 FP per image and < 0.5 FP per CMB in most datasets). The high generalizability of our tool is reinforced by the high correlation of the number of CMB detected by our tool with a visual CMB count by human experts in a much larger evaluation dataset of T2*GRE scans collected in a large multi-center study, also unseen by the model during the training. In the same dataset, the accuracy for clinically relevant patient classification task was 73%, with very little disagreements (~ 2%) in classifying those with moderate (4–9) or high (10 or more) burden of CMB from those with no or a small number (< 4) of CMB. In the context of ARIA-H monitoring, the high accordance with expert raters in distinguishing those with moderate or high burden of CMB from those with no or very few CMB attest to its clinical applicability to monitor those who should be excluded from continuing anti-Aβ treatments[40].

At first glance, the F1 scores we report here may seem less impressive compared to prior works reporting a similar F1 score or accuracy score of 0.9 or higher: in a recent comprehensive review of automated CMB detection[10], at least three studies reported F1 scores > 0.9 and a much higher number of studies reported accuracy scores of > 0.9. However, it should be noted that the many of the reviewed work is patch-based, in which the performance is evaluated on a set of fragmented 2D or 3D patches of original images, typically preselected on the basis of the presence or absence of CMB at the center to balance two classes of patches (note that the calculation of accuracy score itself is only possible with the patch-level classification, since the number of 'true negatives' at the lesion-cluster level per image cannot be meaningfully defined). Thus, it is hard to know how the reported metrics translate when the performance is evaluated at the original image level without selecting fragmented patches from these images. Another, perhaps even more critical point is that many of the reviewed work was trained and evaluated on a single data source or two, as summarized in Table 3 of Ferlin et al., (2023)[10], in stark contrast with six different cohort studies used in the training of our model and additional two datasets included in the independent evaluation sets. The training and evaluation in a single or very limited data source can result in over-estimation of the method performance, and in the case of any supervised learning methods, overfitting to the specific dataset used in a given study. To underscore both of these points, a recently published result of the MICCAI2021 VALDO challenge in which the CMB segmentation performance of all the contestants was evaluated in 147 held-out test set images from three datasets that participating teams did not have access to, using the metrics computed on per-image basis, the best performing method reached a median F1 score of 0.68[18]. It took an ensemble model that combined the top four methods in the competition to reach a median F1 score above 0.7, at 0.76. Thus, although evaluated in different datasets (we did not participate in the challenge and therefore did not have access to the held-out test set), the performance level of the SHIVA-CMB detector is on par with the state-of-the-art methods submitted to the VALDO competition.

As with the best-performing method in the CMB task of the VALDO competition, the SHIVA-CMB detector is based on a single-stage, end-to-end learning using a 3D Unet architecture[15] that takes the entire 3D brain volume as an input. This simple framework has the advantage over typical two-stage frameworks that have been proposed for many earlier CMB detection methods. Two-stage frameworks usually involve the first stage that detects CMB candidates using hand-crafted features, such as intensities, size, and shape, including complex 2D or 3D radial symmetry[41–44], or using DL-based approaches[11,45], followed by the second stage that applies the classification to reduce false positive detections using classic machine learning[41,46,47] or DL-methods[11,43–45]. Although such frameworks can reduce the computational cost by reducing the search space in the first stage, any failure to detect the true candidates at the first stage would propagate to the next stage. Such interdependence of the two stages can complicate retraining with different or larger datasets since both stages need to be optimized. In contrast, our detector can be easily retrained and its weights updated as new training data become available. It can also be repurposed to perform different segmentation tasks that use different input modalities through transfer learning: in fact, the SHIVA-CMB itself initially inherited the weights of our previously described detector for PVS[16] (SHIVA-PVS, T1.PVS/v1 in https://github.com/pboutinaud/SHIVA_PVS/) to speed up the training of the CMB detection task.

