IT UNIVERSITY OF COPENHAGEN

Automatic Segmentation of Cerebral Microbleeds

Jorge del Pozo Lérida

Research Project
KIREPRO1PE
MSc in Data Science
December 15 2023

Jorge del Pozo Lérida (jord@itu.dk)

ITU supervisor:
Veronika Cheplygina (vech@itu.dk)

CEREBRIU supervisors:
Mathias Perslev (mp@cerebriu.com),
Akshay Pai (ap@cerebriu.com),
Silvia Ingala (si@cerebriu.com)

Automatic Segmentation of Cerebral Microbleeds

Jorge del Pozo Lérida
jorgedelpozolerida@gmail.com // jord@itu.dk

***Abstract*—Cerebral Microbleeds (CMBs) are crucial neuroimaging biomarkers associated with medical conditions such as stroke, intracranial hemorrhage, and cerebral small vessel disease. They are detectable as hypointensities on magnetic resonance images (MRI) in T2*-weighted or susceptibility-weighted sequences. Identifying CMBs is a time-consuming and error-prone task for radiologists, making the need for automatic detection critical. Yet, it remains a challenging endeavor due to the small size and quantity of CMBs, scarcity of publicly available annotated data, and their resemblance to various other mimics among other things. This complexity hinders the development of a clinically integrated automated solution. In response to these challenges, this study carefully reviewed the literature on this topic and tested a commonly used architecture, U-Net, for the segmentation and detection of CMBs using the public VALDO dataset. Adhering to the latest research guidelines, the study achieved a recall of 0.71, a precision of 0.44, and an F1 score of 0.54, with an average of 1.5 and 0.9 false positives per subject and per CMB respectively. Concurrently, a new clinically relevant dataset specifically tailored for CMB segmentation was developed, to be utilized in future work.**

***Index Terms*—Medical Imaging, Image Segmentation, Cerebral Microbleeds, Machine Learning, Deep Learning**

I. Introduction

Cerebral Microbleeds (CMBs) or cerebral microhemorrhages are small lesions that result from the accumulation of hemosiderin breakdown products in the brain parenchyma due to previous microscopic hemorrhage [1]. They are detectable as hypointensities on magnetic resonance images (MRI) at T2*-weighted or susceptibility-weighted sequences [2]. CMBs are closely associated with major health conditions such as stroke, Alzheimer's disease, and Diabetes mellitus, all of which are among the top ten causes of death worldwide [3]. Detecting CMBs is crucial for diagnosing these and other conditions and influences treatment decisions. For example, in stroke patients, CMBs indicate an increased risk of hemorrhage following thrombolysis treatment or therapeutic anticoagulation [4]. Moreover, CMBs are often found in association with many other pathologies, for instance indicating several pathological processes in the cerebral vessels.

Nevertheless, identifying CMBs is a challenging and time-consuming task for radiologists, with high variability and error rates, especially when numerous CMBs present [4], [5]. This situation underscores the need for automated solutions, which during the last decade have seen increasing research interest. However, the field is marked by a high degree of complexity, inconsistency in approaches and limited availability of datasets. Initiatives like the VALDO challenge [6] have emerged to accelerate advancements in automated solutions for CMB detection, emulating successful dynamics from public challenges like BRATS [7]. Additionally, some researchers have recently tried to establish guidelines for automated CMB detection [8]. Despite these efforts, a clinically integrated automated solution for CMB detection is yet to be realized, highlighting the ongoing challenges and the necessity for continued research in this area

In collaboration with CEREBRIU — a software company specializing in deep learning applications for MRI data — this project intends to advance in bridging the gap between academic research and commercial application in CMB detection. While CEREBRIU's current technology encompasses segmentation of several types of tumors, infarcts, and hemorrhages, it does not yet include CMBs, so understanding how to correctly succesfully detect these lesions becomes relevant. A key aspect of this collaboration is the creation of a new dataset using the company's internal datasets to address the scarcity of clinically relevant datasets for CMB segmentation.

To effectively address the task of detecting CMBs, we aim to thoroughly understand all facets of CMBs and their automated segmentation. This involves a detailed review of the relevant literature to grasp both the medical and technical aspects of CMB detection, including the challenges encountered and the inherent features of the task. Additionally, we experiment with a 3D U-Net deep learning architecture, adhering to the latest recommendations in literature and using the public dataset from VALDO challenge. This not only provides us with direct experience but also yields valuable insights into the complexities and nuances of the task with a combination of architecture and dataset that can be easily reproduced and compared. The project's main contributions can be summarized into:

1) Generating a new, clinically relevant dataset for CMB segmentation to be used in future work.
2) Testing a U-Net architecture for CMB segmentation on the VALDO dataset, following latest literature guidance.

Our methodology achieved a recall of 0.71, a precision of 0.44, and an F1 score of 0.54, with an average of 1.5 false positives per subject and a rate of 0.9 FPs for each actual CMB. These results suggest that traditional deep learning architectures, when appropriately optimized based on current CMBs research insights, can be capable of effectively detecting CMBs, which lays the groundwork for future research employing more sophisticated models and methods, alongside more clinically diverse and relevant datasets. The latter is tackled through the development of a new dataset for future CMB segmentation research, encompassing various pathologies, a wide range of acquisition parameters and demographics, while including metadata essential for the CMB detection task.

II. Background

A. Magnetic Resonance Imaging (MRI)

MRI is a medical imaging modality that provides detailed images of the body's internal structures. The underlying physics are highly complex, but in a nutshell, MRI operates using two different energy sources: strong magnetic fields and radiofrequency (RF) waves. The technique exploits the natural magnetic properties of some special nuclei present in the body (most of the times hydrogen, found abundantly due to water molecules). When subjected to this magnetic field, these nuclei resonate at a particular frequency, which is then perturbed by RF pulses. The MRI machine's instrumentation, consisting of a main magnet, gradient coils (for spatial encoding), and an RF system, detects the signals emitted as these nuclei return to their baseline states during the signal reading. Data acquisition involves sampling these emitted signals in the frequency domain, denoted 'k-space', which is then converted into the anatomical images using inverse Fourier transform.

