



## OPEN ACCESS

## EDITED BY

Xiaohao Cai,  
University of Southampton, United Kingdom

## REVIEWED BY

Ramón Iglesias Rey,  
Health Research Institute of Santiago de  
Compostela (IDIS), Spain  
Jiong Wu,  
University of Pennsylvania, United States

## \*CORRESPONDENCE

Ludovica Griffanti  
✉ ludovica.griffanti@psych.ox.ac.uk

†These authors have contributed equally to this work and share last authorship

RECEIVED 11 April 2023

ACCEPTED 19 June 2023

PUBLISHED 10 July 2023

## CITATION

Sundaresan V, Arthofer C, Zamboni G, Murchison AG, Dineen RA, Rothwell PM, Auer DP, Wang C, Miller KL, Tendler BC, Alfaro-Almagro F, Sotiropoulos SN, Sprigg N, Griffanti L and Jenkinson M (2023) Automated detection of cerebral microbleeds on MR images using knowledge distillation framework. *Front. Neuroinform.* 17:1204186. doi: 10.3389/fninf.2023.1204186

## COPYRIGHT

© 2023 Sundaresan, Arthofer, Zamboni, Murchison, Dineen, Rothwell, Auer, Wang, Miller, Tendler, Alfaro-Almagro, Sotiropoulos, Sprigg, Griffanti and Jenkinson. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Automated detection of cerebral microbleeds on MR images using knowledge distillation framework

Vaanathi Sundaresan<sup>1,2</sup>, Christoph Arthofer<sup>2,3,4</sup>,  
Giovanna Zamboni<sup>2,5,6</sup>, Andrew G. Murchison<sup>7</sup>,  
Robert A. Dineen<sup>3,4,8</sup>, Peter M. Rothwell<sup>5</sup>, Dorothee P. Auer<sup>3,4,8</sup>,  
Chaoyue Wang<sup>2</sup>, Karla L. Miller<sup>2</sup>, Benjamin C. Tendler<sup>2</sup>,  
Fidel Alfaro-Almagro<sup>2</sup>, Stamatios N. Sotiropoulos<sup>2,3,4</sup>,  
Nikola Sprigg<sup>9</sup>, Ludovica Griffanti<sup>2,10\*†</sup> and Mark Jenkinson<sup>2,11,12†</sup>

<sup>1</sup>Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, Karnataka, India, <sup>2</sup>Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom, <sup>3</sup>National Institute for Health and Care Research (NIHR) Nottingham Biomedical Research Centre, Queen's Medical Centre, University of Nottingham, Nottingham, United Kingdom, <sup>4</sup>Sir Peter Mansfield Imaging Centre, University of Nottingham, Nottingham, United Kingdom, <sup>5</sup>Centre for Prevention of Stroke and Dementia, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom, <sup>6</sup>Dipartimento di Scienze Biomediche, Metaboliche e Neuroscienze, Università di Modena e Reggio Emilia, Modena, Italy, <sup>7</sup>Department of Neuroradiology, Oxford University Hospitals National Health Service (NHS) Foundation Trust, Oxford, United Kingdom, <sup>8</sup>Radiological Sciences, Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, United Kingdom, <sup>9</sup>Stroke Trials Unit, Mental Health and Clinical Neuroscience, University of Nottingham, Nottingham, United Kingdom, <sup>10</sup>Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, Oxford, United Kingdom, <sup>11</sup>South Australian Health and Medical Research Institute, Adelaide, SA, Australia, <sup>12</sup>Australian Institute for Machine Learning, School of Computer Science, The University of Adelaide, Adelaide, SA, Australia

**Introduction:** Cerebral microbleeds (CMBs) are associated with white matter damage, and various neurodegenerative and cerebrovascular diseases. CMBs occur as small, circular hypointense lesions on T2\*-weighted gradient recalled echo (GRE) and susceptibility-weighted imaging (SWI) images, and hyperintense on quantitative susceptibility mapping (QSM) images due to their paramagnetic nature. Accurate automated detection of CMBs would help to determine quantitative imaging biomarkers (e.g., CMB count) on large datasets. In this work, we propose a fully automated, deep learning-based, 3-step algorithm, using structural and anatomical properties of CMBs from any single input image modality (e.g., GRE/SWI/QSM) for their accurate detections.

**Methods:** In our method, the first step consists of an initial candidate detection step that detects CMBs with high sensitivity. In the second step, candidate discrimination step is performed using a knowledge distillation framework, with a multi-tasking teacher network that guides the student network to classify CMB and non-CMB instances in an offline manner. Finally, a morphological clean-up step further reduces false positives using anatomical constraints. We used four datasets consisting of different modalities specified above, acquired using various protocols and with a variety of pathological and demographic characteristics.

**Results:** On cross-validation within datasets, our method achieved a cluster-wise true positive rate (TPR) of over 90% with an average of <2 false positives per subject. The knowledge distillation framework improves the cluster-wise TPR of the student model by 15%. Our method is flexible in terms of the input modality and provides comparable cluster-wise TPR and better cluster-wise precision compared to existing state-of-the-art methods. When evaluating across different

datasets, our method showed good generalizability with a cluster-wise TPR >80 % with different modalities. The python implementation of the proposed method is openly available.

#### KEYWORDS

deep learning, knowledge distillation, detection, susceptibility-weighted image (SWI), quantitative susceptibility mapping (QSM), magnetic resonance imaging, cerebral microbleed (CMB)

## 1. Introduction

Cerebral microbleeds (CMBs) are hemosiderin deposits due to micro-hemorrhages in the brain. CMBs are found in subjects with cerebrovascular diseases, cognitive impairment and dementia, and also found in healthy elderly subjects. CMBs have been associated with white matter damage, and various neurodegenerative diseases including Alzheimer's disease and cerebral amyloid angiopathy (CAA). The presence of CMBs has also been shown to increase the risk of symptomatic intracerebral hemorrhage (ICH) and stroke (Cordonnier et al., 2007). Identification of CMBs and determining their distribution could help in obtaining important biomarkers for various diseases [e.g., lobar CMBs and deep/infratentorial CMBs might indicate CAA and hypertensive vasculopathy, respectively (Greenberg et al., 2009)].

CMBs appear as small, circular, well-defined hypointense lesions ranging from 2 to 10 mm in size on T2\*-weighted gradient recalled echo (GRE) images. Due to the paramagnetic susceptibility of the iron content in the CMBs, modalities such as susceptibility-weighted imaging (SWI) (Haacke et al., 2004) and quantitative susceptibility mapping (QSM) images (Liu et al., 2015) are useful in the identification of CMBs. While all the above modalities are derived from the same scan, they use different aspects of data—T2\*-weighted GRE are derived from magnitude only, QSM and SWI are derived from a combination of phase and magnitude. When compared to T2\*-weighted GRE (T2\*-GRE) images, CMBs appear more prominently on SWI images due to the blooming effect (Greenberg et al., 2009; Charidimou and Werring, 2011). Unlike T2\*-GRE and SWI modalities, CMBs appear hyperintense on QSM images.

Automated detection of CMBs is highly challenging due to their small size, contrast variations, sparse distribution and the presence of imaging artefacts (e.g., ringing effect, susceptibility artefacts at tissue interfaces). Additionally, the presence of various “CMB-like” structures (or *mimics*) with diamagnetic (e.g., calcifications) and paramagnetic (e.g., micrometastases and hemorrhages) properties make the accurate detections of CMBs very difficult (for the list of mimics and their description, refer to Greenberg et al. (2009)). While the use of SWI images generally improves the CMB contrast when compared to GRE magnitude images (Nandigam et al., 2009; Shams et al., 2015), SWI also enhances mimics with magnetic susceptibility differences (both diamagnetic and paramagnetic), making it difficult to identify true CMBs (Greenberg et al., 2009). QSM could be useful to accurately identify true CMBs since it allows to separate diamagnetic tissues (with negative susceptibility, appearing hypointense) from paramagnetic tissues (with positive susceptibility, appearing hyperintense). On QSM images, CMBs

appear hyperintense while diamagnetic mimics (e.g., calcifications) will appear hypointense (Rashid et al., 2021).

### 1.1. Existing literature on CMB detection

Various semi-automated and automated methods have been proposed for CMB detection. Most of the methods follow a common pattern with two steps: CMB candidate detection and post-processing to remove false positives (FPs). The first step generally achieves high sensitivity, while the second step is more challenging and leads to improvement in the precision. In the semi-automated methods, manual intervention has often been used in the cleaning-up step to remove FPs (Barnes et al., 2011; Seghier et al., 2011; van den Heuvel et al., 2016; Morrison et al., 2018). Occasionally, candidate detection (De Bresser et al., 2013; Lu et al., 2021a) and ground truth verification (Kuijf et al., 2012, 2013) also involve manual intervention. Manual detection of CMB candidates is extremely labor-intensive, especially when done on a large number of subjects (e.g., around 8,000 subjects from the UK Biobank Lu et al., 2021a), and might increase the risk of observer error, given the large number of scans and low prevalence rate. Fully automated methods, with high accuracy, could therefore be useful. Various fully automated methods have been proposed, with the candidate detection step often using hand-crafted shapes (Bian et al., 2013; Fazlollahi et al., 2014), intensity (Fazlollahi et al., 2015) and geometric features (Fazlollahi et al., 2014) within supervised classifier frameworks (Pan et al., 2008; Ghafaryasl et al., 2012; Fazlollahi et al., 2014, 2015; Dou et al., 2015). The FP reduction stage is typically based on supervised classifiers (Pan et al., 2008; Dou et al., 2015; Fazlollahi et al., 2015) using local intensity features and shape descriptors [e.g., Hessian-based shape descriptors (Fazlollahi et al., 2015)]. Among the shape descriptors, the radial symmetry transform has been most commonly used (Bian et al., 2013; Liu et al., 2019b), exploiting the circular shape of CMBs. Hence, using structural (e.g., intensity and shape) and anatomical information in combination with the local characteristics (e.g., local contrast) could aid in the reduction of FPs and more accurate detections of CMBs (Dou et al., 2015).

Conventional machine learning (ML) methods require the extraction of meaningful features capable of distinguishing CMBs from the background and mimics. However, due to the small size and variation in shape and intensities of CMBs, designing robust, descriptive and cost-effective features is highly challenging. The use of deep learning models, especially convolutional neural networks (CNNs) could overcome this challenge and provide more accurate CMB detection, since they efficiently extract both local and global

contextual information. For instance, 3D CNN models have been used for feature extraction (Chen et al., 2015) and patch-level CMB detection (Dou et al., 2016). Dou et al. (2016) used a local region-based approach for the segmentation of CMB candidates and discrimination of CMB and non-CMB patches. They initially trained a 3D CNN with true CMB samples and randomly selected background samples. They then applied the initial model on the training set and used the false positive patches for enlarging the training dataset in the discrimination step. Another region-based CNN method using You Only Look Once (YOLO) (Redmon and Farhadi, 2017) was proposed by Al-Masni et al. (2020) (using a 3D CNN for FP reduction). In addition to the above methods, deep ResNets (He et al., 2016) were used for patch-level CMB classification (Chen et al., 2018; Liu et al., 2019b), along with a post-processing step using intensity morphological operations (Liu et al., 2019b). Given the size and sparsity of CMBs, class imbalance between CMBs and background is one of the major problems. Due to this, several methods used equal numbers of CMB patches and non-CMB patches, selected using manually annotated CMB voxels (and a comparable number of non-CMB voxels) for training and evaluation purposes (Zhang et al., 2016, 2018; Wang et al., 2019; Hong et al., 2020; Lu et al., 2021b). Note that patches selected in these methods may contain multiple CMBs.

## 1.2. Existing literature on knowledge distillation

Deep neural networks have been rapidly developing over recent years for accurate medical image segmentation tasks, including CMB segmentation, as mentioned above. However, the improved performance is achieved at the cost of long training times and using resource-intensive complex models (Lan et al., 2018). Hence, training small networks that are computationally efficient and generalizable across datasets is highly desirable. With this aim of model compression (Buciluță et al., 2006), knowledge distillation (KD) (Ba and Caruana, 2013; Hinton et al., 2015) aims to train a smaller network (usually referred as a *student network*) with the supervision (or distillation of knowledge) from a larger network (referred as a *teacher network*). In KD, the student network is typically trained to match the prediction quality of the teacher network, and has been shown to reduce overfitting (Hinton et al., 2015; Lan et al., 2018). KD methods have been successfully used for various object detection tasks (Chen et al., 2017), including lesion segmentation on brain MR images (Lachinov et al., 2019; Hu et al., 2020; Vadacchino et al., 2021). The most commonly used distillation types include response-based (Hinton et al., 2015; Kim and Kim, 2017; Ding et al., 2019; Müller et al., 2019) and feature-based distillation (Romero et al., 2014; Zhou et al., 2018; Jin et al., 2019). In response-based distillation, the output logits from the softmax layer are softened (also known as *soft labels*) using a *temperature* parameter that acts as a regularization factor (Hinton et al., 2015). In the feature-based distillation, outputs of intermediate layers of the teacher model are used to train the student model (e.g., *hint learning* using outputs of hidden layers Romero et al., 2014; Jin et al., 2019 and parameter sharing of intermediate layers Zhou et al., 2018).

Based on the training methods, offline distillation (using a pretrained teacher model to train the student models)

(Romero et al., 2014; Hinton et al., 2015), online distillation (training teacher and student models together) (Zhou et al., 2018; Guo et al., 2020) and self-distillation (where the student models from prior epochs become the teacher for the subsequent epochs) (Yang et al., 2019; Zhang et al., 2019) are most commonly used. Various techniques have also been proposed to improve the generalizability and the performance of the student models including using noisy data (Li et al., 2017; Sarfraz et al., 2019), adaptive regularization of distillation parameters (Ding et al., 2019) and adversarial perturbation of data for training (Xie et al., 2020). Multi-task learning methods have also been shown to provide good regularization, reducing the risk of over-fitting (Liu et al., 2019a; Ye et al., 2019). The auxiliary task could be a related task [e.g., auxiliary classification network in lesion segmentation (Yang et al., 2017)] or an adversarial task [e.g., adversarial training of domain predictor in domain adaptation networks (Ganin et al., 2016)].