Another advantage of our framework is that by using the entire 3D volume of the brain as an input, it is able to use both local (e.g. intensity, 3D shape) and global (e.g. relative anatomical relationships of different structures)

features of CMB in the training data. The 3D contextual information is likely to be crucial for distinguishing the possible mimics from CMB: The most prominent source of these mimics are the flow voids from the vascular structures, most commonly pial vessels, which are difficult to differentiate from cortical CMB when only the 2D slice showing the cross-section is viewed[3]. We note, however, our approach does not completely resolve these difficult mimics, as these were still the most common source of false detection (an example in Fig. 4d). Another common source is the mineral deposits prevalent during aging[48], such as calcifications and iron deposits frequently located in basal ganglia. Although calcifications in lobar locations may not be readily distinguished from CMB without additional source of information such as phase information or quantitative susceptibility maps (QSM) derived from multi-echo T2*GRE sequences[12], typical bilateral deposits in basal ganglia (both iron and calcium) and occasional calcification in choroid plexus or in pineal gland can be distinguished based on their anatomical context and shapes, given enough examples of them in the training data (see Fig. 2d for an example of non-detection of mineral deposits in basal ganglia by the SHIVA-CMB, but also see Fig. 3c for a failure to detect a potential CMB). The advantage of using the 3D input with the whole-brain context over the 2D or partial 3D input is reinforced by the VALDO CMB task competition results, in which the winning team using the full-resolution 3D input out-performed other teams using partial 3D or 2D inputs[18], despite the fact they all used the same basic Unet architecture included in the 'nn (no-new)- Unet' tool[49].

We acknowledge some limitations of the current work. First, in addition to lobar calcifications, there are other potential mimics that our detector was not trained to disambiguate from CMB. They include small cavernous malformations, metastatic melanoma and diffuse axonal injury from head trauma, which however can usually be suspected based on the clinical context or would require multi-modal information (T1-, T2-weighted and gadolinium injection in addition to the primary T2*GRE-based sequence) to be distinguished from CMB[3]. How critical these distinctions are would depend on the context in which our detector is used, i.e. clinical characteristics of the target population and specific research questions. Second, even though we used diverse data sources for training our model, there is always a possibility that they do not sufficiently represent CMB encountered in different medical conditions. Some false negatives in subjects exhibiting particularly high numbers of CMB in the evaluation-only SHIVA and MEMENTO datasets (Fig. 4d–f) may have resulted from the relatively few examples of images with very severe forms of CMB in the training dataset. Incorporating diverse examples of specific clinical conditions known to be associated with higher incidence and severity of CMB in the training dataset could further ameliorate the sensitivity in severe cases and improve generalizability. In the context of cSVD, the enhanced training set from patients with CAA or hereditary conditions related to cSVD like cerebral autosomal dominant arteriopathy[50] could further improve the sensitivity of the SHIVA-CMB. Lastly, we made pragmatic choices during the ground truth CMB label generation across the multiple data sources that rendered the quality of these ground truth labels somewhat heterogeneous. In particular, those for T2*GRE scans in BBS used in the training were suboptimal since they were generated by simply aligning the ground truth labels created for SWAN scans from the same subject to the T2*GRE space, without independent reviewing of T2*GRE scans (although those used as held-out test sets were reviewed in the T2*GRE independently to accurately assess the quality of predicted CMB segmentations). Both the higher resolution and susceptibility-enhancement processing of SWAN images are likely to result in some CMB only visible in these scans and not in the 2D T2*GRE images from the same subjects[51]. We note, however, even when using consistent protocols, there can be considerable uncertainties in whether any given hypointense objects in the T2*GRE-based images are judged to be a true CMB by human expert raters, and there is always the possibility that plausible CMB clusters are missed, as demonstrated by some examples of questionable false positive and negative cases in Fig. 3. Even though individual studies can strive to build the gold standard CMB labels in a given database as rigorously as possible, ultimately, we believe that it will be necessary for the field to collectively build an annotated image database of CMB from diverse data sources, possibly with voting from multiple experts, to establish a consensus gold standard, akin to the collectively created STRIVE guidelines[2,33] but with actual images and annotations, that can be used to train and evaluate future generations of automated methods.