The contrast observed in an MRI image is determined by the weighting assigned to three key elements: proton density, T1 or longitudinal relaxation time, and T2 or transverse relaxation time. Through various combinations of gradients, echo times (TE), repetition times (TR), pulses, echoes, and other parameters, a range of distinct sequence types can be produced. Each of these sequences yields different contrasts, effectively highlighting various tissue types based on their unique characteristics in the imaging process. There are over a hundred sequence types that can have varying acronyms based on manufacturer.

Gradient Recalled Echo (GRE) sequences are a fundamental type of MRI sequence that are characterized by their use of lower flip angles, usually below 90 degrees (which is the go-to in Spin Echo sequences), and the absence of a 180-degree radiofrequency rephasing pulse. This setup allows for faster image acquisition and highlights tissue properties related to magnetic susceptibility. Within this family lie the two most common sequences used to detect CMBs: Gradient-echo T2*-weighted Imaging (T2S for short from now on) and susceptibility-weighted imaging (SWI). In Figure II-B, II-B one can see how CMBs are seen in these two sequence tpyes.

T2S sequence is created using GRE pulse sequences, where the selected echo time plays a crucial role in determining the contrast level. In these images, cerebral microbleeds (CMBs) are typically seen as areas of signal loss in the magnitude image [9], due to changes in local magnetic susceptibility that reflect the pathologic iron accumulation. Although T2S has been a longstanding standard for CMB detection [10], SWI has demonstrated greater reliability and sensitivity in identifying CMBs [11], [12]. However, this comes with a trade-off of increased visibility of other structures that mimic CMBs. What is more, there is an increased blooming effect on SWI, which makes CMB look more irregular in shape [13]

SWI employs a mix of advanced acquisition and post-processing techniques to enhance susceptibility contrast, typically using three-dimensional GRE imaging with extended echo times to improve the detection of hemosiderin deposits. While other sequences, such as Phase imaging for distinguishing mimics or Quantitative Susceptibility Mapping (QSM), can be employed for CMB detection, T2S and SWI stand out for their routine use both in automated detection systems and in daily clinical practice.

B. Cerebral Microbleeds

CMBs are typically identified as small areas of signal void, often accompanied by a blooming artifact on MRI image, excluding larger haematomas (macrobleeds), specific secondary causes of bleeding, and non-haemorrhagic causes of signal void [10]. The observed blooming effect of CMBs is protocol dependent (resolution, signal-to-noise, echo time, field strength and susceptibility) [14]. CMBs can also be seen in computed tomography (CT) images, where sensitivity is the highest within the first few days. However, they fade from the image in about 7 to 10 days, which makes MRI the preferred technique for detecting CMBs, since they remain way longer. Cerebral microbleeds (CMBs) are observed in approximately 7–13% of the general population aged between 40 and 69 years, with this prevalence increasing in individuals over 80 years of age [15]. Among those with CMBs, about 70–80% have a single microbleed, corresponding to a prevalence rate of 35.7%. Notably, the occurrence of CMBs is more prevalent in stroke patients, with a reported prevalence of 29.4% [16]. As explained before, CMBs are better detected on SWI than on T2S, disclosing up to 6 times more CMBs [13]. However, both sequences present FNs, and CMBs are normally detected with a true positive rate of 48%–89% across several diseases and even optimized MR imaging sequences at 1.5 and 3.0 tesla (T) detect only about 50% [2].

Visual rating scales have demonstrated effectiveness in enhancing the consistency of evaluations among different raters, particularly in determining the presence and anatomical positioning of CMBs. Two of the most prominent and validated scales are the Microbleed Anatomic Rating Scale (MARS) [17] and the Brain Observer MicroBleed Scale (BOMBS) [18]. Both these scales provide a structured approach to classifying the location of CMBs into three key areas: the lobar, deep, and infratentorial regions, including the brainstem. The design of these scales takes into account the fact that diagnostic information is contained in the quantity and location of CMBs. For instance, a lobar distribution of CMBs is commonly seen in cerebral amyloid angiopathy (CAA), whereas a more widespread distribution, including deep or central areas, is indicative of hypertensive CMBs. Additionally, a higher count of CMBs correlates with an increased risk of cognitive decline, dementia, and stroke [2].

CMBs have several **mimics**, including calcium and iron deposits, flow voids from pial blood vessels, paramagnetic deoxyhemoglobin in cerebral venules, partial volume artifacts from bone, cavernous malformations, metastatic melanoma, and diffuse axonal injury [10]. Calcium and iron deposits often appear as small foci of low signal intensity on T2S. Flow voids, visible in cross-sectional views of cortical sulci,

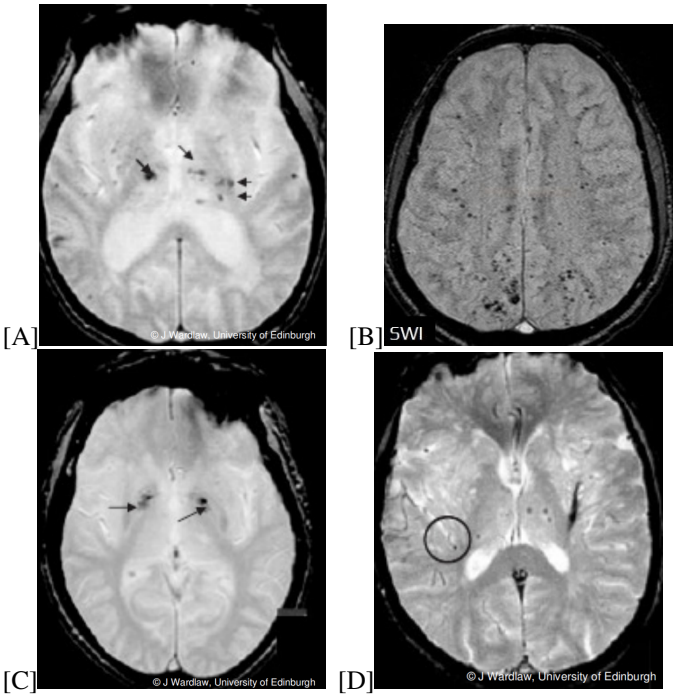

Fig. 1. Cerebral Microbleeds and some of its common mimics. A) CMB on T2S; B) CMB on SIW; C) Basal ganglia calcification (mimic); D) Cortical Vessels (mimic) on T2S. All pictures taken from [18], except for B which is taken from https://www.stroke-manual.com/