So far, KD has never been used for CMB detection to the best of our knowledge. However, in this context, a teacher-student network could be highly beneficial. The teacher network is trained to differentiate CMBs from non-CMBs and then distil this knowledge for the student model to distinguish CMBs from various mimics. In this work, we, therefore, used for the first time a knowledge-distillation framework for accurate and fully automated detection of CMBs, given a single image modality. We propose a 3-step approach: in the first two steps we used 3D CNN models for CMB candidate detection and discrimination. In the third post-processing step, we used appearance-based attributes to reduce false positives. We tested our approach in the presence of mimics, across different datasets with different modalities and pathological conditions. Our main contributions are as follows:

- In the initial CMB candidate detection step (Section 2.2), we utilize the radial symmetry property of CMBs for more efficient candidate detection with high sensitivity.
- In the candidate discrimination step (Section 2.3), we use a knowledge distillation framework to create a light-weight student model from a multi-tasking teacher model, which overcomes the class imbalance between CMBs and the background, leading to the effective removal of false positives without reducing sensitivity.
- In the final post-processing step (Section 2.4), we exploit the structural properties of CMBs to further reduce false positives and improve precision by rejecting CMB mimics from the discrimination step.
- We evaluated our method on four different datasets (details are provided in Section 3). Through the experiments described in Section 4, we studied the contribution of the individual steps on the CMB detection performance, and also the effect of various modalities and different pathological conditions on the results. We also performed an indirect comparison of our results with existing methods at various stages of detection.

## 2. Materials and methods

In the following sections we describe the details of our method. We initially preprocess the input data and remove blood vessels in the images as specified in Section 2.1. The proposed method consists of three steps. (1) 3D CMB initial

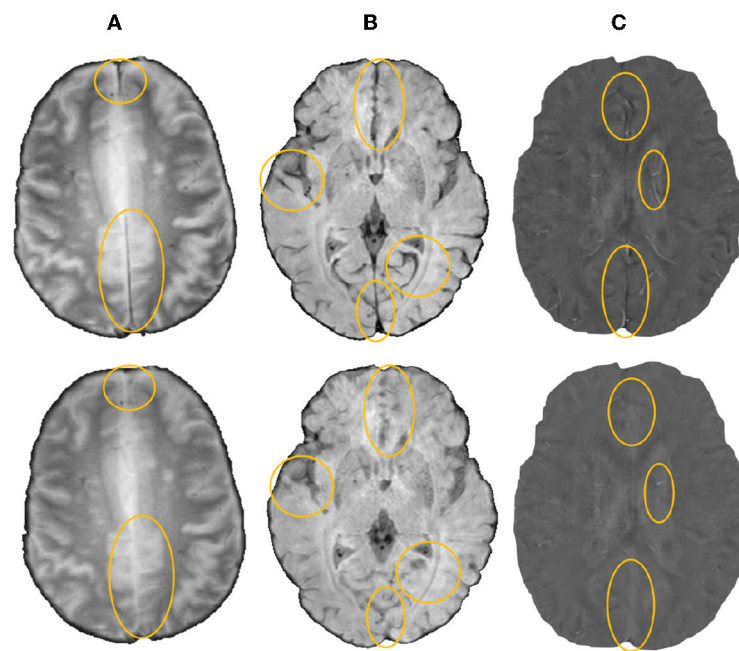


FIGURE 1

Sample images after removal of blood vessels and sulci (regions with major changes indicated by circles in the bottom row) shown for (A) T2\*-GRE, (B) SWI, and (C) QSM modalities.

candidate detection (Section 2.2): this step takes in the preprocessed input images, performs fast radial symmetry transform (FRST), and applies a deep learning model on both input image and FRST output to generate an initial CMB candidate detection map. (2) CMB candidate discrimination (Section 2.3): the initial candidate detection map is taken as input and a knowledge distillation framework, involving multi-tasking teacher and student networks, is used on this input image as well as the FRST output, to discriminate between CMBs and non-CMB candidates (obtained from step 1). (3) Post-processing (Section 2.4): finally, the CMB discrimination map obtained from step 2 is fed into a post-processing step, which uses anatomical constraints to further reduce false positives from the discrimination map. Sections 2.5 and 2.6 provide information regarding the training and implementation details of the method.

## 2.1. Data pre-processing

We reoriented the T2\*-GRE, SWI and QSM images to match the orientation of the standard MNI template, and skull-stripped the images using FSL BET (Smith, 2002). For T2\*-GRE and SWI, we performed bias field correction using FSL FAST (Zhang et al., 2001). We also inverted the intensity values of the input volume by subtracting the intensity-normalized image (obtained by dividing intensity values by the maximum intensity) from 1, so that CMBs have higher intensities (a design choice to facilitate our choice of CNN layers—e.g., max-pooling layers). For QSM images, we only normalized the intensity values without inverting their intensity values since CMBs already appear hyperintense with respect to the

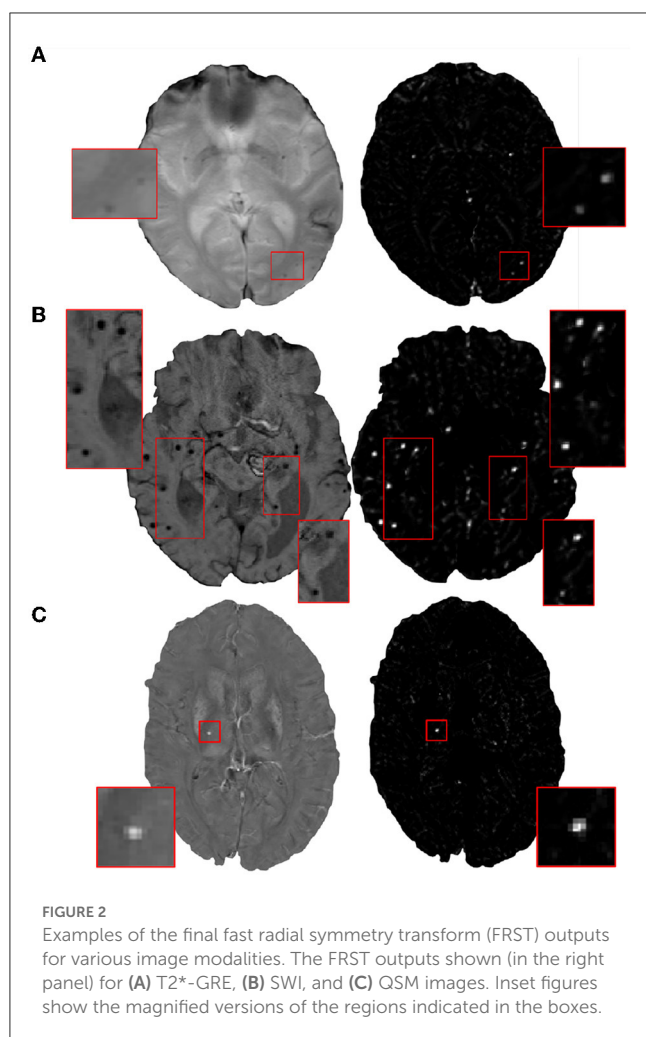
background. We cropped the skull-stripped images closer to the brain edges to make the FOV tighter.

We then removed blood vessels, sulci and other elongated structures from the input image to reduce the appearances of CMB mimics using the method described in Sundaresan et al. (2022). Briefly, the method involves the extraction of edge and orientation-based features, using Frangi filters (Frangi et al., 1998) and eigenvalues of the structure tensor (Förstner, 1994), followed by  $K$ -means clustering to obtain the vessel masks. The masked regions were then inpainted using the mean of intensity values from the immediate non-masked neighboring voxels (within a 26-connected neighborhood). Figure 1 shows a few sample images (from various modalities) after the removal of vessels and sulci.

## 2.2. 3D CMB initial candidate detection

In the initial candidate detection step, our main aim is to detect CMB candidates with maximum sensitivity, despite variations in their intensity characteristics and presence of mimics. The shape and size of CMBs are the main characteristics that could help differentiate them from mimics [e.g., flow voids, micro-hemorrhages, partial volume artefacts Greenberg et al. (2009)]. Since CMBs are circular, for the initial CMB candidate detection, in addition to the intensity characteristics, we also use the radial symmetry property of CMBs. We performed a fast radial symmetry transform (FRST) (Loy and Zelinsky, 2002) which uses a gradient-based operator to detect voxels with high radial symmetry. We calculated FRST at four radii (2, 3, 4, and 6 voxels) and then used their mean as the final FRST output (shown for different modalities in Figure 2).





During the training phase, for both the input modality and the FRST output, we split the 3D volumes into patches of size  $48 \times 48 \times 48$  voxels and provided them as two input channels to the 3D patch-based encoder-decoder model for initial candidate detection. We selected the patch-size of 48 voxels empirically—at this scale, the patches were large enough to overcome the effect of local noise and assign higher probabilities to CMB-like regions on experimented datasets described in Section 3. Note that we use patches only for training. During testing, we apply the trained model on whole 3D images.

Figure 3 shows the block diagram of this initial candidate detection step and the architecture of the 3D encoder-decoder model. The architecture of the 3D encoder-decoder network at a scale  $N$  is based on a shallow U-Net. We trimmed the U-Net to a shallow architecture with two pooling layers. Since CMBs are small and sparse, the model is required to detect low-level features in limited-size patches, rather than global features generally learnt for larger lesions (e.g., stroke lesions) at the image-level. The choice of a shallow architecture is also in line with prior literature supporting their use when data is scarce (Amiri et al., 2019; Du et al., 2020). The input channels are converted into three channels by the initial  $1 \times 1 \times 1$  projection layer, followed by  $3 \times 3 \times 3$  convolution to get the initial filter channel depth of 64. The architecture consists of

two consecutive  $3 \times 3 \times 3$  convolutional layers followed by the  $2 \times 2 \times 2$  max-pooling layer (in the encoder) or  $2 \times 2 \times 2$  upsampling layer (in the decoder). We added a  $1 \times 1 \times 1$  convolutional layer before the final softmax layer for predicting the probability maps  $P_{Cdet}$ .

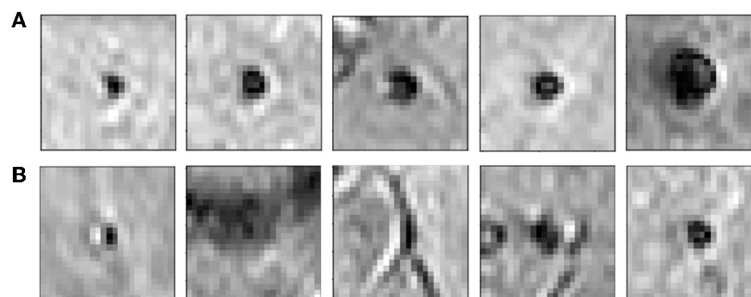
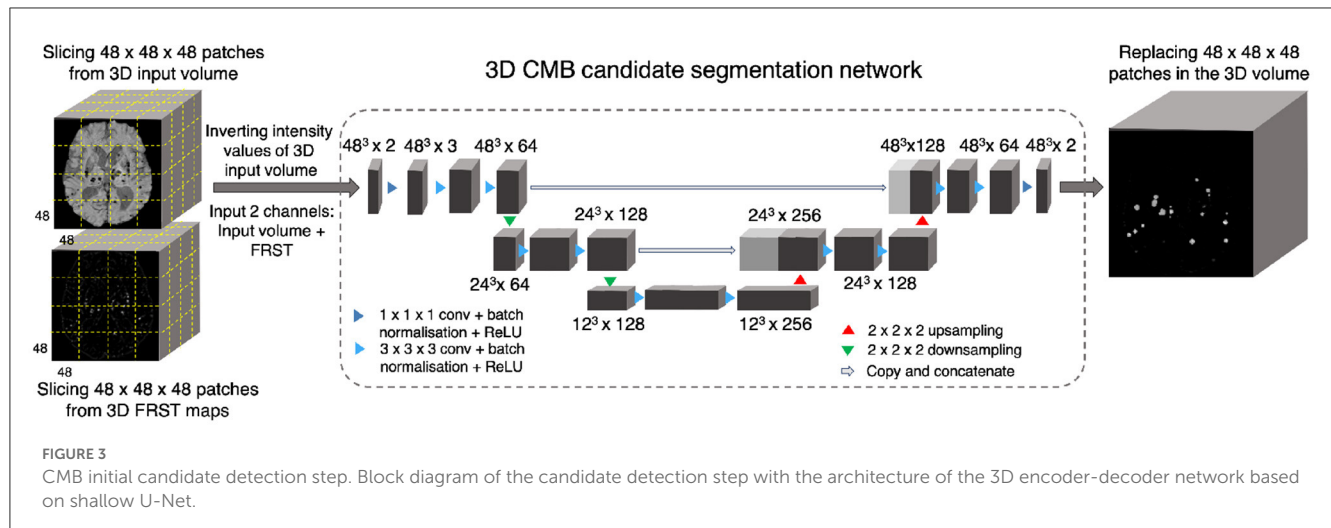
We used a combination of cross-entropy (CE) and Dice loss functions as the total loss function. In the CE loss function, we upweight the CMB voxels 10 times compared to the non-CMB voxels during training to compensate for the imbalance in the classes. Dice loss is based on the voxel-wise Dice similarity measure and aids in the accurate detection of edges and small CMBs in the patches.

## 2.3. CMB candidate discrimination

The candidate discrimination step is more challenging than the initial candidate detection step, since the discrimination step needs to learn the subtle features to detect CMBs and discriminate them from other CMB mimics. To illustrate the complexity of the problem, Figure 4 shows instances of CMB and non-CMB patches that were all identified as CMB candidates in the initial detection step. In this step, we use a student-teacher framework for classifying true CMB candidates from FPs. We use two networks: (1) a teacher network that has a multi-tasking architecture and learns the task-based characteristics (in our case, CMB-related features) from a larger dataset of patches extracted in a sliding manner from the image; (2) a student model that has comparatively simpler architecture and is trained directly on more contextual patches centered at candidates detected from the initial candidate detection step (Section 2.2). For training the teacher model (and candidate detection stage), non-overlapping patches are used. They provide sufficient sample CMB candidates to achieve the main aim at this stage, which is to train the model with examples of CMBs in a more comprehensive manner (especially with a multitasking teacher model performing both classification and segmentation tasks). We aim to improve the classification accuracy of the student model, by guiding its training by using the information from the teacher model with response-based knowledge distillation. For both candidate detection and discrimination stages, patches were extracted only for training, whereas the trained models were applied on whole 3D images during testing. Figure 5 shows the proposed overall architecture, while the details of the student-teacher architecture and training are provided in the sections below.