In summary, we presented an openly accessible SHIVA-CMB detector, a 3D Unet-based model with pre-trained weights from diverse training data sources, that can be used out-of-the-box on new T2*GRE-based scans to detect CMB. To our knowledge, this is one of the few functional CMB detection tools that can be applied to new datasets without a need for retraining with an additional set of annotated images, although its flexible design allows further refinement of performance if new annotated images are available. As we have done so for our previously published tools that detect other markers of cSVD[16,17], we plan to upgrade our models and make them available online continuously (https://github.com/pboutinaud/SHIVA_CMB/) as any new datasets become available to retrain and refine the performance. In the era of self-configuring DL tools like nn-Unet[49], we believe that the field should shift away from achieving the highest possible performance metrics on a limited and study-specific dataset and instead focus on improving the practical applicability of the proposed methods through sharing of pre-trained models in the case of supervised methods. In addition, we provide a unified, end-to-end tool that can segment all the principal cSVD markers from raw input MRI images by preprocessing them and applying the SHIVA-CMB and other models we have previously described (SHIVA-PVS[16],SHIVA-WMH[17]) and generate anatomically-specific quantifications for each (https://github.com/pboutinaud/SHiVAi/). Together with the future refinement of our models, we believe this effort can accelerate the characterization of cSVD in both research and clinical settings. By allowing the automated and comprehensive quantification of cSVD brain lesions in large-scale studies, it should help elucidate its genetic and environmental risk factors, and ultimately aid with the intervention.

Data availability

The data from MICCAI 2021 VALDO challenge (SABRE, RSS, and ALFA) are available under a CC BY NC-SA license at the challenge website (https://valdo.grand-challenge.org, registration required). The DOU dataset is

available on the institutional homepage of Dr. Qui Dou (http://www.cse.cuhk.edu.hk/~qdou/cmb-3dcnn/cmb-3dcnn.html). AIBL CMB dataset used in the study is available at (https://doi.org/10.25919/aegy-ny12) under a CISRO data license. BBS, SHIVA and MEMENTO are not publicly available due to French regulations regarding the sharing of medical imaging data for protection of privacy. However, de-identified data may be available by request to the principal investigators in charge of respective studies (BBS: Thomas Tourdias; thomas.tourdias@u-bordeaux.fr, SHIVA: Stephanie Debette, stephanie.debette@u-bordeaux.fr, MEMENTO: Carole Dufouil, carole.dufouil@inserm.fr). Source codes for the SHIVA-CMB detector and the links to pre-trained models are available on GitHub (https://github.com/pboutinaud/SHIVA_CMB/) under a CC BY-NC-SA license.

Received: 25 July 2024; Accepted: 29 November 2024
Published online: 28 December 2024