differ from CMBs in their linear structure and consistent appearance across different MRI sequences. Paramagnetic deoxyhemoglobin in cerebral venules also creates a blooming effect, but unlike CMBs, these venules have a distinct tubular structure. Partial volume artifacts, particularly from bone near sinuses, can obscure or be confused with CMBs, especially in areas close to the temporal and frontal lobes. Cavernous malformations, sometimes considered secondary causes of CMB-like appearances, are identifiable through stagnant blood in their sinusoidal lumen and a characteristic hemosiderin rim on MRI sequences. Metastatic melanoma manifests as hypointense areas on T2S due to melanin and bleeding, distinguishable from CMBs by additional T1 hyperintensity or surrounding edema. Lastly, diffuse axonal injury (DAI) following head trauma presents as a secondary cause of CMB-like lesions, discernible through clinical history and concurrent imaging abnormalities. All these False-positive cases or mimics occur in about 11% to 24% of instances [2]. Figure II-B, II-B show some of these mimics on T2S.

CMBs findings on MRI data have been **associated with more than 30 conditions** [19]. In the context of intracerebral hemorrhage (ICH), CMBs serve as a strong independent predictor for the first onset of ICH [19]. CMBs are crucial for strokes. Their presence, can predict hemorrhagic and ischemic stroke, even in healthy persons and the may be helpful in distinguishing between stroke (deep brain) and degenerative diseases (lobar) [20], [21]. In patients with atrial fibrillation (AF) on antithrombotic therapy, CMBs, especially more than five, increase the risk of ICH and its associated mortality [22]. CMBs also serve as markers for cerebral amyloid angiopathy (CAA), where they are typically found peripherally as opposed to being located in the basal ganglia or infratentorial regions as seen in hypertensive arteriopathy [23]. Again, this distinction is crucial in assessing hemorrhage risk after treatments involving antiplatelet, antithrombotic, or thrombolytic therapies. In traumatic brain injury (TBI), the number and severity of CMBs correlate with the injury's severity and can increase a week after the incident [24]. They are also a concern in radiation therapy, with their incidence increasing post-treatment and linked to higher radiation doses [25]. Alzheimer's disease (AD) and other cognitive dysfunctions exhibit a correlation between the number of CMBs and declining cognitive function, particularly marked when more than ten CMBs are present [26]. Furthermore, CMBs are associated with worse outcomes in conditions like coronary artery disease (CAD), infective endocarditis (IE), chronic obstructive pulmonary disease (COPD), and chronic renal disease, among others [19].

C. Deep Learning in Medical Imaging

In the last decades the use of Machine Learning (ML) and more specifically Deep Learning (DL) have seen extensive adoption into several fields of society [27]. This is no less for Medical Image Analysis (MIA), where the use of Deep Learning algorithms, in particular Convolutional Neural Networks (CNNs), has permeated the entire field and has rapidly become a methodology of choice for analyzing medical images, having proved to be the state-of-the-art foundation for some time already [27]. DL applications in MIA are diverse, covering tasks such as classification (e.g. image/exam classifications or identifying detected objects), detection (e.g. localizing organs, regions, landmarks, objects, or lesions), and segmentation (segmenting organs, substructures, and lesions). Additionally, DL contributes to tasks like registration, content-based image retrieval, image generation and enhancement, and the combination of image data with reports, among others [28]–[30].

DL methods excel with large datasets for training. However, available and relevant data from medical institutions is often limited. Despite the vast amounts of medical data held by healthcare institutions, accessing it for research — and especially for commercial applications — remains a significant hurdle, and often the available data is either inaccurately labeled or not pertinent due to variations in patient cohorts, demographics, pathologies, or technical factors like differing imaging modalities and sequences. This limitation hinders the development of highly accurate models without the risk of overfitting [28]. Different ways in which community has tried to make the most of these imperfect datasets for medical image segmentation are outlined by Tajbakhsh et al. [31]. For these reasons and due to compelxity of medical domain, in MIA a key factors is still professional knowledge in the task at hand, which can offer significant benefits for the model, extending beyond merely increasing the complexity of a CNN's layers

[32]. But these difficulties do not seem to stop DL models from developing in the medical imaging, where significant progress in their research and development is expected [29].

D. Medical Image Segmentation

Image segmentation is the process of partitioning an image into distinct regions. In the context of medical imaging, segmentation becomes crucial as it enables the identification of specific anatomical structures or regions of interest within the image that can help clinicians with decision-making and treatment or surgery planning. Methods for segmentation are diverse and can be grouped into six categories: a) Thresholding, b) Region growing, c) Region merging-splitting, d) Clustering, e) Edge detection and f) Model-based methods [33]. While some methods use the resemblance among pixels to create a segment (similarity approach), others isolate segments based on their differences (discontinuity approach) [34]. There are two subdivision of image segmentation task: a) Semantic segmentation, which classifies each pixel in an image into a category without differentiating between individual objects of the same class and b) Instance segmentation, that not only categorizes each pixel but also distinguishes between different instances of the same category.

Also in medical image segmentation, the same holds and state-of-the-art methods predominantly utilize Deep Learning (DL) approaches, significantly outperforming previous techniques [30]. A pivotal development in this field was the introduction of the **U-Net** architecture [35], which then became the backbone for many subsequent segmentation models. This architecture begins with convolutional layers that extract initial low-level features from the input image. Following this, there is a process of downsampling, commonly using max pooling, to decrease the spatial dimensions of the feature maps. Concurrently, the depth of feature maps is often increased in the subsequent convolutional layer. This design aims to transform the large spatial feature maps into more compact but deeper representations, preserving essential information while reducing spatial dimensions. The deeper layers of this 'contracting' path are intended to capture several high-level, potentially image-wide features (e.g. the presence of a tumor). In U-Net architecture, each upsampling step incorporates 'skip connections' by concatenating feature maps from the downsampling phase. These connections are crucial for recovering detailed spatial information lost during max pooling. This process ensures that the network retains fine details necessary for precise segmentation. The architecture culminates in a 1x1 convolution layer that maps deep feature representations to segmentation classes. U-Net's balanced approach, efficiently combining global context and local details, is particularly effective for medical image segmentation tasks. Figure 2 depicts the architecture just explained.