### 2.3.1. Teacher network with multi-task training

The teacher model uses a multi-tasking architecture consisting of three parts (1) feature extractor ( $T_f$ ), (2) voxel-wise CMB segmentor ( $T_s$ ) and (3) patch-level CMB classifier ( $T_c$ ). For the multi-tasking teacher model, we took the pretrained weights from the initial CMB candidate detection for  $T_s$  and added a patch-level CMB classifier arm ( $T_c$ ) using truncated normal variables for weight initialization (refer to Section 2.6). This helps to train the teacher model with more relevant initial weights for  $T_s$ , since pretrained weights for  $T_s$  were already available from the candidate



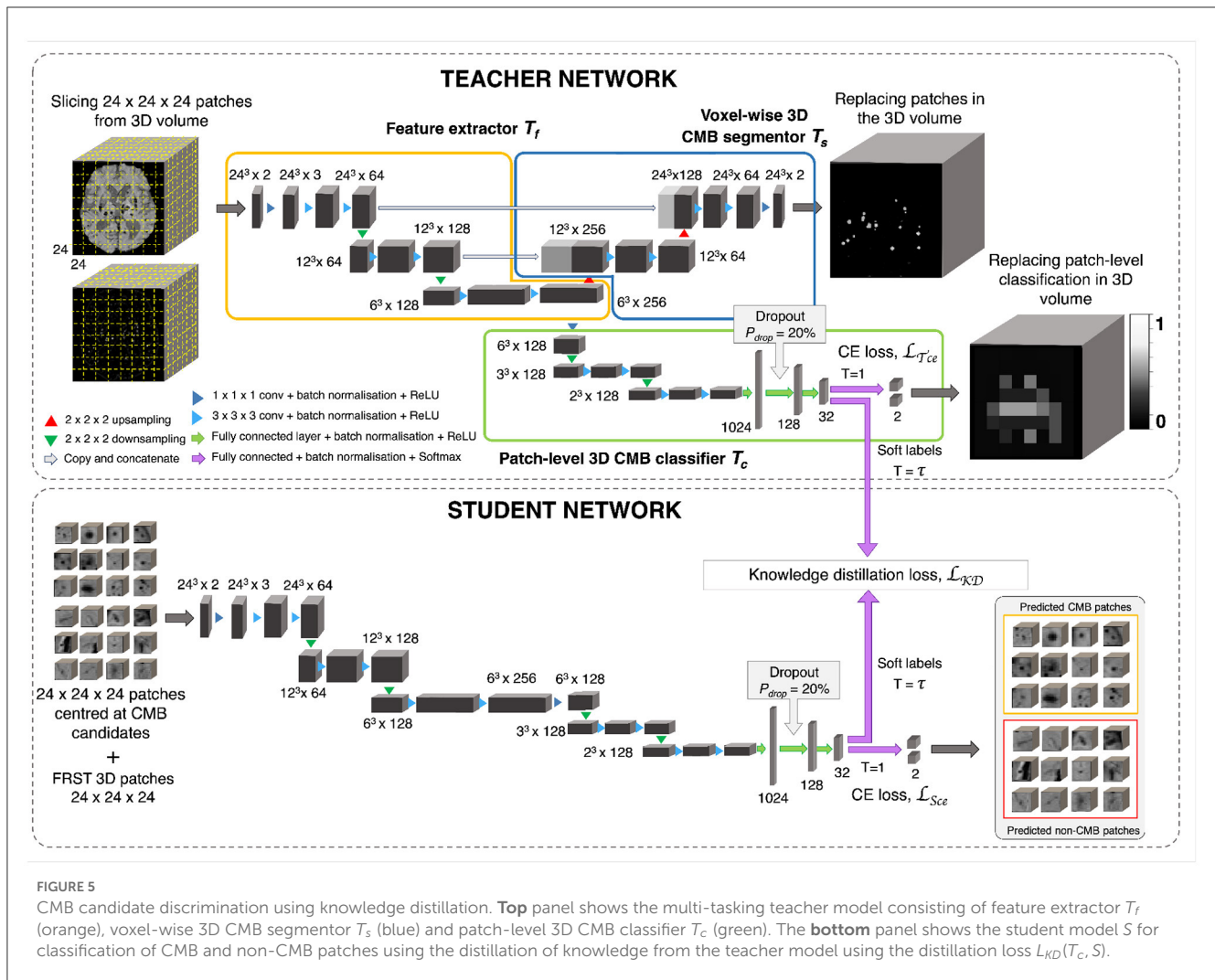
**FIGURE 4**  
Examples of initial candidates detected in the first step. (A) CMB and (B) non-CMB patches are shown separately. Note that in most of the cases, non-CMB instances are quite similar to CMBs.

detection step. However, note that the teacher model can also be trained with random initialization for both  $T_s$  and  $T_c$ . While the architecture of feature extractor + segmentor is the same as that of the model used in the initial candidate detection stage, the classifier arm consists of a projection layer with  $1 \times 1 \times 1$  kernel, followed by two consecutive  $3 \times 3 \times 3$  convolutional layers followed by a pooling layer in each level of abstraction. The output of the third layer of the encoder is fed into dense fully connected layers (FC). Three fully connected layers (FC-1024, FC-128, and FC-32) with 1,024, 128, and 32 nodes are then followed by a softmax layer. We added a dropout layer with a *drop probability* of 20% before the FC-128 layer. We extracted  $24 \times 24 \times 24$  adjacent patches from the input modality and FRST images, and provide them as 2-channel inputs for training this model. While we used a patch-size of 48 for the detection stage, we used a smaller patch-size of 24 for this stage. This is because our main aim was to adapt the segmentor arm  $T_s$ , initialized with pretrained weights (trained on patch-size of 48), to learn the lesion-level characteristics of initial CMB candidates from the local neighborhood obtained using smaller patches. The  $T_f$ , made of a series of convolutional layers, extracts features that are helpful for both  $T_s$  and  $T_c$ . Therefore, both  $T_s$  and  $T_c$  learn to improve the CMB segmentation and classification in a progressive manner since both are trained simultaneously with shared weights in  $T_f$ . This means that  $T_s$  assigns high probability values to the CMB

voxels in the CMB patches, while reducing the probability values of CMB-like mimics on the non-CMB patches. At the same time,  $T_c$  detects the patches with more CMB-like features (regions that are assigned higher probabilities by  $T_s$ ) as CMB patches with higher confidence and vice versa. In addition to the loss function to train  $T_s$  (specified in Section 2.2), we used a binary cross-entropy loss function for  $T_c$ .

### 2.3.2. Knowledge distillation using student network

The student model consists of a feature extractor and a patch-level classifier parts ( $T_f + T_c$ ), as the teacher model. However, while we provided non-overlapping, adjacent  $24 \times 24 \times 24$  patches for the teacher model, we extracted more meaningful input patches for the student model, centered at the detected initial CMB candidates for quicker learning. We trained the student model in an offline manner using response-based knowledge distillation (KD). For determining the centroids of the patches, we thresholded  $P_{Cdet}$  from the first step at a specific threshold  $Th_{Cdet}$  based on the performance values (for more details refer to Section 5.2). During testing, patches centered at candidates detected from the initial candidate detection steps are classified as CMB or non-CMB by the student model. Let the student model and teacher model classifier be  $S$  and  $T_c$ ,



respectively. For the distillation of knowledge from the teacher model for training the student model, the loss function is given by

$$L = \alpha * L_S + \beta * L_{KD}(T_c, S) \quad (1)$$

where  $L_S$  is the student loss function,  $L_{KD}(T_c, S)$  is the KD loss and  $\alpha, \beta$  are weighing parameters. We used the cross-entropy loss function as the student loss. For determining the KD loss, the targets are the class outputs predicted by the classifier of the teacher model (in the inference mode) on the same input as that of the student model. A temperature ( $\tau$ ) parameter is used in the softmax function to soften the target distribution. While  $\tau = 1$  provides the usual softmax outputs, higher values of  $\tau$  soften the softmax outputs (as shown in Equation 2). The softmax function with  $\tau$  is given by,

$$\sigma(z_i, \tau) = \frac{\exp(z_i/\tau)}{\sum_{j=1}^N \exp(z_j/\tau)} \quad (2)$$

where  $N$  is the number of classes. Compared to hard target distributions (closer to 0 or 1 for individual classes), softer target distributions (between 0 and 1) have been shown to aid in training

a generalizable student model (Hinton et al., 2015), however, having very high  $\tau$  might also be counter-productive in some cases. The optimal value of  $\tau$  and the level of softness in the target distribution depends on specific applications, student/teacher network architectures and dataset characteristics. Temperature  $\tau$  values between 2.5 and 4 have been shown to provide better results, while models with more units in the hidden layers may require higher  $\tau$  values (Hinton et al., 2015). Using the temperature  $\tau$  parameter, the KD loss is given by,

$$L_{KD}(T_c, S) = KL(\sigma(z_S, \tau), \sigma(z_{T_c}, \tau)) \quad (3)$$

where  $KL$  is the KL-divergence (distance between the class probability distributions of student and teacher classifier models). From Equation 1, the loss function is,

$$L = \alpha * CE(y_S, \sigma(z_S, \tau = 1)) + \beta * KL(\sigma(z_S, \tau), \sigma(z_{T_c}, \tau)) \quad (4)$$

where  $y_S$  are the target labels of the student model and  $z_S$  and  $z_{T_c}$  are the logits (inputs to the softmax layer) of the student and teacher classifier model, respectively.

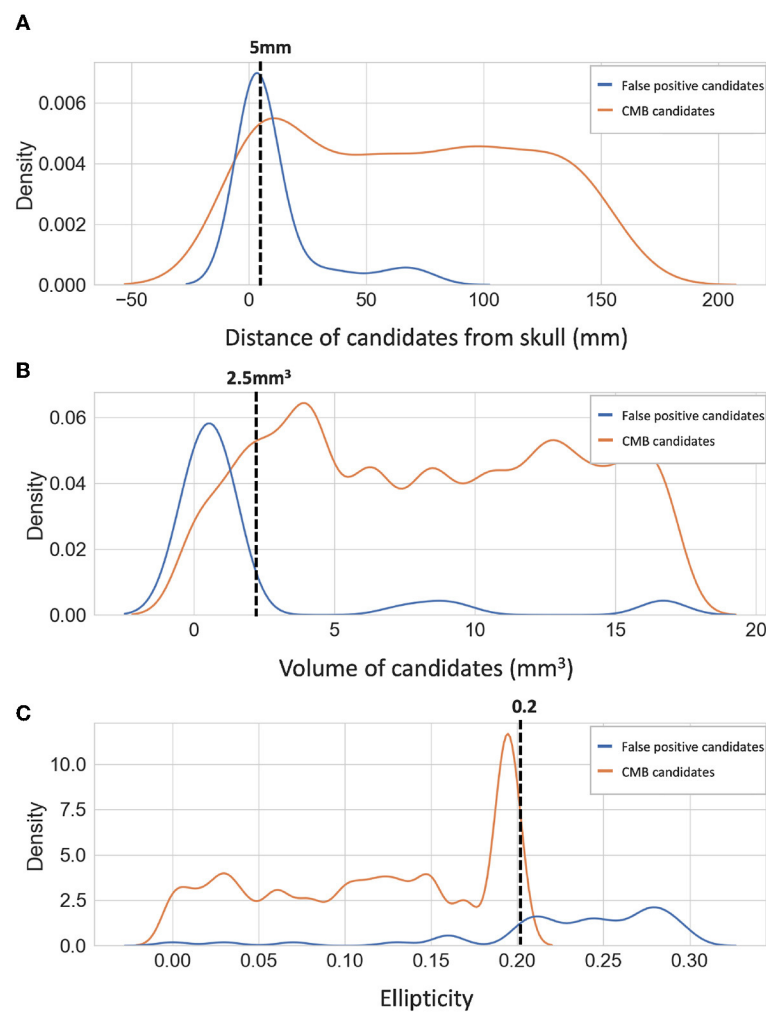


FIGURE 6

Density plots of candidate-level attributes. Density plots of (A) distance from the skull (mm), (B) volume ( $\text{mm}^3$ ), and (C) ellipticity of true CMB candidates (orange) and false positive candidates (blue) on an independent dataset (see Section 2.6). The filtering criteria are marked on the density plots with dotted lines.

## 2.4. Post-processing

We applied a threshold  $Th_{Cdisc}$  on the probabilities to discriminate CMB and non-CMB candidates. We set  $Th_{Cdisc}$  values empirically based on the performance metric values (refer to Section 5.2). Additionally, we removed the noisy stray voxels by filtering out the candidates with volume  $< 2.5\text{mm}^3$ , removed the tubular structures (e.g., fragments of sulci near the skull) by filtering out candidates having higher ellipticity ( $> 0.2$ ) and removed the CMB candidates that are closer to the skull ( $< 5\text{mm}$  from the brain mask boundary) to reject the FPs due to the sulci in the brain. The density plots for the above attributes for false positive and true CMB candidates on an independent dataset (that was used for hyperparameter tuning as specified in Section 2.6) are shown along with the cut-off criteria values in Figure 6. Note that this dataset was not later used for training or testing in the evaluation. The cut-off criteria values were determined empirically based on the attribute values as shown in density plots.

## 2.5. Data augmentation

Due to the small size of CMBs, transformations such as rotation and down-scaling could result in the loss of CMBs in the augmented data patches. Hence, we chose our data augmentation carefully, to inject variations in the data with minimal interpolation of intensity values. For the initial candidate detection step, we performed augmentation on the patches, increasing the dataset size by a factor of 10, using random combinations of the following transformations: translation, random noise injection and Gaussian filtering (with a small  $\sigma$  value). The parameters for the above transformations were chosen randomly from the ranges as specified as follows: (1) Translation:  $x$ -offset:  $[-15, 15]$ ,  $y$ -offset:  $[-15, 15]$  voxels, (2) Random noise injection: Distribution - Gaussian,  $\mu = 0$ ,  $\sigma^2 = [0.01, 0.04]$ , (3) Gaussian filtering:  $\sigma = [0.1, 0.2]$  voxels. We used similar augmentation for the discrimination step, increasing training data by a factor of 5.