References

1. Yates, P. A. et al. Cerebral microbleeds: A review of clinical, genetic, and neuroimaging associations. *Front. Neurol.* **4**, 205 (2014).
2. Duering, M. et al. Neuroimaging standards for research into small vessel disease-advances since 2013. *Lancet Neurol.* **22**, 602–618 (2023).
3. Greenberg, S. M. et al. Cerebral microbleeds: A guide to detection and interpretation. *Lancet Neurol.* **8**, 165–174 (2009).
4. Charidimou, A. et al. Brain hemorrhage recurrence, small vessel disease type, and cerebral microbleeds: A meta-analysis. *Neurology* **89**, 820–829 (2017).
5. Filippi, M. et al. Amyloid-related imaging abnormalities and β-amyloid-targeting antibodies: a systematic review. *JAMA Neurol.* **79**, 291–304 (2022).
6. Sperling, R. A. et al. Amyloid-related imaging abnormalities in amyloid-modifying therapeutic trials: recommendations from the Alzheimer's association research roundtable workgroup. *Alzheimers Dement.* **7**, 367–385 (2011).
7. Hampel, H. et al. Amyloid-related imaging abnormalities (ARIA): Radiological, biological and clinical characteristics. *Brain* **146**, 4414–4424 (2023).
8. Cordonnier, C. et al. Improving interrater agreement about brain microbleeds: Development of the brain observer microbleed scale (BOMBS). *Stroke* **40**, 94–99 (2009).
9. Haller, S., Haacke, E. M., Thurnher, M. M. & Barkhof, F. Susceptibility-weighted imaging: Technical essentials and clinical neurologic applications. *Radiology* **299**, 3–26 (2021).
10. Ferlin, M., Klawikowska, Z., Grochowski, M., Grzywińska, M. & Szurowska, E. Exploring the landscape of automatic cerebral microbleed detection: A comprehensive review of algorithms, current trends, and future challenges. *Expert Syst. Appl.* **232**, 120655 (2023).
11. Dou, Q. et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Trans. Med. Imaging* **35**, 1182–1195 (2016).
12. Rashid, T. et al. DEEPMIR: A deep neural network for differential detection of cerebral microbleeds and iron deposits in MRI. *Sci. Rep.* **11**, 14124 (2021).
13. Fan, P. et al. Cerebral microbleed automatic detection system based on the deep learning. *Front. Med. (Lausanne)* **9**, 807443 (2022).
14. Sundaresan, V. et al. Automated detection of cerebral microbleeds on MR images using knowledge distillation framework. *Front. Neuroinformatics.* **17**, 1204186 (2023).
15. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (eds Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G. & Wells, W.) vol 9901 424–432 (Springer International Publishing, (2016).
16. Boutinaud, P. et al. 3D segmentation of perivascular spaces on T1-weighted 3 Tesla MR images with a convolutional autoencoder and a U-shaped neural network. *Front. Neuroinformatics* **15**, 641600 (2021).
17. Tsuchida, A. et al. Early detection of white matter hyperintensities using SHIVA-WMH detector. *Hum. Brain Mapp.* https://doi.org/10.1002/hbm.26548 (2023).
18. Tillin, T. et al. The relationship between metabolic risk factors and incident cardiovascular disease in europeans, South asians, and African caribbeans: SABRE (Southall and Brent Revisited)—A prospective population-based study. *J. Am. Coll. Cardiol.* **61**, 1777–1786 (2013).
19. Ikram, M. A. et al. The Rotterdam scan study: Design update 2016 and main findings. *Eur. J. Epidemiol.* **30**, 1299–1315 (2015).
20. Ikram, M. A. et al. The Rotterdam study: 2018 update on objectives, design and main results. *Eur. J. Epidemiol.* **32**, 807–850 (2017).
21. Molinuevo, J. L. et al. The ALFA project: A research platform to identify early pathophysiological features of Alzheimer's disease. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **2**, 82–92 (2016).
22. Coutureau, J. et al. Cerebral small vessel disease MRI features do not improve the prediction of stroke outcome. *Neurology* **96**, e527–e537 (2021).
23. Momeni, S. et al. Synthetic cerebral microbleed on SWI images. *CSIRO* https://doi.org/10.25919/aegy-ny12 (2021).
24. Momeni, S. et al. Synthetic microbleeds generation for classifier training without ground truth. *Comput. Methods Progr. Biomed.* **207**, 106127 (2021).
25. Ellis, K. A. et al. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int. Psychogeriatr.* **21**, 672–687 (2009).
26. Fowler, C. et al. Fifteen years of the Australian imaging, biomarkers and lifestyle (AIBL) study: Progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to Alzheimer's disease. *J. Alzheimers Dis. Rep.* **5**, 443–468 (2021).
27. Rowe, C. C. et al. Amyloid imaging results from the Australian imaging, biomarkers and lifestyle (AIBL) study of aging. *Neurobiol. Aging* **31**, 1275–1283 (2010).
28. Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I. & Zimmerman, R. A. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *AJR Am. J. Roentgenol.* **149**, 351–356 (1987).
29. Dufouil, C. et al. Cognitive and imaging markers in non-demented subjects attending a memory clinic: study design and baseline findings of the MEMENTO cohort. *Alzheimers Res. Ther.* **9**, 67 (2017).
30. Kaaouana, T. et al. 2D harmonic filtering of MR phase images in multicenter clinical setting: toward a magnetic signature of cerebral microbleeds. *Neuroimage* **104**, 287–300 (2015).
31. Gregoire, S. M. et al. The microbleed anatomical rating scale (MARS): Reliability of a tool to map brain microbleeds. *Neurology* **73**, 1759–1766 (2009).
32. Wardlaw, J. M. et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* **12**, 822–838 (2013).
33. Sudre, C. H. et al. Where is VALDO? VAscular lesions detection and segmentatiOn challenge at MICCAI 2021. *Med. Image Anal.* **91**, 103029 (2024).
34. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).

35. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) vol. 9351 234–241Springer International Publishing, (2015).
36. Zhou, Z., Sodha, V., Pang, J., Gotway, M. B. & Liang, J. Models genesis. *Med. Image Anal.* **67**, 101840 (2021).
37. Lutnick, B. et al. An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nat. Mach. Intell.* **1**, 112–119 (2019).
38. Cummings, J. et al. Lecanemab: Appropriate use recommendations. *J. Prev. Alzheimers Dis.* **10**, 362–377 (2023).
39. Tsivgoulis, G. et al. Risk of symptomatic intracerebral hemorrhage after intravenous thrombolysis in patients with acute ischemic stroke and high cerebral microbleed burden: A meta-analysis. *JAMA Neurol.* **73**, 675–683 (2016).
40. Charidimou, A. et al. Microbleeds, cerebral hemorrhage, and functional outcome after stroke thrombolysis. *Stroke* **48**, 2084–2090 (2017).
41. Ateeq, T. et al. Ensemble-classifiers-assisted detection of cerebral microbleeds in brain MRI. *Comput. Electr. Eng.* **69**, 768–781 (2018).
42. Morrison, M. A. et al. A user-guided tool for semi-automated cerebral microbleed detection and volume segmentation: Evaluating vascular injury and data labelling for machine learning. *Neuroimage Clin.* **20**, 498–505 (2018).
43. Liu, S. et al. Cerebral microbleed detection using susceptibility weighted imaging and deep learning. *Neuroimage* **198**, 271–282 (2019).
44. Chen, Y., Villanueva-Meyer, J. E., Morrison, M. A. & Lupo, J. M. Toward automatic detection of radiation-induced cerebral microbleeds using a 3D deep residual network. *J. Digit. Imaging* **32**, 766–772 (2019).
45. Al-Masni, M. A., Kim, W. R., Kim, E. Y., Noh, Y. & Kim, D. H. Automated detection of cerebral microbleeds in MR images: A two-stage deep learning approach. *Neuroimage Clin.* **28**, 102464 (2020).
46. Fazlollahi, A. et al. Computer-aided detection of cerebral microbleeds in susceptibility-weighted imaging. *Comput. Med. Imaging Graph.* **46 Pt 3**, 269–276 (2015).
47. Qi et al. Automatic cerebral microbleeds detection from MR images via independent subspace analysis based hierarchical features. In *Conference of the Proceedings IEEE Engineering in Medicine and Biology Society*, vol. 2015 7933–7936 (2015).
48. Acosta-Cabronero, J., Betts, M. J., Cardenas-Blanco, A., Yang, S. & Nestor, P. J. In vivo mri mapping of brain iron deposition across the adult lifespan. *J. Neurosci.* 36, 364–374 (2016).
49. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
50. Puy, L. et al. Cerebral microbleeds: From depiction to interpretation. *J. Neurol. Neurosurg. Psychiatry* https://doi.org/10.1136/jnnp-2020-323951 (2021).
51. Nandigam, R. N. K. et al. MR imaging detection of cerebral microbleeds: Effect of susceptibility-weighted imaging, section thickness, and field strength. *AJNR Am. J. Neuroradiol.* **30**, 338–343 (2009).

Acknowledgements
We thank Carole Dufouil, Geneviève Chêne and Vincent Bouteloup for providing the MEMENTO cohort and CMB manual counting.

Author contributions
Ami Tsuchida: Conceptualization, Formal analysis, Investigation, Data Curation, Writing-Original Draft, Visualization; Philippe Boutinaud: Conceptualization, Methodology, Software, Writing-Review & Editing; Martin Goubet: Expertise in CMB detection (manual tracing), Writing-Review & Editing. Iana Astafeva : Data Curation; Victor Nozais: Software Development. Pierre-Yves Hervé: Data Curation and Analysis. Thomas Tourdias: Expertise in CMB detection (manual tracing), Writing-Review & Editing. Stéphanie Debette: Funding acquisition, Project administration, Writing-Review & Editing; Marc Joliot: Conceptualization, Supervision, Project administration, Writing-Review & Editing.

Funding
This work was supported by a grant overseen by Agence National de la Recherche Française (ANR) as part of the "Investment for the Future Programme" ANR-18-RHUS-0002 and as part of the France2030-funded Precision and Global Vascular Brain Health Institute (IHU VBHI, ANR-23-IAHU-0001). This work was also supported by a grant from Agence National de la Recherche Française (ANR-16-LCV2-0006-01, LABCOM Ginesislab). This study was also conducted in the framework of the University of Bordeaux's France 2030 program RRI "IMPACT" that received financial support from the French government.

Declarations

Conflict of interest
The authors have nothing to declare.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-81870-5.

Correspondence and requests for materials should be addressed to M.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.