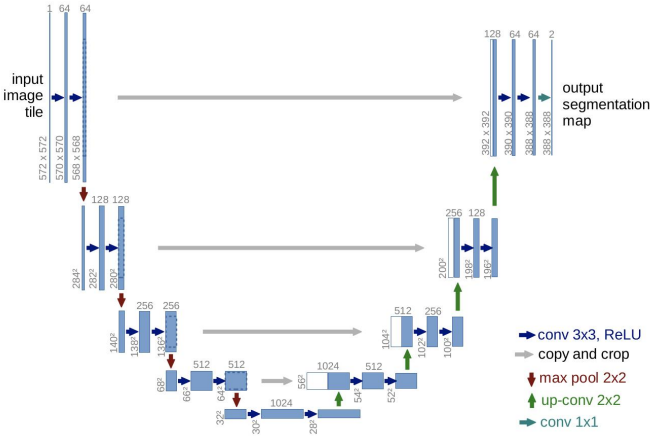

Fig. 2. 2D U-net architecture (example for 32x32 pixels in the lowest resolution). Taken from original U-Net paper [35].

III. Related Work

A. Summary of reviews

To the best of our knowledge, there is only one comprehensive review article that explores solely the approaches taken in the past to automatically detect CMBs done by Ferlin et al. [8], which will be published in December 2023 and also attempts to establish guidelines on how to approach the task. This study identified 67 studies containing some sort of automatic CMB detection (both using traditional methods and ML) until early 2023. The closest attempt is made by Jiang et al. [36], who also perform an somewhat extensive review of automated methods for CMB detection — along with methods for other imaging biomarkers of Cerebral Small Vessel disease (CSVD) — and identified 36 papers for the task of CMB detection until November 2021, also for both type of approaches. Then, it is worth mentioning effort by Matsoukas et al. [37], where they investigate the performance of AI systems to detect ICH and CMBs. In this case they focus only on ML-based approaches and identify 18 relevant papers until early 2021. Of course, the aforementioned studies differ on the inclusion and exclusion criteria, but still provide an idea of the task dimensions on literature. Apart from these, other articles or books have also reviewed vaguely some approaches to the task [38], many times in relation to bigger contexts like CSVD [39], [40] and TBI or DAI [41], [42]. Finally, it is worth mentioning that the first studies on automatic CMB detection date back to 2011.

All reviews agree on existing approaches facing the same key challenges: defining the problem scope (classification, detection, or segmentation), handling limited data and class imbalance, dealing with high false positive (FP) rates mainly produced by the presence of various mimics and the availability and limitations of datasets. They also agree on some general recommendations like using transfer learning to account for datasets scarcity, incorporating spacial 3D information to distinguish CMB from mimics and to use a diverse enough datasets for robustness, testing on some benchmark for comparability. Regarding those that propose specific approaches, Ferlin et al. [8] suggest combining 2D analysis with 3D neural network-based detection and adding an additional CMB verification step. They also propose exploring the state-of-

the-art (SOTA) architectures for small object detection like RoI Transformer [43] and Oriented R-CNN [44]. Jiang et al. [36], however, suggest the use of knowledge distillation and introducing image artifacts like bias field correction to enhance model generalization, as well as pooling data from multiple sources to increase data size and diversity. They also put special focus on also evaluating performance using clinical measures.

B. Summary of ML-based approaches

Most research in the field of CMB detection divides the process into two distinct stages: the initial detection of CMBs and the subsequent verification phase to reduce FPs.

For CMB detection, two primary strategies are observed using deep learning: designing custom neural networks and employing transfer learning with pre-trained models. Custom neural networks in CMB detection have utilized various architectures like traditional Artificial Neural Networks (ANNs), Back-Propagation Neural Networks (BPNNs), Sparse Auto-Encoders (SAEs), and CNNs — the most common ones. Some studies have also incorporated Extreme Learning Machines (ELMs) for greater efficiency. Transfer learning approaches adapt well-known pre-trained architectures like AlexNet, ResNet50, Faster-RCNN, VGG, U-Net, YOLOv2, DenseNet 201, and SSD to the specific task of CMB detection.

For CMB verification, automated methods use machine learning classifiers (such as SVM, LDC, QDC, Parzen, and RFC) and algorithms (like 2D CNN, 3D ISA network, 3D Radon Transform) based on predefined CMB features. Some approaches utilize brain masks and region-growing techniques, while others employ 3D CNNs to incorporate spatial information, reducing false positives after initial 2D analysis.

C. Most recent approaches

Given the presence of a comprehensive review in the field, our attention will be directed towards summarizing the latest advancements postdating the review's submission. This entails an analysis of six new and pertinent papers, either published or awaiting publication on automatic segmentation of cerebral microbleeds using deep learning techniques

Ali et al. [45] employ a variant of U-Net model within an IoMT framework for CMB detection and segmentation. Their model is notable for its end-to-end design, which eliminates the need for any pre-processing or post-processing steps. In contrast, Wu et al. [46] developed a dual-task approach using Mask R-CNN for CMB detection, followed by a Multi-Instance Learning (MIL) network for cerebral small vessel disease (CSVD) classification. This method combines semantic segmentation with classification, and used internal datasets. Ferrer et al. [47] implemented the MultiResUNet architecture, trained on various heterogeneous datasets, including public datasets for normal aging, stroke, Alzheimer's disease, and an in-house dataset for COVID-19 assessment. The framework's obtained 78% sensitivity, 80% precision, and an average of only 1.6 false positives per scan, being particularly effective in low-resolution images. Sundaresan et al. [48] used a knowledge distillation framework within a multi-tasking teacher-student network. Their method, applied across four different datasets, achieved a cluster-wise true positive rate of over 90% with less than 2 false positives per subject. Lastly, Fang et al. [49] proposed a 2.5D convolutional neural network, utilizing a body plane detection framework, achieving a sensitivity of 98.24%, an accuracy of 94.10%, and maintaining an average of only 1.72 false positives per patient.