## 2.6. Implementation details

For both candidate detection and discrimination steps, patches were extracted only for training. During the testing phase, the trained models were applied to the whole 3D images. For both CMB candidate detection and discrimination steps, we trained the networks using the Adam Optimiser (Kingma and Ba, 2014) with an epsilon ( $\epsilon$ ) value of  $1 \times 10^{-4}$ . We used a batch size of 8, with an initial learning rate of  $1 \times 10^{-3}$  and reducing it by a factor  $1 \times 10^{-1}$  every 2 epochs, until it reaches  $1 \times 10^{-6}$ , after which we maintain the fixed learning rate value. For both candidate detection and candidate discrimination training, we also empirically set the total number of epochs to 100 and used a criterion based on a patience value (number of epochs to wait for progress on validation set) of 20 epochs to determine model convergence for early stopping. We used the truncated normal variable (with  $\sigma = 0.05$ ) for weights initialization and the biases were initialized as constants with a value of 0.1. For the student model in the discrimination step, we used a temperature  $\tau = 4$ ,  $\alpha = 0.4$ , and  $\beta = 0.6$ , determined empirically using a trial-and-error method. We used a subset of a publicly available dataset (<https://valdo.grand-challenge.org/Data/>) consisting of a random sample of 20 subjects for hyperparameter tuning and the empirical determination of parameter values in the post-processing step and loss functions. The initial candidate detection model took  $\sim 15$  min per epoch for training and converged in around 80 epochs. For the candidate discrimination step, the teacher model took  $\sim 20$  min per epoch for training and converged in around 80 epochs (this model was trained only once and was always used in inference mode for training student model on various datasets during cross-validation). The student model took only  $< 5$  min per epoch and converged in  $\sim 50$  epochs, thereby effectively reducing the training time for various datasets. We implemented the networks on NVIDIA Tesla V100 GPU, in Python 3.6 using Pytorch 1.2.0.

## 3. Datasets used

We used four datasets for the evaluation of our proposed method. The datasets consist of images from different modalities, that were acquired using different scanners with variations in acquisition protocols and from subjects with different pathological conditions and demographic characteristics. Histograms of subject-level CMB counts for the individual datasets (along with zoomed-in histograms for subjects with CMB count  $< 10$ ) are shown in Supplementary Figure S1.

### 3.1. The UK Biobank (UKBB) dataset

From 14,521 subjects with usable imaging and non-imaging variables of interest (out of  $\sim 20,000$  subjects from the January 2018 release of UKBB), we preselected 78 CMB candidate subjects using the method proposed in Sundaresan et al. (2022). Manual segmentations in the form of coordinates were annotated on SWI images for these 78 subjects by a trained radiologist (A.G.M). From those coordinates, the ground truth segmentation for each CMB was obtained by a region-growing-based method that in addition

to a voxel's intensity also takes into account its distance from the seed voxel, and is constrained by a maximum radius of five voxels in-plane and three voxels through-plane. The age range of subjects is 50.8–74.8 years, mean age  $59.9 \pm 7.2$  years, median age 57.8 years, female to male ratio F:M = 37:41. For SWI, 3D multi-echo GRE images were acquired using 3T Siemens Magnetom Skyra scanner with TR/TE = 27/9.4/20 ms, flip angle  $15^\circ$ , voxel resolution of  $0.8 \times 0.8 \times 3$  mm, with image dimension of  $256 \times 288 \times 48$  voxels. The QSM images were generated using a multi-step post-processing of phase data as described in Wang et al. (2021). Briefly, the method involved a combination of phase data of individual channels, phase unwrapping, background field removal, followed by dipole inversion. Total number of CMBs in this dataset: 186, mean:  $2.4 \pm 7.0$  CMBs/subject, median: 1 CMB/subject.

### 3.2. The Oxford Vascular Study (OXVASC) dataset

The dataset consists of T2\*-GRE images from 74 participants from the OXVASC study (Rothwell et al., 2004), who had recently experienced a minor non-disabling stroke or transient ischemic attack. The 2D single-echo T2\*-GRE images were acquired using 3T Siemens Verio scanner with GRAPPA factor = 2, TR/TE = 504/15 ms, flip angle  $20^\circ$ , voxel resolution of  $0.9 \times 0.8 \times 5$  mm, with image dimension of  $640 \times 640 \times 25$  voxels. Age range 39.6–91.2 years, mean age  $69.8 \pm 14.6$  years, median age 67.3 years, female to male ratio F:M = 36:38. Out of 74 subjects, 36 subjects had CMBs, and manual segmentations, labeled using T2\*-GRE images, were available for all 36 subjects. Total number of CMBs: 366, mean:  $10.2 \pm 33.3$  CMBs/subject, median: 3 CMBs/subject.

### 3.3. The tranexamic acid for intraCerebral hemorrhage 2 (TICH2) trial MRI sub-study dataset

The dataset consists of a subset of the MRI data used in (Pszczolkowski et al., 2022) obtained as part of the TICH2 trial (Sprigg et al., 2018). The dataset consists of images with variations in image dimension, spatial resolution and MR acquisition parameters (details in Dineen et al., 2018). The dataset used in this work consists of 115 SWI from the subjects with spontaneous intracerebral hemorrhage (ICH). Age range 29–88 years, mean age  $64.76 \pm 15.5$  years, median age 66.5 years, female to male ratio F:M = 24:26. Out of 115 subjects, 71 subjects had CMBs and manual segmentations for CMBs were available for all 71 subjects. Additionally, microbleed anatomical rating scale (MARS, Gregoire et al., 2009) values were provided for the CMB subjects. For evaluation purposes, we included in the manual segmentation maps used in all our experiments all CMBs that were labeled as either “definite” or “possible”. Total number of CMBs: 849, mean:  $11.9 \pm 22.0$  CMBs/subject, median: 3 CMBs/subject.

### 3.4. The stroke dataset from Hong Kong (SHK)

Originally, the dataset used in Dou et al. (2016) consisted of 320 SWI images in total, out of which 126 are subjects with stroke (mean age:  $67.4 \pm 11.3$ ) and 194 are from normal aging subjects (mean age:  $71.2 \pm 5.0$ ). In this work, we used a subset of 20 subjects that were publicly available from this dataset. Manual annotations in the form of CMB coordinates were available along with the dataset. From coordinates, ground truth segmentations were obtained with the same method used for the UKBB data (refer to Section 3.1). Another rater independently provided the manual segmentations on SWI images on the dataset, and we considered the union of both manual masks as our final ground truth. Total number of CMBs: 126, mean:  $6.3 \pm 8.8$  CMBs/subject, median: 3 CMBs/subject.

## 4. Experiments

### 4.1. Performance evaluation metrics

We evaluated the CMB detection results at the lesion-level using the following metrics for a total number of CMBs over the individual datasets, as done in the existing literature:

- **Cluster-level TPR:** the number of true positive clusters (i.e., CMBs) divided by the total number of true clusters as given by,

$$\text{cluster-wise TPR} = \frac{TP_{clus}}{(TP_{clus} + FN_{clus})} \quad (5)$$

where  $TP_{clus}$  and  $FN_{clus}$  are true positive (overlaps with a ground truth cluster by at least one voxel) and false negative clusters, respectively.

- **Average number of FPs per subject (FPavg):** for a given dataset  $D$ , FPavg is defined as the ratio of the total number of detected FP clusters ( $FP_{clus}$ , has no overlap with a ground truth cluster) to the number of subjects (or images) in the dataset, as given by,

$$FPavg = \frac{\text{Total number of } FP_{clus}}{\text{Number of subjects in } D} \quad (6)$$

- **Cluster-wise precision:** the number of true positive clusters divided by the total number of detected clusters as given by,

$$\text{cluster-wise precision} = \frac{TP_{clus}}{(TP_{clus} + FP_{clus})} \quad (7)$$

We used 26-connectivity to form the clusters. In general, for cluster-wise TPR and cluster-wise precision, the higher the values the better while for FPavg, lower values are better. We used TPR and FPavg values for plotting a free-response receiver operating characteristics (FROC) curve, which is a plot of cluster-wise TPR vs. the average number of false positives per image/subject.

### 4.2. Ablation study: effect of knowledge distillation on CMB detection within the UKBB dataset

In this study, we evaluate the effect of individual steps, including the teacher-student distillation framework and the post-processing step on the CMB detection performance (using metrics specified in Section 4.1) on the UKBB dataset (using a training-validation-test split of 44-10-24 subjects, with 40 CMBs in the test data). To this aim, we calculated the above performance evaluation metrics at the following stages: after the initial CMB candidate detection (i) without using FRST output, (ii) using FRST output as an additional input channel, after candidate discrimination (iii) using the teacher model  $T_c$  alone, (iv) a classification network trained without the teacher model (trained independently using only CE loss function  $L_S$ ), (v) with knowledge distillation using student-teacher training and (vi) after final post-processing. For the classification model in (iv), we used only the student model architecture and provided as inputs the patches centered at the detected initial CMB candidates. For the student-teacher architecture used in (v), we provided adjacent patches sampled in a sliding manner as inputs for the teacher model and patches centered at initial CMB candidates as inputs for the student model.

### 4.3. Cross-validation of CMB detection on T2\*-GRE and SWI images within individual datasets

We performed 5-fold cross-validation separately on T2\*-GRE images from the OXVASC dataset and SWI images from the UKBB dataset, and evaluated the cluster-wise performance using the metrics specified in Section 4.1 for the whole dataset across subjects. Note that in both cases, for the candidate discrimination step, we used the teacher model pretrained on the UKBB dataset (for response-based distillation), while only the student model was trained on the individual datasets. Also, for this cross-validation, we used the hyper-parameters that were determined separately using an independent dataset specified in Section 2.6.

### 4.4. Evaluation of the generalizability of the proposed method across different datasets

For this experiment, we trained the proposed method on SWI images from 78 subjects from the UKBB dataset. We chose UKBB as training data since it is a large, open-access epidemiological study. The manually annotated subset of UKBB dataset used in this work provided larger training data than the other datasets. Moreover, UKBB dataset consists of both SWI and QSM modalities, the former is commonly used for detecting CMBs and the latter is effective for removing false positives. We used the hyperparameters mentioned in Section 2.6 and evaluated the trained model on data from different domains (e.g., variations in intensity profiles, scanners and acquisition protocols and demographics), using

performance metrics specified in Section 4.1, under the following three scenarios:

1. Evaluation on the same dataset with different modalities: training data—UKBB (SWI), test data: UKBB (QSM)
2. Evaluation on different datasets with the same modality: training data—UKBB (SWI), test data: TICH2 (SWI), SHK (SWI)
3. Evaluation on different datasets with different modalities: training data—UKBB (SWI), test data: OXVASC (T2\*-GRE)

### 1. Evaluation on the same dataset with different modalities:

We evaluated the effect of change in the modality only on CMB detection by applying the method, that was trained on SWI images from the UKBB dataset, to the QSM images (intensities non-inverted for QSM, hence CMBs appear brighter than the background, similar to the preprocessed SWI) from the same subjects from the UKBB dataset.

### 2. Evaluation on different datasets with the same modality:

We evaluated our method on different test datasets to observe the effect of scanner-related and population-level pathological variations on the CMB detection. We applied our method trained on SWI images from the UKBB dataset to SWI images from 115 subjects with intra-cerebral hemorrhages from the TICH2 dataset and SWI images from 10 healthy controls and 10 subjects with stroke from the SHK dataset.

### 3. Evaluation on different datasets with different modality:

We evaluated our method trained on SWI images from the UKBB dataset on the T2\*-GRE images from 74 subjects from the OXVASC dataset. The OXVASC data is quite different from the UKBB data not only in terms of modality, but also in terms of resolution, scanner and demographic/pathological factors. Hence, this scenario would provide a better indication of the method's generalizability in real world clinical applications.

For the above experiments, for the CMB candidate detection and discrimination steps, we used the threshold values ( $Th_{Cdet}$  and  $Th_{Cdisc}$ ) determined during 5-fold cross-validation on the UKBB dataset.

## 4.5. Indirect comparison of our results with the existing literature

Finally, we performed an indirect comparison of our results from the UKBB and OXVASC datasets with those of existing CMB detection methods in the literature.