IV. Dataset Creation

The inherent problem in MIA of data availability explained in section II-C also holds for CMBs. As Ferlin et al. [8] point out, this is the main obstacle for most of the current automated CMB methods. What is more, in order to assess to what extent solutions can be used for clinical settings, it is required that datasets present simultaneously other markers of pathology to ensure that data represents the actual variability expected in real clinical scenarios [6]. In light of this, we developed a new annotated dataset of CMBs using diverse internal data from CEREBRIU, intended as groundwork for future studies. This effort highlights the challenges in creating clinically annotated datasets. For clarity, the current project encompasses only the data collection and establishment of annotation protocols, with the actual application of this dataset reserved for subsequent research.

A. Selected Studies

Within a dataset of 2080 case-level annotated studies from an internal study at CEREBRIU, we have selected **70 patients** that also feature microhemorrhages as incidental findings. Because the source dataset was originally curated to identify infarcts, tumors, and macro-hemorrhages, CMBs will predominantly coexist with these pathologies. Table I details the diverse pathologies and their subtypes present in this subset. It should be highlighted that the conditions listed are not mutually exclusive, which means a single patient may present with more than one of these. The distribution is as follows: 23 studies with infarct only, 18 with hemorrhage only, 3 with tumor only, 8 with both infarct and hemorrhage, and 18 with neither condition. Additionally, while all cases have been evaluated for other pathologies, such as White Matter Hyperintensities (WMH), vascular lesions, cortical atrophy, brain atrophy, demyelinating disease, Developmental Venous Anomaly (DVA), and cavernoma, these annotations may be incomplete as their reporting was not mandatory.

TABLE I
COUNTS OF PATHOLOGIES PRESENT IN SUBJECTS INCLUDED IN DATASET

Type	Subtype	Count
Infarct	Chronic infarct	20
	Hyperacute/acute infarct	26
	Subacute infarct	3
	Hemorrhagic infarct	5
Hemorrhage	Intra-axial chronic hemorrhage	12
	Intra-axial acute hemorrhage	19
	SDH/EDH acute	3
	SAH chronic	2
	SAH acute	4
	IVH	1
	SAH chronic	1
Tumor	Supratentorial extra-axial tumor w/o hemorrhage	2
	Intraventricular tumor w/o hemorrhage	1

Each study in the dataset is accompanied by a medical radiological report that provides insights on the findings for the various MRI sequences available. These reports were compiled by a radiologist distinct from the one who conducted the image-level annotations. Because of internal product requirements, the studies are ensured to comply with exclusion criteria, which rule out participants: under 18 or over 90 years old, with a history of neurosurgical procedures, diagnosed with cancer outside the brain, suffering from demyelinating or inflammatory conditions, and cases where the quality of the relevant MRI sequences (SWI/T2S) was inadequate for a reliable radiological evaluation. There is either a T2S or SWI sequence for every study. The data, sourced from various providers, comes from hospitals of three continents outside of Europe. Due to confidentiality regulations, data is anonymized and no patient-level metadata is available. The diverse MRI acquisition parameters presented in Table II represent a variety of combinations commonly employed in clinical practice.

B. Annotation process

Creating manual segmentation masks — strong annotations — is highly time-consuming, specially for CMBs. Instead, we decied to generate weak annotations, which are less resource-intensive and still effective for training [31]. Weak annotations can be categorized into three types of increasing information: image-level annotations, sparse annotations involving partial marking of slices or pixels, and model-generated or noisy annotations prone to under- and over-segmentation. We decided to use **scribbles**, a type of sparse annotations where annotators draw lines over the surface of CMBs, which we believe should suffice for estimating relevant pixel features given the distinct and limited characteristics of CMBs — typically round, low-intensity voids with a maximum size constraint.

We designed the annotation process to follow the visual rating scale **BOMBs** [18]. During the annotation process, we collect BOMBS-related metadata for each individual CMB as well as the possible cause of the microbleed. Each CMB's corresponding scribble is saved as a separate segmentation map or grouped with others in cases of CAA (to streamline metadata collection since CAA typically presents numerous CMBs). To account for future experimentation, we also collect data on the presence of mimics in patients, which might help in tuning the model's false positive reduction strategies. This includes a landmark point on the image indicating the approximate location and additional metadata specifying the type of mimic. Similarly, other relevant findings can also be marked with distinct landmarks. Finally, we collect quality-related data (e.g., presence of artifacts). For every study or patient being annotated, the radiologist can access its radiological report. We implemented all these requirements into an annotation project using the RedBrick framework. An example of this setup is shown in Figure 3. Cases have been assigned some priority based on the pathology present: first, cases with no extra pathology, then tumors and finally the rest.

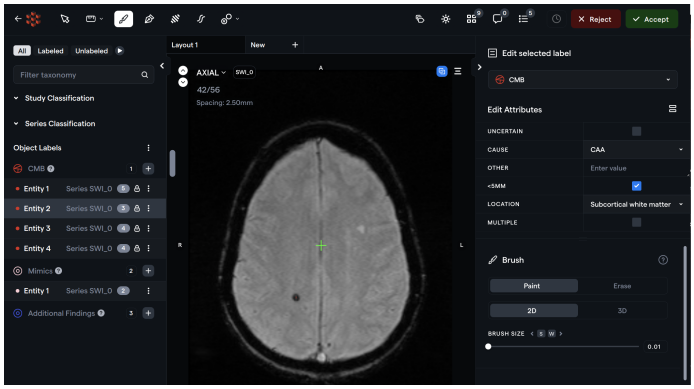

Fig. 3. **Example of annotation on Redbrick**. For a randomly selected annotated case, an axial slice of the SWI sequence is show for CMB finding number 2. It can be seen that guess on possible cause is CAA, that it has less than 5mm diameter and that location of CMB is in subcortical white matter. If looking closely at the CMB, one can see the scribble generated with the smallest brush size possible to avoid over-segmenting. Also, to the left one can see one instance of mimic annotated somewhere in the image as a landmark.