## 5. Results

### 5.1. Ablation study: effect of knowledge distillation on CMB detection within the UKBB dataset

Figure 7 shows the FROC curves for the initial CMB candidate detection and candidate discrimination steps of our method on the UKBB dataset. Table 1 reports the best performance points at the 'knee-point' on the FROC curves for the first two steps, along with the performance metrics after the third step (post-processing). In the candidate detection step, the aim was to achieve

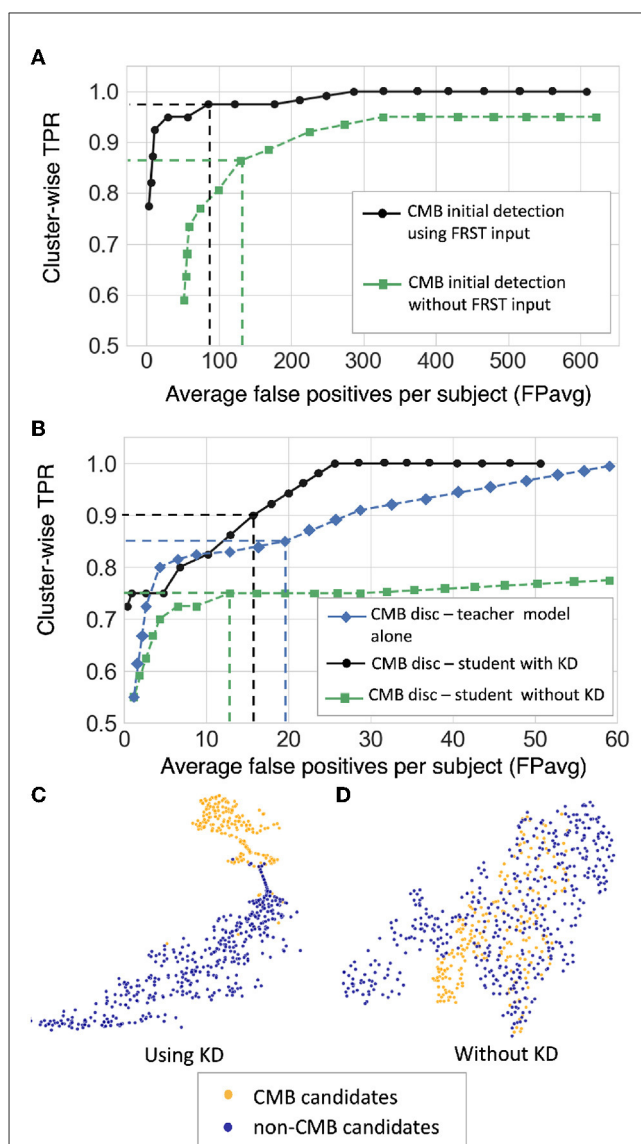


FIGURE 7

Results of the ablation study. (A) FROC curves at the CMB initial candidate detection stage using FRST output as an additional input channel (black solid ●) and without using FRST output (green dashed ■), (B) FROC curves comparing the classification performance of the teacher model  $T_c$  (blue dashed ◆), the student model trained using KD from a teacher model (black solid ●) and the same model trained independently without KD (green dashed ■). The horizontal and vertical dashed lines on the FROC curves indicate the best performance points determined at the knee-point of the curve with higher TPR (reported in Table 1) at  $Th_{Cdet} = 0.5$  at the candidate detection step,  $Th_{Cdisc} = 0.29, 0.3$ , and  $0.35$  for the teacher model, models with and without KD, respectively in the candidate discrimination step. T-SNE plots showing feature embeddings at the FC-32 layer for CMB (orange) and non-CMB (dark blue) cases for (C) the student model trained using KD and (D) the model without using KD. In (C), the plot shows better separability of features corresponding to CMB and non-CMB classes, indicating the ability of the model to discriminate well between these two classes.

higher cluster-wise TPR, to detect as many true CMBs as possible. Hence, the number of FPs was higher at this step (with the highest cluster-wise TPR of 0.975 at the  $Th_{Cdet} = 0.5$ ), when compared to the subsequent steps. Using the FRST output as an additional

TABLE 1 Ablation study: performance metrics after candidate detection, discrimination, and post-processing steps.

Steps	Cl. TPR	FPavg	Cl. prec
(i) Cand. det. without FRST	0.86	129.3	0.02
(ii) Cand. det. using FRST	0.975	85.3	0.03
(iii) Cand. disc. using teacher model $T_c$	0.85	19.6	0.09
(iv) Cand. disc. without KD	0.75	12.8	0.09
(v) Cand. disc. using K	0.9	14.7	0.11
(vi) After postproc.	0.83	0.5	0.74

Cl. TPR and Cl. prec indicate cluster-wise TPR and cluster-wise precision, respectively. FPavg values at a fixed cluster-wise TPR value of 95% are provided in the [Supplementary Table S3](#).

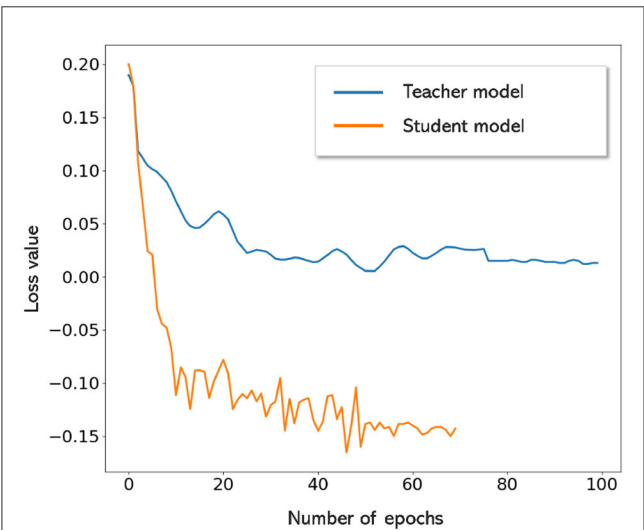


FIGURE 8  
Convergence of teacher and student models. Student model trained with KD converges quicker and to lower loss values when compared to the teacher model.

input channel improved the sensitivity with slightly lower FPavg (Figure 7A). For the candidate discrimination step (Figure 7B), the FROC curves are shown for the comparison of the teacher model classification arm  $T_c$ , student network trained with KD framework from the teacher model and classification network (with the same architecture as that of the student network) trained independently without KD from the teacher model. The performance at the candidate discrimination step is better with KD (cluster-wise TPR of 0.9 at  $Th_{Cdisc} = 0.3$ ) than the model trained without KD (cluster-wise TPR of 0.75 at  $Th_{Cdisc} = 0.35$ ), with the former showing an improvement of 0.02 in the cluster-wise precision (see Table 1). Also, the student model trained with KD performs better than the teacher model  $T_c$ , with the improvement of 0.02 in cluster-wise precision. Moreover, the student model trained with KD converges quicker and to a much lower loss value when compared to the teacher model as shown in Figure 8.

The t-stochastic neighbor embedding (t-SNE) plots show the separability between the features of the CMB and non-CMB classes as a measure of the model's ability to discriminate the classes from the learnt feature embeddings. For instance, the more separable the two classes are, the better the ability of the student model

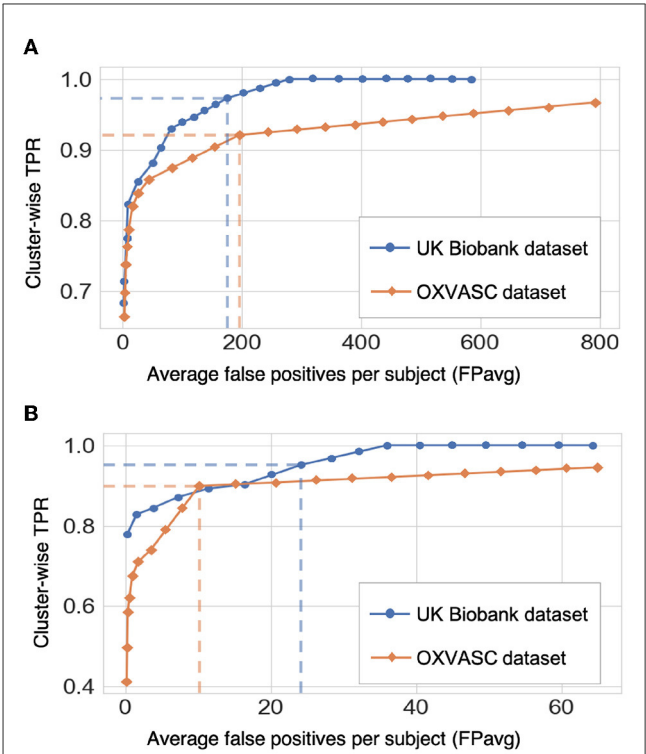


FIGURE 9  
Results of the 5-fold cross-validation. FROC curves at (A) the CMB initial candidate detection stage and (B) the candidate discrimination stage on the UKBB (blue) and OXVASC (orange) datasets. The dashed lines on the FROC curves indicate the best performance points (reported in Table 2) at specific threshold points (threshold = 0.3 and 0.2 for the UKBB and OXVASC datasets, respectively for both candidate detection and candidate discrimination steps).

to discriminate between CMB class and mimics. Figures 7C, D show the feature embeddings of the last fully connected layer (FC-32) in the classification network (in the CMB discrimination step) for the CMB and non-CMB cases, trained with and without KD. The feature embeddings for the student model using KD were quite separable between the CMB and non-CMB cases, indicating the capability of the student model to learn the subtle differences in the features between CMB and non-CMB classes, using the information from the multi-tasking teacher model. The classification model without KD, on the other hand, showed substantially more overlap between the feature embeddings. The post-processing step improves the cluster-wise precision. Upon visual inspection, the main reductions in FPs were near the skull (e.g., sulci), penetrating blood vessels and stray noisy voxels. Regarding the contribution of individual attributes (e.g., shape, area and proximity to the skull) in FP reduction, we observed around 65, 25, and 15% of FP reduction after applying thresholds on distance of candidates from the skull, area and shape of candidates successively. Since the individual thresholds were determined as a part of hyper-parameter tuning on an independent dataset (Section 2.4) and the interaction between the three attributes' thresholds on FPavg is difficult to visualize, a separate FROC curve for the post-processing step is not shown. The majority of FPs rejected at this stage consists of candidates closer to the skull -



**TABLE 2** Cross-validation on the UKBB and OXVASC datasets: performance metrics at candidate detection, discrimination, and post-processing steps.

Datasets	Steps	Cl. TPR	FPavg	Cl. prec
UKBB (SWI)	C. det.	0.97	175.4	0.01
	C. disc.	0.95	24.2	0.09
	Postproc.	0.93	1.5	0.59
OXVASC (T2*-GRE)	C. det.	0.93	195.7	0.02
	C. disc.	0.91	10.1	0.29
	Postproc.	0.90	0.9	0.84

Cl. TPR and Cl. prec indicate cluster-wise TPR and cluster-wise precision, respectively. C. det - candidate detection, C. disc - candidate discrimination. FPavg values at a fixed cluster-wise TPR value of 95% are provided in the [Supplementary Table S4](#).

these candidates passed the discrimination step since most of the CMBs in the training data (for the student model) were lobar CMBs and were closer to the skull. Hence the discrimination step (despite removing a large number of FPs near sulci) allows false predictions in this region. Having said that, it is worth noting that, in the post-processing step, a few true CMBs closer to the skull were also rejected as FPs, hence leading to a slight decrease in the cluster-wise TPR values.

## 5.2. Cross-validation of CMB detection on T2\*-GRE and SWI images within individual datasets

[Figure 9](#) shows the FROC curves for CMB candidate detection and candidate discrimination steps of 5-fold cross-validation on whole images across all subjects on the UKBB and OXVASC datasets. [Table 2](#) reports the best performance metrics at different steps of the cross-validation on the UKBB and OXVASC datasets. The proposed method achieved cluster-wise TPR values of 0.93 and 0.90 with FPavg of 1.5 and 0.9 at  $Th_{Cdet}$ ,  $Th_{Cdisc}$  = 0.3 and 0.2 on the UKBB and OXVASC datasets, respectively. The method provides higher cluster-wise TPR and FPavg values on SWI images (from the UKBB dataset) when compared to the T2\*-GRE images from the OXVASC dataset. Even though the FPavg values were comparable at the candidate detection step for both datasets, the student model at the candidate discrimination step provided much lower FPavg on T2\*-GRE images from the OXVASC dataset, thus providing a higher cluster-wise precision value. The FPavg values reduced substantially after the post-processing step with only a slight reduction in the cluster-wise TPR values.

[Figure 10](#) shows sample results of the cross-validation at various steps of CMB detection on the UKBB and the OXVASC datasets. In both UKBB and OXVASC datasets, the main sources of FPs in the initial candidate detection step are sulci, minor intensity/contrast variations in the brain tissue and small vessel fragments. While most of the penetrating blood vessels are segmented correctly as part of the background even at the candidate detection step (due to the vessel removal step, especially in the OXVASC dataset), the remaining FPs on/near the vessels are removed at the discrimination step. The post-processing step

further reduced the stray noisy voxels and sulci regions closer to the skull, resulting in very few FPs on both datasets.

## 5.3. Evaluation of the generalizability of the proposed method across different datasets

[Table 3](#) reports the performance metrics of the proposed method, when trained on the UKBB dataset and applied on the same dataset but different modality (UKBB QSM data), different datasets but the same modality (SWI from the TICH2 and SHK datasets) and different datasets and modality (T2\*-GRE from the OXVASC dataset). We used  $Th_{Cdet}$  and  $Th_{Cdisc}$  values of 0.3 (determined from the cross-validation on the UKBB dataset) on the probability maps at the candidate detection step and on the patch-level probabilities at the discrimination step. [Figure 11](#) shows sample results of the method, when applied on various datasets at various steps of CMB detection.

Out of all datasets, the method achieved the highest cluster-wise TPR on the QSM dataset. On this dataset, the results were on par with the cross-validation results on the UKBB SWI data (with a slight decrease in the cluster-wise TPR and precision on QSM data). We obtained FPavg values of 1.8 FPs/subject—the FPs candidates were mainly due to minor susceptibility changes in the tissue and penetrating small blood vessels.

The method gave a cluster-wise TPR of 0.82 on the TICH2 dataset, despite the presence of ICH lesions (third row in [Figure 11](#)) in all subjects. The method provided the highest FPavg values in the initial candidate detection step (FPavg = 289.3 FPs/subject), possibly due to ICH edges and texture. However, the candidate discrimination step reduced the number of FPs and lowered the FPavg value to 42.8 FP/subject, which is comparable with other datasets. Even then, we obtained the highest FPavg after post-processing on this dataset with a cluster-wise precision of 0.62. Susceptibility artefacts at tissue interfaces and sulci were mainly detected as FPs in this dataset.

On the SHK dataset, while the first two steps (CMB detection and discrimination) provided consistently good cluster-wise TPR values (>0.90), the TPR value decreased at the post-processing step. Even then, on this dataset our method provided the lowest FPavg (0.5 FPs/subject) and the highest cluster-wise precision among all the datasets. The FPs in the candidate detection step were mostly regions of intersections of blood vessels closer to the sulci, especially the central sulcus.

On the OXVASC dataset, the method achieved the lowest cluster-wise TPR of 0.81. The high number of false negatives in this dataset, as suggested by the lower cluster-wise TPR, could be due to the reduced contrast between CMBs and normal brain tissue, unlike the SWI data used for training. Also, occasionally true CMB candidates quite close to the skull were removed in the post-processing step, having been mistaken as sulci.

## 5.4. Indirect comparison with existing methods

[Table 4](#) provides a comparison of the proposed CMB detection method with existing fully automated methods. From the

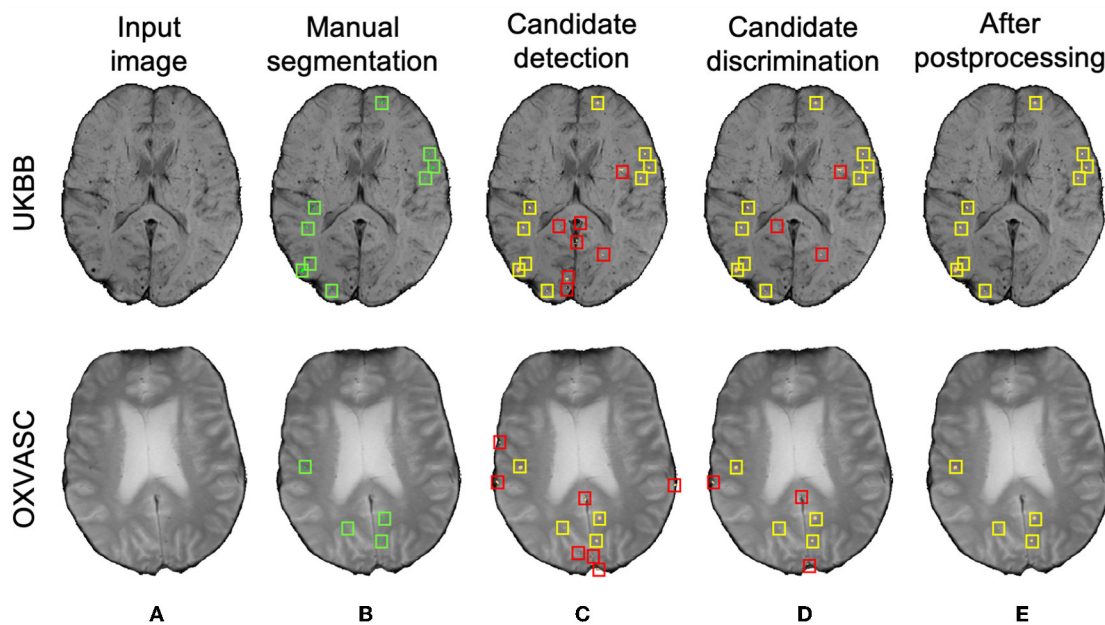


FIGURE 10

Sample cross-validation results on the UKBB (top panel) and the OXVASC (bottom panel) datasets. **(A)** Input image and **(B)** manual segmentations shown along with results at **(C)** CMB initial candidate detection, **(D)** candidate discrimination, and **(E)** after post-processing steps. True positive and false positive candidates are shown in yellow and red boxes, respectively for each step. Note that the displayed samples also include mimics (e.g., penetrating vessels and sulci) that appear similar to manually marked CMBs (in green boxes).

TABLE 3 Evaluation of the generalizability of the proposed method - trained on the UKBB SWI data and evaluated on the UKBB QSM, TICH2, SHK, and OXVASC datasets: performance metrics at candidate detection, discrimination and post-processing steps.

Datasets	Steps	Cl. TPR	FPavg	Cl. prec
UKBB (QSM)	C. det.	0.99	138.0	0.02
	C. disc.	0.91	40.3	0.04
	Postproc.	0.90	1.8	0.44
TICH2 (SWI)	C. det.	0.88	289.3	0.02
	C. disc.	0.83	42.8	0.10
	Postproc.	0.82	3.1	0.62
SHK (SWI)	C. det.	0.98	254.7	0.01
	C. disc.	0.94	43.6	0.09
	Postproc.	0.87	0.5	0.89
OXVASC (T2*-GRE)	C. det.	0.88	147.1	0.03
	C. disc.	0.85	53.7	0.07
	Postproc.	0.81	2.0	0.71

Cl. TPR and Cl. prec indicate cluster-wise TPR and cluster-wise precision, respectively. C. det, candidate detection; C. disc, candidate discrimination. FPavg at cluster-wise TPR value of 95% could not be provided since specific values of thresholds were applied.

table, generally deep-learning-based methods performed better compared to conventional machine learning methods. Also, the methods using multiple modalities or using phase information in addition to SWI (Ghafaryasl et al., 2012; Liu et al., 2019b; Al-Masni et al., 2020; Rashid et al., 2021) showed better results. In

fact, [Al-Masni et al. \(2020\)](#) showed that using phase in addition to SWI images improves the cluster-wise TPR by 5.6% (with only SWI: 91.6% and with SWI and phase: 97.2%) and [Rashid et al. \(2021\)](#) achieved the best CMB detection performance by using T2-weighted, SWI and QSM modalities. However, our proposed method uses a single modality (SWI or T2\*-GRE), along with the FRST images (obtained from the input modality itself) and gives comparable results to state-of-the-art methods such as [Liu et al. \(2019b\)](#) and [Al-Masni et al. \(2020\)](#), and with lower FPavg compared to [Dou et al. \(2016\)](#). Also, our precision values on both UKBB and OXVASC datasets are better than existing methods including [Bian et al. \(2013\)](#); [Fazlollahi et al. \(2014, 2015\)](#); [Dou et al. \(2016\)](#). Even though not directly relevant to our work (since we performed all our evaluations at the image-level rather than at patch-level), we have reported an additional comparison of patch-level methods in Supplementary Table S5. While the methods using patch-wise evaluation ([Chen et al., 2018](#); [Lu et al., 2021b](#)) provided good patch-level sensitivity and accuracies (shown in the Supplementary material), we cannot compare the performance of our method with those, since they used preselected CMB patches (from manually annotated CMB voxels) and comparable numbers of non-CMB patches as inputs. Also, the input CMB patches occasionally contained multiple CMBs, which makes the fair comparison with cluster-wise metrics highly difficult.

## 6. Discussion and conclusions

In this work, we proposed a fully automated, deep-learning-based, 3-step method for accurate lesion-level detection of CMBs

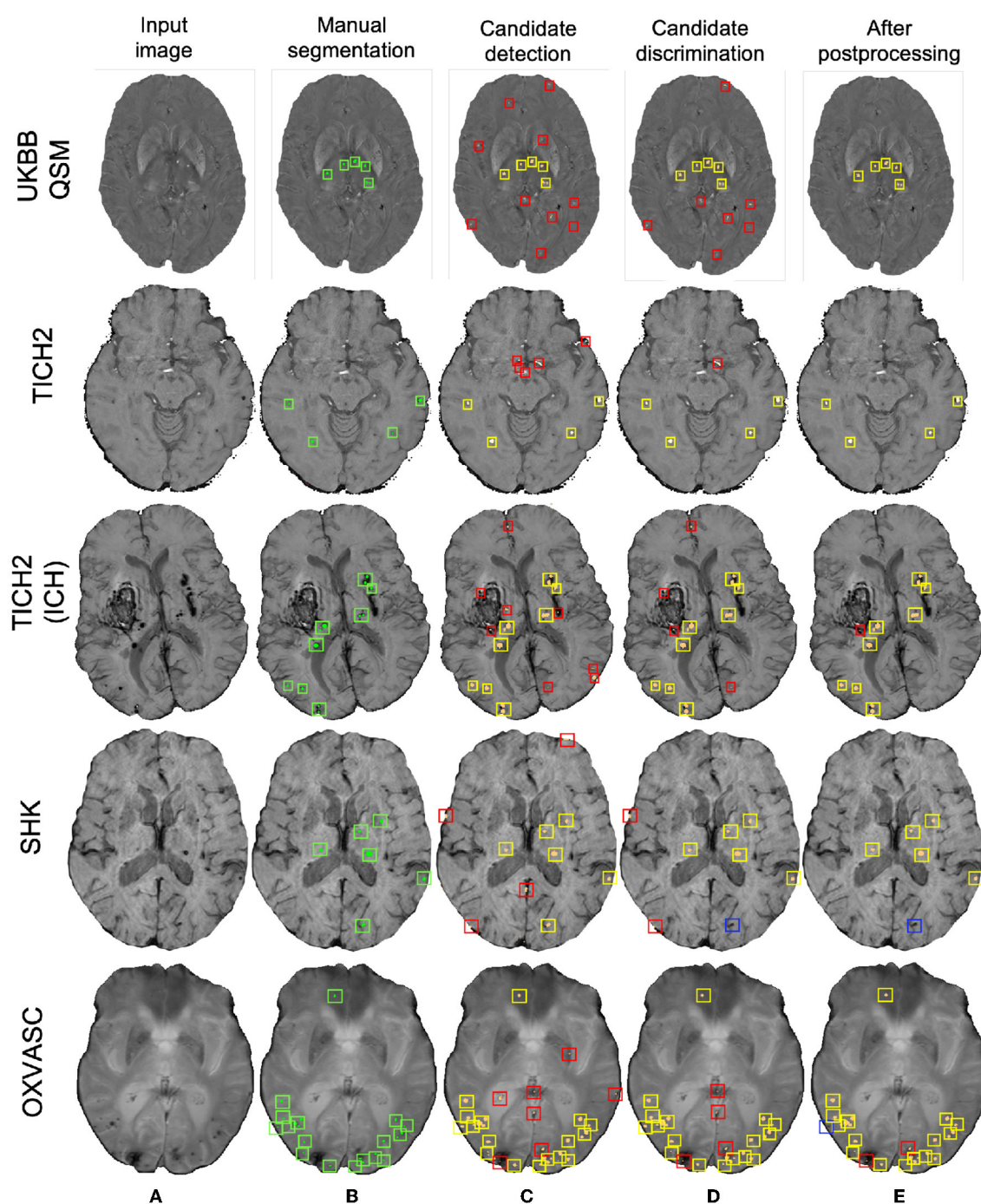


FIGURE 11

Sample results of the UKBB-trained method on the UKBB QSM, TICH2 (without and with ICH), SHK, and OXVASC datasets (from top to bottom panels). (A) Input image and (B) manual segmentations shown along with results at (C) CMB initial candidate detection, (D) candidate discrimination and (E) after post-processing steps. True positive, false positive and false negative candidates are shown in yellow, red, and blue boxes, respectively for each step. Note that the displayed samples also include mimics (e.g., penetrating vessels and sulci) that appear similar to manually marked CMBs (in green boxes).

on various datasets, irrespective of variations in population-level, scanner and acquisition characteristics. Our method uses a single input modality and the radial symmetry property of CMBs for the detection of CMB candidates with high sensitivity in the initial candidate detection step. For the candidate discrimination step, we trained a student classification network with the knowledge

distilled from a multi-tasking teacher network for accurate classification of CMB candidates from non-CMB candidates. Our ablation study results show that the candidate discrimination and post-processing steps drastically reduce the number of FPs, and the use of the KD framework improves cluster-wise TPR values at the discrimination step. Our method achieved cluster-wise TPR



**TABLE 4** Comparison of the performance of the proposed CMB detection method with existing conventional machine learning (ML) and deep learning (DL) methods.

Methods	Datasets		Performance		
	Sequence(s) (# test subjects)	Total # CMBs	Cl. TPR	FPavg	Cl. Prec
<b>ML methods</b>					
Bian et al. (2013)	SWI (10)	304	86.5%	44.9	
Fazlollahi et al. (2014)	SWI (41)	103	92%	FPavg <sub>CMB</sub> - 6.7 FPavg <sub>nCMB</sub> - 16.8	
Fazlollahi et al. (2015)	SWI (66)	231	87%	FPavg <sub>D</sub> - 10.28, FPavg <sub>P+D</sub> - 27.8	
Ghafaryasl et al. (2012)	T2*-GRE + PD (81)	183	91%	4.1	
Dou et al. (2015)	SWI (19)	161	80%	7.7	49%
Chesebro et al. (2021)	T2*-GRE, SWI (78)	64	95%	9.7 (SWI), 17.1 (T2*-GRE)	11% (SWI), 7% (T2*-GRE)
<b>DL methods</b>					
Chen et al. (2015)	SWI (5)	55	89%	6.4	56%
Dou et al. (2016)	SWI (50)	117	93%	2.74	44%
Liu et al. (2019b)	Phase + SWI (41)	168	96%	1.8	(5-fold CV)
Al-Masni et al. (2020)	Phase + SWI (72)	188	94.3%	1.4	61.9%
Rashid et al. (2021) (Leave-one-out validation)	QSM + SWI + T2w (24)	~172	89%		49%
Proposed method	UKBB - SWI (78)	186	93%	1.5	59%
	OXVASC - T2*-GRE (74)	366	90%	0.9	84%

Cl.TPR, cluster-wise TPR; FPavg, average false positives per image/subject; Cl.Prec, cluster-wise precision; FPavg<sub>CMB</sub>, FPavg for CMB subjects; FPavg<sub>nCMB</sub>, FPavg for non-CMB subjects; FPavg<sub>D</sub> - FPavg for “definite” CMB subjects; FPavg<sub>P+D</sub>, FPavg for “definite and possible” CMB subjects.

values >90% with FPavg of <1.6 FPs/subject during initial cross-validation on the UKBB and OXVASC datasets consisting of SWI and T2\*-GRE images, respectively. On training the models on the UKBB dataset and applying them on different datasets with different demographic and scanner-related variations, the method showed a good generalizability across datasets, providing cluster-wise TPR values >80% on all datasets.

The initial vessel removal pre-processing step helped in reducing the number of FPs since blood vessels (especially the small ones closer to sulci) are one of the common mimics of CMBs. One of the main challenges in the vessel removal step is the potential removal of true CMBs that are very close to vessels. Hence, we removed only linear segments with low uniform width in this step. Therefore, this step removed the vessels and sulci that were more prominent and could lead to obvious FPs. This was especially effective for the SHK dataset, where the blood vessels and sulci had higher contrast and were distinctly different from CMBs (Figure 1B). While removing the linear, elongated structures from the images, we also aimed to leverage the radial symmetry property of CMBs. Toward that aim, using FRST maps, obtained from the input modality, as an additional input channel helped the candidate detection model in learning contextual features, leading to the detection of more true positive CMBs as shown in Figure 7A in the ablation study.