The annotations were conducted by an experienced rater (SI), who not only has expertise in annotating microbleeds — in fact one of the annotators of the VALDO dataset — but also has conducted scientific research on CMBs. While ideally multiple annotators would be involved given the task's complexity, the radiologist's extensive experience in CMB annotation enables us to treat these annotations as ground truth with sufficient confidence. As of the submission of this report, a total of 10 cases have already been annotated.

C. Generation of masks

To create volumetric masks from the sparse scribble annotations, which only cover a fraction of a 2D image's pixels (since not all slices are individually annotated), we employed a custom **Region Growing** algorithm. This algorithm treats every voxel within the scribble as a seed point and expands to adjacent voxels based on intensity similarity, using a 6-connectivity approach. It operates on a breadth-first search method, utilizing a FIFO queue initially filled with the seed points and their corresponding initial mask.

The algorithm examines each element's neighbors within the image boundaries, adding those with an intensity difference

TABLE II
SUMMARY OF MRI ACQUISITION PARAMETERS AND DEMOGRAPHICS FROM THE PROVIDED DATASET

Hospital-Location	Scanner Type	Scanner Model	Seq. Type	TR/TE (ms)	Flip Angle	Resolution	Voxel Size (mm^3)	n
Source1, Brazil	GE 1.5T	40%: Brivo MR355, 40%: GENESIS_SIGNA, 20%: Optima MR450w	T2S	8%: 267/19, 20%: 300-350 /23, 52%: 367-667 /20, 20%: 75/48	20%: 15, 80%: 20	12%: (256, 256, 22-4) 88%: (512, 512, 20-80)	88%: 0.47-0.51 x0.47-0.51 x3-5.5 12%: 0.94x0.94x5	25
Source2, India	Siemens 1.5T	ESSENZA	T2S	667-774/19	20	(256, 200-224, 25-27)	20%: 0.98x0.98x5, 80%: 0.9x0.9x5	5
Source3, India	Siemens 1.5T	80%: Symphony, 20%: ESSENZA	SWI	48-49/40	80%: 12, 20%: 15	(256, 176-208, 52-60)	10%: 0.86x0.86x2.5, 90%: 0.9x0.9x2.5	10
Source4, India	71%: Siemens 1.5T, 30%: Siemens 3T	64%: Sempra, 29%: Spectra, 6%: ESSENZA	T2S	36%: 657-711 /19, 64%: 807-936 /25	20	6%: (256,224,25), 30%: (320,270, 29-30), 64%: (512, 400-416, 25-30)	65%: 0.44-0.49 x0.44-0.49 x5, 24%: 0.72-0.98 x0.72-0.98 x4-5	17
Source5, U.S.A	Siemens 3T	MAGNETOM Vida	SWI	27/20	15	83%: (256, 208, 80), 17%: (256, 208, 88)	0.86-0.98 x0.86-0.98 x1.75-1.9	12

below a set tolerance threshold to both the queue and the mask. A maximum size limit for the mask is imposed to control unanticipated growth, based on the known maximum diameter of microbleeds (10mm) and the image resolution. For instance, with 1mm isotropic voxels, the volume of a 5mm radius sphere is approximately 105 voxels, which can serve as a conservative maximum size. The resulting masks can be considered as a form of noisy annotations, generated synthetically from the sparse data. Pseudo-code is shown in Algorithm 1.

The region growing algorithm employs three methods for assessing intensity differences: a) "Parent-son," comparing the intensity of each voxel with its neighboring 'parent' voxel; b) "Seed Average," where a voxel's intensity is compared to that of the initial seed point; and c) "Running Average," averaging the intensities of seed points and newly added voxels. To determine the most effective strategy, we applied the algorithm to the VALDO dataset, using the center of mass of each microbleed as a single seed point. The algorithm's performance was evaluated visually, as shown on Figure 6 for one case; and through the Dice score against actual label maps, as depicted on Figure 4.

To set the tolerance hyperparameter, crucial for our region growing algorithm, we implemented an automatic method for tuning it. This process iterates through tolerance values from 0 to 100%, observing changes in the region's size. Due to CMBs' distinct intensity contrast, a significant increase in region size is expected when tolerance is big enough to include surrounding pixels not in microbleed. We identify the optimal tolerance — which is computed for every image individually — as the point just before this marked size increase, that is, when going over maximum size allowed parameter. This approach is illustrated in Figure 5 where we clearly see for this subject an 'elbow' representing the ideal tolerance level.

Algorithm 1: Region Growing in 3D Volume

Input: Volume data V, Seed points S, Tolerance T, Connectivity type C, MaxSize threshold M, Intensity mode $modeI$

```
Initialize queue Q with S ;
Initialize region mask R with S;
Initialize reference intensity refI based on modeI;
Initialize total points totalVox = 0;
while queue Q is not empty do
    Dequeue a point p;
    foreach neighbor n of the current point p do
        if n is within bounds of V and not visited then
            Calculate intensity difference diff between
             n and refI;
            if diff < T then
                Add n to grown region R;
                Enqueue neighbor n in Q;
                Update refI based on modeI;
                Update totalVox;
                if totalVox > M then
                    break;
                end
            end
        end
    end
    Mark p as visited ;
endw
return R and metadata;
```

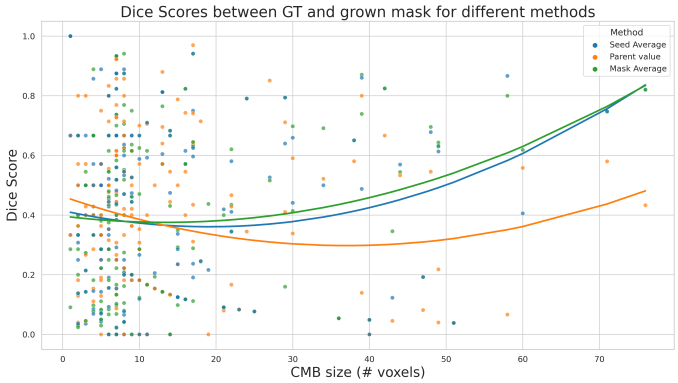

Fig. 4. **Comparison of Dice Scores for Different Region Growing Methods**. This plot illustrates the variation in Dice scores between ground truth (GT) of VALDO dataset and synthetic masks generated by the region growing algorithm across different sizes of cerebral microbleeds (CMBs) measured in totol number of voxels. Three methods of calculating intensity differences - Parent value, Seed Average, and Mask Average - are evaluated. The scatter plot shows individual data points, while the line plots indicate the trend for each method after smoothing with polynomial interpolation.

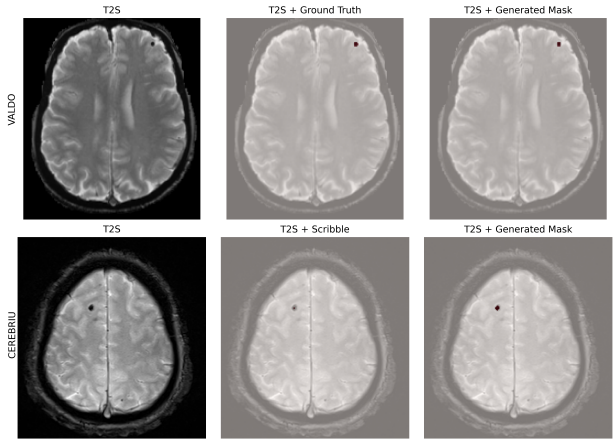

Fig. 6. **Comparative visualization of region growing results**. The top row illustrates data from sub-302 of VALDO dataset, showcasing a T2S sequence image (left), the same image with the ground truth overlay for a CMB (center), and the image with a generated mask using custom region-growing algorithm (right). The bottom row displays one study from CEREBRIU dataset, with a T2S sequence image (left), and the generated mask (right). The middle image displays here however the overlayed annotated scribble

This method not only ensures precise segmentation of CMBs but also enhances computational efficiency by avoiding exploring unnecessary tolerance levels. Upon completion of the region growing process, the algorithm employs morphological operations to refine the segmentation mask. This includes a closing operation (two iterations each of dilation and erosion) followed by hole filling. These steps ensure a coherent and accurate representation of the regions of interest, yielding a single connected component per CMB.

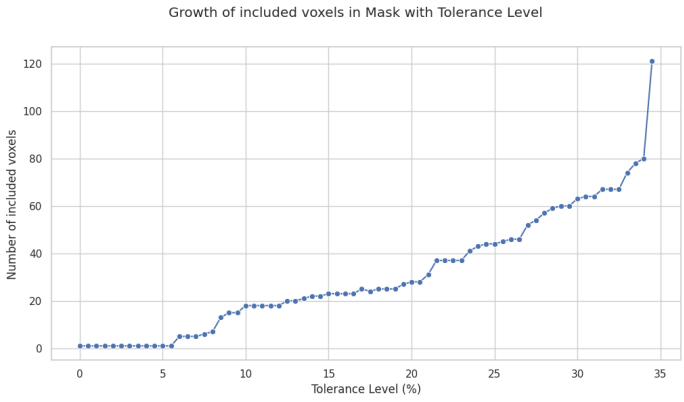

Fig. 5. **Growth of total voxels in mask as a function of tolerance level** for subject "sub-302" from VALDO dataset. Ground truth masks has 76 voxels, while grown regions 80, with a Dice score of 0.82. Optimal otlerance level is 34.5%. Method showed on this example is Running Average

Dice scores depicted in Figure 4 demonstrate varying effectiveness depending on the CMB size. This variability is expected, especially for smaller CMBs, where even minor misclassifications or slight over- or under-segmentation can significantly impact the metric. For larger CMBs, the 'Running Average' method appears most effective. However, it's important to note that comparisons with the 'Seed Points Average' are somewhat skewed, as these were based on a single voxel seed point, unlike the multiple voxel points typical in scribble annotations. For smaller CMBs, the 'Parent-Son' comparison method seems more appropriate. Adjusting the method based on CMB size, for which metadata will be available, might be advantageous. This approach, though effective in T2S sequences, needs further testing in SWI sequences for validation and maybe other approaches like Random Walker should be investigated.

V. Methods

A. Task setup

We will adopt the approach outlined in Metrics Reloaded framework [50] to select the appropriate problem category for our study. For the VALDO dataset we deal with 3D MRI images accompanied by 3D label maps that indicate the presence or absence of CMBs at the pixel level. A critical decision we face is whether to predict labels at the pixel or object (being a single CMB one object instance) level. As we previously discussed, this remains an unresolved issue in CMB segmentation. Clinically, the count and location of CMBs hold more significance than precise segmentation maps. Yet, we believe segmentation maps offer a clearer indication of the model's ability to identify objects accurately. Results for the VALDO challenge — which was assessed at the detection level mostly — indicate that segmentation performance aligned closely with detection performance, even though most participants treated detection as secondary to segmentation [6]. Consequently, we decided to classify our task as **Instance Segmentation**, which cobmines both object detection and semantic segmentation.

In the task of detecting CMBs, it's crucial to leverage three-dimensional data to distinguish them from their primary mimics — notably blood vessels, with their characteristic tubular structure — and DL-based systems have consistently outperformed other methods in this area. To address the previous aspects effectively, we have selected a **3D U-Net** as the model. This architecture is the backbone of many state-of-the-art models so it provides an excellent basis for generating results that can be readily compared with other methodologies.

Finally, we chose the **VALDO dataset** since it facilitates comprehensive benchmarking against existing methods and already includes domain variation by including 3 different population-based cohorts with different MRI acquisition protocols and scanner characteristics. In its original form, most subjects with a CMB contained only one, with an average of 13 voxels per CMB and a median of 7. The largest CMB comprised 147 voxels. Following recommendations from Horien et al. [51], we examined demographics and imaging measures of the dataset as shown in Table III), more info can be obtained at VALDO dataset. This dataset includes three retrospective cohort subsets (SABRE, RSS, ALFA) with annotations adhering to STRIVE guidelines. T2S, T2 and T1 weighted sequences are available for a total of 72 subject, along with pixel-level annotations.