The main objective of the candidate detection step is to detect as many true positive candidates as possible, with a trade-off of high FPavg, because any of the CMB candidates missed in this stage

cannot be recovered in the subsequent steps. From the ablation study, given high FPavg in the candidate detection step, the student model trained using KD reduced FPavg approximately by a factor of 4 with a smaller decrease in cluster-wise TPR, when compared to the classification model trained without KD. The advantage of the teacher model in the proposed discrimination step was 2-fold: (1) its ability to learn the contextual features that are salient for both voxel-wise CMB detection and patch-level CMB/non-CMB classification and (2) the use of a multi-tasking framework with  $T_s$  providing high cluster-wise TPR, while  $T_c$  providing a patch-level regularization by penalizing false positive detection, reducing the chance of over-fitting and false classifications. We observed that the multi-tasking framework, together with the upweighting of the CMB classes in the loss function, reduced the effect of class imbalance between CMB and non-CMB patches (note that the model trained without KD is slightly biased toward the non-CMB class, evident from the lower cluster-wise TPR and FPavg values in Table 1). The  $T_c$  component of the teacher model classifies CMB from non-CMB mimics, which enhances the capability of the student networks to differentiate CMBs from non-CMBs, evident from better cluster-wise TPR and precision for the student model with KD. This is also shown in t-SNE plots in Figures 7C, D, where the classes are more separable for the KD case. Additionally, we also provided the input patches centered on the detected initial CMB candidates to the student model. This, in addition to the distilled knowledge from the teacher, enabled the model to focus on the pattern at the center of the patches for accurate classification of



CMB patches. This was especially useful to remove the fragments of blood vessels (e.g., intersections and branching points) missed in the vessel removal step as seen in [Figures 10D, 11D](#). Regarding the parameters used in KD, using a higher temperature ( $\tau$ ) results in softened softmax values between classes and has been shown to typically provide the knowledge (also known as *dark knowledge*) for training a generalizable student network ([Hinton et al., 2015](#)). However, given the similarities in the characteristics of CMBs and mimics, having very high  $\tau$  values could lead to misclassifications. Our main aim was to achieve a good hard prediction to differentiate the CMB class from the non-CMB class, while at the same time to transfer the knowledge from the teacher model to the student model. Hence we empirically chose an optimal  $\tau$  value of 4 (that provided smoother softmax values without affecting the CMB/non-CMB prediction) based on manual tuning. Further removal of FPs in the post-processing resulted in the drastic improvement in the cluster-wise precision. The use of a post-processing step based on shape and spatial criteria has been shown to be beneficial in other studies (see [Supplementary Table S2](#) for an overview). Given the big performance improvement after this step, we also tested the effect of applying post-processing at different stages of our proposed method. We found that, despite improving cluster-wise precision in all cases, the most beneficial effect was to use it as the final step of our 3-stage method (see [Supplementary Table S1](#)). Noise reduction or smoothing during pre-processing might lead to a loss of CMBs (even for data augmentation, very small  $\sigma$  values were chosen carefully). Therefore, small intensity and texture variations (mainly in the sub-cortical and lobar regions) led to the detection of FPs, which were removed in the post-processing step.

As for the cross-validation results within individual datasets, the method achieved the highest cluster-wise TPR values on SWI images (from the UKBB dataset), while providing the lowest FPavg and the highest cluster-wise precision on T2\*-GRE images (from the OXVASC dataset). This could be due to the fact that CMBs appear with a higher contrast on SWI compared to T2\*-GRE images due to the blooming effect. This also affects most of the CMB mimics as well, increasing their contrast on SWI, leading to high cluster-wise TPR but also high FPavg. Also, T2\*-GRE images had a smoother texture when compared to SWI ([Figure 10](#)), resulting in less noisy FRST maps, hence leading to the improved performance metrics at the candidate discrimination step in the OXVASC dataset. However, the FPavg value at the initial candidate detection step was higher for the OXVASC dataset due to the lower voxel resolution in the z-direction (5 mm), leading to partial volume artefacts and making it highly difficult to differentiate between small sulci closer to the skull and CMBs.

On evaluating the generalizability of our method across various datasets, our method trained on the SWI data from the UKBB dataset showed good generalizability on QSM images from the same dataset, with comparable performance to the cross-validation results on the SWI data. Regarding the performance after individual steps, in the initial candidate detection step, the method provided the highest cluster-wise TPR values with the lowest FPavg values on the QSM data (even lower than with UKBB SWI data) since QSM shows a better separation of diamagnetic mimics from CMBs. However, due to local tissue susceptibility variations (which is quite different from the SWI training data), the FPavg in the

candidate discrimination step was higher than it was when using the SWI data. Finally, the post-processing step effectively removed the stray voxels due to noisy susceptibility variations (that were extremely small and hence were below the 2.5 mm<sup>3</sup> threshold) and reduced the FPavg value to 1.8 FPs/subject. It is worth noting that, since the same subjects were used for training (SWI data for training and QSM data for testing), the results are likely to be biased. That is, the model could have learnt the overall locations of CMBs for the training subjects, rather than the modality-invariant features. However, we believe that the use of patches, rather than whole slices or volumes, at both steps would reduce the chance of biased assessment.

For the datasets consisting of the same modality as that of the training data (SWI) but from different populations, the method was affected by the presence of additional pathological signs (e.g., ICH in the TICH2 dataset). In the TICH2 dataset, the noisy texture of the hemorrhage regions and their edges led to the highest FPavg value in the initial candidate detection step. In terms of FPs, we found that additional pathological signs, that were not the part of training, affected the method more than the modalities. For instance, among the OXVASC (different modality from the training SWI data) and TICH2 datasets (same modality), even though both are pathological datasets, the greater prevalence of confounding “CMB-like” signs in TICH2 resulted in higher FPavg in the TICH2 dataset. Among all the datasets we used, the SHK dataset had high contrast, low noise and a better than average resolution making vessels and sulci easy to remove in this dataset. Moreover, this dataset has the same modality as that of the training data, and hence both candidate detection and discrimination step models performed well (and cluster-wise TPR values comparable even with that on the UKBB SWI data). However, during the post-processing step, a few CMBs near the sulci, closer to the skull were misclassified as FPs resulting in lower cluster-wise TPR. The OXVASC dataset was quite different from the training SWI data and from other datasets, since it shows lower contrast between CMB and background as shown in [Figure 11](#). Hence, providing FRST as the second input channel was particularly useful for this dataset, since the FRST relies more on the radial symmetry nature of CMBs at different radii (we used 2, 3, 4, and 6 as specified in [Section 2.2](#)) rather than its intensity contrast with respect to the background. Hence, on the OXVASC dataset the FRST maps had the same contrast as that of other modalities (as seen in [Figure 2A](#)) aiding in the detection of subtle CMBs. Since the estimation of FRST maps does not require any additional modality other than the input modality, our method effectively uses a single image modality and provided results comparable to existing methods that use multiple modalities ([Ghafaryasl et al., 2012](#); [Al-Masni et al., 2020](#)).

Concluding, we proposed a fully automated method using deep learning for CMB candidate detection, and candidate discrimination with a knowledge distillation framework, followed by post-processing filtering using structural and anatomical properties. Our method achieved cluster-wise TPR values of >90% with FPavg <1.6 FPs/subject on T2\*-GRE and SWI modalities, on par with the state-of-the-art, and gave better precision than existing methods. When the models were trained on SWI data and applied on QSM images from the same dataset, the method achieved a cluster-wise TPR ~90%. On applying the trained method to

other datasets consisting of data from different populations and acquired using different scanners and protocols, our method gave a cluster-wise TPR > 81%, despite the presence of other major pathologies. The Python implementation of the proposed method is currently available in <https://github.com/v-sundaresan/microbleed-detection>. The user guide (readme.md) provided via the above link provides more information regarding the scripts, input file formats and prediction times of the implementation. The tool yields high predictive performance on various modalities used in clinical settings. Also, given its short prediction time (<5 min/scan), it has the potential to be used to assist clinicians by reducing the time taken for assessing individual scans, which would also benefit patients in the long run. Also, the CMB segmentation maps obtained from the tool could be used for obtaining an automated rating of CMBs (i.e., total CMB count, spatial distribution and size). One of the future directions of this research would be to improve the generalizability of the proposed method using various domain adaptation techniques, to overcome the effect of scanner- and population-related variations. Another clinically focused avenue of this research could be to develop automated algorithms to rate the CMBs based on their size and distribution, which would be useful in studying their clinical impact.

## Data availability statement

The datasets presented in this article are not readily available because Requests for data from the OXVASC Study will be considered by PR in line with data protection laws. The TICH-2 MRI sub-study data can be shared with bona fide researchers and research groups on written request to the sub-study PI RD ([rob.dineen@nottingham.ac.uk](mailto:rob.dineen@nottingham.ac.uk)). Proposals will be assessed by the PI (with advice from the TICH-2 trial Steering Committee if required) and a Data Transfer Agreement will be established before any data are shared. The UK Biobank and Hong Kong (HK) datasets are available to researchers through an open applications via <https://www.ukbiobank.ac.uk/register-apply/> and <http://www.cse.cuhk.edu.hk/~qdou/cmb-3dcnn/cmb-3dcnn.html>, respectively. Requests to access the datasets should be directed to [rob.dineen@nottingham.ac.uk](mailto:rob.dineen@nottingham.ac.uk); [peter.rothwell@ndcn.ox.ac.uk](mailto:peter.rothwell@ndcn.ox.ac.uk).

## Ethics statement

OXVASC was approved by the South Central—Oxford A Research Ethics Committee (Research Ethics Committee reference number: 05/Q1604/70). Human subjects: UK Biobank has approval from the North West Multi-center Research Ethics Committee (MREC) to obtain and disseminate data and samples from the participants (<http://www.ukbiobank.ac.uk/ethics/>), and these ethical regulations cover the work in this study. Written informed consent was obtained from all participants. The TICH-2 trial obtained ethical approval from East Midlands (Nottingham 2) NHS Research Ethics Committee (Reference: 12/EM/0369) and the amendment to allow the TICH2 MRI sub-study was approved in April 2015 (amendment number SA02/15). The

patients/participants provided their written informed consent to participate in this study.

## Author contributions

VS contributed to conceptualization, methodology, software, validation, formal analysis, investigation, visualization, and wrote the original draft. CA performed investigation, methodology, and software. GZ, AM, RD, PR, DA, and NS contributed to resources. AM and RD also contributed to data curation. DA contributed to data curation and investigation. CW, KM, BT, FA-A, and SS contributed to resources and data curation. LG contributed to conceptualization, methodology, data curation, and project administration. MJ contributed to conceptualization, methodology, funding acquisition, and project administration. GZ, MJ, and LG contributed to supervision. All authors contributed to manuscript revision, read, and approved the submitted version.

## Funding

This research was funded in part by the Wellcome Trust (203139/Z/16/Z). For the purpose of open access, the author has applied a CC-BY public copyright licence to any Author Accepted manuscript version arising from this submission. This work was also supported by the Engineering and Physical Sciences Research Council (EPSRC), Medical Research Council (MRC) (grant number EP/L016052/1), NIHR Nottingham Biomedical Research Centre and Wellcome Centre for Integrative Neuroimaging, which has core funding from the Wellcome Trust (203139/Z/16/Z). The computational aspects of this research were funded from National Institute for Health Research (NIHR) Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The Oxford Vascular Study was funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), Wellcome Trust, Wolfson Foundation, the British Heart Foundation and the European Union's Horizon 2020 programme (grant 666881, SVDs@target). The TICH-2 MRI sub-study was funded by a grant from British Heart Foundation (grant reference PG/14/96/31262) and the TICH-2 trial was funded by a grant from the NIHR Health Technology Assessment programme (project code 11\_129\_109). VS was supported by the Wellcome Centre for Integrative Neuroimaging (203139/Z/16/Z). CA was supported by NIHR Nottingham Biomedical Research Centre and is now supported by Wellcome Trust Collaborative Award (215573/Z/19/Z). GZ is supported by the Italian Ministry of Education (MIUR) and by a grant Dipartimenti di eccellenza 2018-2022, MIUR, Italy, to the Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia. PR is in receipt of a NIHR Senior Investigator award. KM and BT are funded by a Senior Research Fellowship from the Wellcome Trust (202788/Z/16/Z). MJ was supported by the NIHR Oxford Biomedical Research Centre (BRC), and this research was funded by the Wellcome Trust (215573/Z/19/Z). LG was supported by an Alzheimer's Association Grant (AARF-21-846366) and by the National Institute for Health

and Care Research (NIHR) Oxford Health Biomedical Research Centre (BRC).

## Acknowledgments

The UKBB data used in this work was obtained from UK Biobank under Data Access Applications (8107, 43822). We are grateful to the UK Biobank for making the resource data available, and are extremely grateful to all UK Biobank study participants, who generously donated their time to make this resource possible. We acknowledge all the OXVASC study participants. For the OXVASC dataset, we acknowledge the use of the facilities of the Acute Vascular Imaging Centre, Oxford. We also thank Dr. Chiara Vincenzi and Dr. Francesco Carletti for their help on generating the manual masks used in our experiments.

## Conflict of interest

MJ and LG receive royalties from licensing of FSL to non-academic, commercial parties.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Al-Masni, M. A., Kim, W.-R., Kim, E. Y., Noh, Y., and Kim, D.-H. (2020). "A two cascaded network integrating regional-based yolo and 3d-cnn for cerebral microbleeds detection," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Montreal, QL: IEEE), 1055–1058.
- Amiri, M., Brooks, R., and Rivaz, H. (2019). "Fine tuning u-net for ultrasound image segmentation: which layers?," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3D 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings I* (Shenzhen: Springer), 235–242.
- Ba, L. J., and Caruana, R. (2013). Do deep nets really need to be deep? *arXiv*. 27.
- Barnes, S. R., Haacke, E. M., Ayaz, M., Boikov, A. S., Kirsch, W., and Kido, D. (2011). Semiautomated detection of cerebral microbleeds in magnetic resonance images. *Magn. Reson. Imaging* 29, 844–852. doi: 10.1016/j.mri.2011.02.028
- Bian, W., Hess, C. P., Chang, S. M., Nelson, S. J., and Lupo, J. M. (2013). Computer-aided detection of radiation-induced cerebral microbleeds on susceptibility-weighted MR images. *NeuroImage Clin.* 2, 282–290. doi: 10.1016/j.nicl.2013.01.012
- Bucilucă, C., Caruana, R., and Niculescu-Mizil, A. (2006). "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and Data Mining* (Philadelphia, PA), 535–541.
- Charidimou, A., and Werring, D. J. (2011). Cerebral microbleeds: detection, mechanisms and clinical challenges. *Fut. Neurol.* 6, 587–611. doi: 10.2217/fnl.11.42
- Chen, G., Choi, W., Yu, X., Han, T., and Chandraker, M. (2017). "Learning efficient object detection models with knowledge distillation," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (New York, NY), 742–751.
- Chen, H., Yu, L., Dou, Q., Shi, L., Mok, V. C., and Heng, P. A. (2015). "Automatic detection of cerebral microbleeds via deep learning based 3D feature representation," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* (New York, NY: IEEE), 764–767.
- Chen, Y., Villanueva-Meyer, J. E., Morrison, M. A., and Lupo, J. M. (2018). Toward automatic detection of radiation-induced cerebral microbleeds using a 3d deep residual network. *J. Dig. Imaging* 32, 766–772. doi: 10.1007/s10278-018-0146-z
- Chesebro, A. G., Amarante, E., Lao, P. J., Meier, I. B., Mayeux, R., and Brickman, A. M. (2021). Automated detection of cerebral microbleeds on t2\*-weighted mri. *Sci. Rep.* 11, 1–13. doi: 10.1038/s41598-021-83607-0
- Cordonnier, C., Wardlaw, J., and Al-Shahi Salman, R. (2007). Spontaneous brain microbleeds: systematic review, subgroup analyses and standards for study design and reporting. *Brain* 130, 1988–2003. doi: 10.1093/brain/awl387
- De Bresser, J., Brundel, M., Conijn, M., Van Dillen, J., Geerlings, M., Viergever, M., et al. (2013). Visual cerebral microbleed detection on 7T MR imaging: reliability and effects of image processing. *Am. J. Neuroradiol.* 34, E61–E64. doi: 10.3174/ajnr.A2960
- Dineen, R. A., Pszczolkowski, S., Flaherty, K., Law, Z. K., Morgan, P. S., Roberts, I., et al. (2018). Does tranexamic acid lead to changes in MRI measures of brain tissue health in patients with spontaneous intracerebral haemorrhage? protocol for a MRI substudy nested within the double-blind randomised controlled TICH-2 trial. *BMJ Open* 8, e019930. doi: 10.1136/bmjopen-2017-019930
- Ding, Q., Wu, S., Sun, H., Guo, J., and Xia, S.-T. (2019). Adaptive regularization of labels. *arXiv*.
- Dou, Q., Chen, H., Yu, L., Shi, L., Wang, D., Mok, V. C., et al. (2015). "Automatic cerebral microbleeds detection from MR images via independent subspace analysis based hierarchical features," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan: IEEE), 7933–7936.
- Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., et al. (2016). Automatic detection of cerebral microbleeds from MR images via 3d convolutional neural networks. *IEEE Trans. Med. Imaging* 35, 1182–1195. doi: 10.1109/TMI.2016.2528129
- Du, G., Cao, X., Liang, J., Chen, X., and Zhan, Y. (2020). Medical image segmentation based on u-net: a review. *J. Imaging Sci. Technol.* doi: 10.2352/J.ImagingSci.Technol.2020.64.2.020508
- Fazlollahi, A., Meriaudeau, F., Giancardo, L., Villemagne, V. L., Rowe, C. C., Yates, P., et al. (2015). Computer-aided detection of cerebral microbleeds in susceptibility-weighted imaging. *Comp. Med. Imaging Graph.* 46, 269–276. doi: 10.1016/j.compmedimag.2015.10.001
- Fazlollahi, A., Meriaudeau, F., Villemagne, V. L., Rowe, C. C., Yates, P., Salvado, O., et al. (2014). "Efficient machine learning framework for computer-aided detection of cerebral microbleeds using the radon transform," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (Beijing: IEEE), 113–116.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Author disclaimer

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fninf.2023.1204186/full#supplementary-material>

- Förstner, W. (1994). "A framework for low level feature extraction," in *European Conference on Computer Vision* (Stockholm: Springer), 383–394.
- Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). "Multiscale vessel enhancement filtering," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Boston, MA: Springer), 130–137.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030. doi: 10.1007/978-3-319-58347-1\_10
- Ghafariyasl, B., van der Lijn, F., Poels, M., Vrooman, H., Ikram, M. A., Niessen, W. J., et al. (2012). "A computer aided detection system for cerebral microbleeds in brain MRI," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)* (Barcelona: IEEE), 138–141.
- Greenberg, S. M., Vernooij, M. W., Cordonnier, C., Viswanathan, A., Salman, R. A.-S., Warach, S., et al. (2009). Cerebral microbleeds: a guide to detection and interpretation. *Lancet Neurol.* 8, 165–174. doi: 10.1016/S1474-4422(09)70013-4
- Gregoire, S., Chaudhary, U., Brown, M., Yousry, T., Kallis, C., Jäger, H., et al. (2009). The Microbleed Anatomical Rating Scale (MARS): reliability of a tool to map brain microbleeds. *Neurology* 73, 1759–1766. doi: 10.1212/WNL.0b013e3181c34a7d
- Guo, Q., Wang, X., Wu, Y., Yu, Z., Liang, D., Hu, X., et al. (2020). "Online knowledge distillation via collaborative learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 11020–11029.
- Haacke, E. M., Xu, Y., Cheng, Y.-C. N., and Reichenbach, J. R. (2004). Susceptibility weighted imaging (swi). *Magn. Reson. Med.* 52, 612–618. doi: 10.1002/mrm.20198
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778.
- Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*.
- Hong, J., Wang, S.-H., Cheng, H., and Liu, J. (2020). Classification of cerebral microbleeds based on fully-optimized convolutional neural network. *Multimed. Tools Appl.* 79, 15151–15169. doi: 10.1007/s11042-018-6862-z
- Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., et al. (2020). "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Lima: Springer), 772–781.
- Jin, X., Peng, B., Wu, Y., Liu, Y., Liu, J., Liang, D., et al. (2019). "Knowledge distillation via route constrained optimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Long beach, CA), 1345–1354.
- Kim, S. W., and Kim, H.-E. (2017). *Transferring Knowledge to Smaller Network With Class-Distance Loss*.
- Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv*. 1412, p.80.
- Kuijff, H. J., Brundel, M., de Bresser, J., van Veluw, S. J., Heringa, S. M., Viergever, M. A., et al. (2013). Semi-automated detection of cerebral microbleeds on 3.0 T MR images. *PLoS ONE* 8, e66610. doi: 10.1371/journal.pone.0066610
- Kuijff, H. J., de Bresser, J., Geerlings, M. I., Conijn, M. M., Viergever, M. A., Biessels, G. J., et al. (2012). Efficient detection of cerebral microbleeds on 7.0 T MR images using the radial symmetry transform. *Neuroimage* 59, 2266–2273. doi: 10.1016/j.neuroimage.2011.09.061
- Lachinov, D., Shipunova, E., and Turlapov, V. (2019). "Knowledge distillation for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen: Springer), 324–332.
- Lan, X., Zhu, X., and Gong, S. (2018). Knowledge distillation by on-the-fly native ensemble. *arXiv*. 31
- Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., and Li, L.-J. (2017). "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 1910–1918.
- Liu, C., Li, W., Tong, K. A., Yeom, K. W., and Kuzminski, S. (2015). Susceptibility-weighted imaging and quantitative susceptibility mapping in the brain. *J. Magn. Reson. Imaging* 42, 23–41. doi: 10.1002/jmri.24768
- Liu, L., Wang, H., Lin, J., Socher, R., and Xiong, C. (2019a). Mkd: a multi-task knowledge distillation approach for pretrained language models. *arXiv*.
- Liu, S., Utriainen, D., Chai, C., Chen, Y., Wang, L., Sethi, S. K., et al. (2019b). Cerebral microbleed detection using susceptibility weighted imaging and deep learning. *Neuroimage* 198, 271–282. doi: 10.1016/j.neuroimage.2019.05.046
- Loy, G., and Zelinsky, A. (2002). "A fast radial symmetry transform for detecting points of interest," in *European Conference on Computer Vision* (Springer), 358–368.
- Lu, D., Liu, J., MacKinnon, A. D., Tozer, D. J., and Markus, H. S. (2021a). Prevalence and risk factors of cerebral microbleeds: analysis from the uk biobank. *Neurology* 97, e1493–e1502. doi: 10.1212/WNL.00000000000012673
- Lu, S.-Y., Nayak, D. R., Wang, S.-H., and Zhang, Y.-D. (2021b). A cerebral microbleed diagnosis method via featurenet and ensemble randomized neural networks. *Appl. Soft Comp.* 109, 107567. doi: 10.1016/j.asoc.2021.107567
- Morrison, M. A., Payabvash, S., Chen, Y., Avadiappan, S., Shah, M., Zou, X., et al. (2018). 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. doi: 10.1016/j.nicl.2018.08.002
- Müller, R., Kornblith, S., and Hinton, G. (2019). When does label smoothing help? *arXiv*.
- Nandigam, R., Viswanathan, A., Delgado, P., Skehan, M., Smith, E., Rosand, J., et al. (2009). MR imaging detection of cerebral microbleeds: effect of susceptibility-weighted imaging, section thickness, and field strength. *Am. J. Neuroradiol.* 30, 338–343. doi: 10.3174/ajnr.A1355
- Pan, Z.-W., Xiang, D.-H., Xiao, Q.-W., and Zhou, D.-X. (2008). Parzen windows for multi-class classification. *J. Complex.* 24, 606–618. doi: 10.1016/j.jco.2008.07.001
- Pszczolkowski, S., Sprigg, N., Woodhouse, L. J., Gallagher, R., Swinton, D., Law, Z. K., et al. (2022). Effect of tranexamic acid administration on remote cerebral ischemic lesions in acute spontaneous intracerebral hemorrhage: a substudy of a randomized clinical trial. *JAMA Neurol.* 79, 468–477. doi: 10.1001/jamaneurol.2022.0217
- Rashid, T., Abdulkadir, A., Nasrallah, I. M., Ware, J. B., Liu, H., Spincemille, P., et al. (2021). Deepmir: a deep neural network for differential detection of cerebral microbleeds and iron deposits in mri. *Sci. Rep.* 11, 1–14. doi: 10.1038/s41598-021-93427-x
- Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 7263–7271.
- Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). Fitnets: Hints for thin deep nets. *arXiv*.
- Rothwell, P., Coull, A., Giles, M., Howard, S., Silver, L., Bull, L., et al. (2004). Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study). *Lancet* 363, 1925–1933. doi: 10.1016/S0140-6736(04)16405-2
- Sarfraz, F., Arani, E., and Zonooz, B. (2019). *Noisy Collaboration in Knowledge Distillation*.
- Seghier, M. L., Kolanko, M. A., Leff, A. P., Jäger, H. R., Gregoire, S. M., and Werring, D. J. (2011). Microbleed detection using automated segmentation (MIDAS): a new method applicable to standard clinical MR images. *PLoS ONE* 6, e17547. doi: 10.1371/journal.pone.0017547
- Shams, S., Martola, J., Cavallin, L., Granberg, T., Shams, M., Aspelin, P., et al. (2015). SWI or T2\*: which MRI sequence to use in the detection of cerebral microbleeds? The Karolinska Imaging Dementia Study. *Am. J. Neuroradiol.* 36, 1089–1095. doi: 10.3174/ajnr.A4248
- Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062
- Sprigg, N., Flaherty, K., Appleton, J. P., Salman, R. A.-S., Bereczki, D., Beridze, M., et al. (2018). Tranexamic acid for hyperacute primary intracerebral haemorrhage (TICH-2): an international randomised, placebo-controlled, phase 3 superiority trial. *Lancet* 391, 2107–2115. doi: 10.1016/S0140-6736(18)31033-X
- Sundaresan, V., Arthofer, C., Zamboni, G., Dineen, R. A., Rothwell, P. M., Sotiropoulos, S. N., et al. (2022). Automated detection of candidate subjects with cerebral microbleeds using machine learning. *Front. Neuroinform.* 15, 777828. doi: 10.3389/fninf.2021.777828
- Vadacchino, S., Mehta, R., Sepahvand, N. M., Nichyporuk, B., Clark, J. J., and Arbel, T. (2021). Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. *Medical Imaging with Deep Learning. arXiv*. 787–801.
- van den Heuvel, T., Van Der Eerden, A., Manniesing, R., Ghafoorian, M., Tan, T., Andriessen, T., et al. (2016). Automated detection of cerebral microbleeds in patients with traumatic brain injury. *NeuroImage Clin.* 12, 241–251. doi: 10.1016/j.nicl.2016.07.002
- Wang, C., Martins-Bach, A. B., Alfaro-Almagro, F., Douaud, G., Klein, J. C., Llera, A., et al. (2021). Phenotypic and genetic associations of quantitative magnetic susceptibility in uk biobank brain imaging. *bioRxiv*. 25, 818–831. doi: 10.1101/2021.06.28.450248
- Wang, S., Tang, C., Sun, J., and Zhang, Y. (2019). Cerebral micro-bleeding detection based on densely connected neural network. *Front. Neurosci.* 13, 422. doi: 10.3389/fnins.2019.00422
- Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 10687–10698.
- Yang, C., Xie, L., Su, C., and Yuille, A. L. (2019). "Snapshot distillation: teacher-student optimization in one generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long beach, CA), 2859–2868.
- Yang, X., Zeng, Z., Yeo, S. Y., Tan, C., Tey, H. L., and Su, Y. (2017). A novel multi-task deep learning model for skin lesion segmentation and classification. *arXiv*.
- Ye, J., Wang, X., Ji, Y., Ou, K., and Song, M. (2019). Amalgamating filtered knowledge: learning task-customized student from multi-task teachers. *arXiv*. Macao. 4128–34. doi: 10.24963/ijcai.2019/573



Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., and Ma, K. (2019). "Be your own teacher: improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 3713–3722.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

Zhang, Y.-D., Hou, X.-X., Chen, Y., Chen, H., Yang, M., Yang, J., et al. (2018). Voxelwise detection of cerebral microbleed in cadasil patients by leaky rectified linear unit and early stopping.

*Multimed. Tools Appl.* 77, 21825–21845. doi: 10.1007/s11042-017-4383-9

Zhang, Y.-D., Hou, X.-X., Lv, Y.-D., Chen, H., Zhang, Y., and Wang, S.-H. (2016). "Sparse autoencoder based deep neural network for voxelwise detection of cerebral microbleed," in *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)* (Seoul: IEEE), 1229–1232.

Zhou, G., Fan, Y., Cui, R., Bian, W., Zhu, X., and Gai, K. (2018). "Rocket launching: a universal and efficient framework for training well-performing light net," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32 New Orleans, LA.