B. Data Preprocessing

We optimized preprocessing for parallel execution on CPUs, applying it to T2S, T2, T1, and mask original NIfTI files, already registered to T2S sequence. The preprocessing produced two NIfTI files: one for MRI images and another for masks. The final image resolutions ranged between 250 and 395 voxels. Key steps, utilizing libraries such as *nibabel, nilearn, skimage, scipy.ndimage, numpy*, included:

Data Cleaning: Adjusted CMB masks to represent background as 0 and CMBs as 1, and replaced NaN values in MRI data with median background value or 0.

Resampling and Standardization: The primary sequence T2S was resampled isotropic 1x1x1mm voxel size (lower resolution version) and 0.5x0.5x0.5mm voxel size (higher resolution version), to which the rest of the images were then resampled. We used linear interpolation for MRI sequences and nearest-neighbor interpolation for the CMB mask.

Cropping Using Brain Mask: Applied Otsu's thresholding and morphological operations to crop MRI scans and annotations, focusing on the brain region.

Concatenation of Sequences: MRI sequences were then concatenated into a single multi-channel NIfTI, preserving headers and affine matrices.

C. Patch Sampling Strategy

To effectively manage memory constraints, we employ a patch sampling strategy for our 3D network, which processes small sub-volumes or patches of the original image. This approach still allows to detect small and localized nature of CMBs, although it requires tuning some extra hyperparameters. The key components of our strategy include:

- Patch Extraction: breaking down 3D volume into several patches of defined size, requiring to sometimes pad image to match dimensions. Params: *patch_size*
- Patch Sampling - training: not all patches from one patient can be included in a batch, so we randomly sample and allow to set a proportion for each type of patch (with and w/o CMB present). The position of the CMB inside a CMB-patch is randomly shifted. Params: *class_props*, *num_patches*
- Patch Sampling - evaluation: deterministic extraction of patches with a defined degree of overlap. When reconstructing volume, we combine logits in overlapping regions by adding them up. Params: *overlap_frac*

D. Training Metrics

The data suffers from huge class imbalance between the "background" label and the "cmb" label. For instance, in the original VALDO dataset, CMB pixels constitute merely 0.0007% of the total voxels. We account for this imbalance, as well as for the small size of the lesions, in our loss function during training by using the **Focal Tversky Loss** — a modification of the focal loss function based on the Tversky index — that was designed to achieve a better precision-recall balance for small lesion segmentation [52]. The Tversky index (TI_c) and the Focal Tversky Loss (FTL) are defined as follows:

$$TI_c = \frac{TP}{TP + \alpha \cdot FP + \beta \cdot FN} \tag{1}$$

$$FTL = \sum_{c=1}^{C} (1 - TI_c)^{\gamma} \tag{2}$$

Here, TI_c measures the similarity for each class c, with α and β in TI_c adjustable to balance false positives (FP) and false negatives (FN), crucial for handling the prevalent class imbalance. High TI_c values means better agreement with the ground truth. Note that with $\alpha = \beta = 0.5$, TI_c resembles the Dice Coefficient, and for $\alpha + \beta = 1$, we obtain a specific F_β score. The FTL parameter γ, when ranging [1, 3], increasingly emphasizes the correction of misclassified predictions, focusing the algorithm on improving the accuracy of predictions where it currently makes mistakes.

To further refine our approach, we created a weighted loss function that combines a class-weighted Categorical Cross-entropy Loss with the Focal Tversky Loss.

E. Validation Metrics

Due to the dual nature of the task, we evaluated both segmentation and detection aspects. For segmentation accuracy, we used the Dice Score, calculated as Dice Score $= \frac{2 \times |X \cap Y|}{|X| + |Y|}$, where X represents the predicted mask, and Y denotes the ground truth mask.

For detection quality, we used connected components analysis. This involved computing connected components in the predicted masks, with the total number representing positive predictions. We compared these against the connected components in the ground truth, which indicate the actual count

TABLE III
SUMMARY OF MRI ACQUISITION PARAMETERS AND DEMOGRAPHICS IN SABRE, RSS, AND ALFA COHORTS PRESENT IN VALDO DATASET FOCUSING ON T2S SEQUENCES.

Study	Demographics	Location	Scanner Type	TR/TE (ms)	Flip Angle	Voxel Size (mm^3)
SABRE	Tri-ethnic, high cardiovascular risk, mean age 72	London, UK	Philips 3T	1288/21	18°	0.45 x 0.45 x 3.0
RSS	Aging population	Netherlands	GE 1.5T	45/31	13°	0.5 x 0.5 x 0.8
ALFA	Enriched for APOE4, family risk of Alzheimer's	Barcelona, Spain	Philips 3T	1300/23	15°	1.0 x 1.0 x 3.0

of CMBs. For each predicted component, we checked for an overlap with any true CMB. If at least one voxel overlapped, it was considered a successful detection (True Positive, TP). The False Positives (FPs) were computed as the difference between successful and expected overlaps, and False Negatives (FNs) as the discrepancy between expected and actual overlaps.

Additionally, classification metrics were employed to evaluate the model's ability to detect at least one microbleed in patients. In this context, a True Positive (TP) indicated the correct detection of at least one CMB in a patient with CMBs, whereas a False Positive (FP) represented a false call for healthy patients.

Finally, we adhere to recommendations from [8] and add two additional metrics to inspect the CMB false positive issue: $FP_{avg} = \frac{FP}{n}$ and $FP_{cmb} = \frac{FP}{m}$, where n represents the number of subjects and m the number of CMBs in the test set. These metrics respectively quantify the average false positives per subject and per ground truth sample.

VI. EXPERIMENTS AND RESULTS

A. Experimental Setup

We divided the subjects from the VALDO dataset into training and validation sets, maintaining a 70% to 30% ratio and preserving the same proportion of 49:72 healthy-unhealthy patients in each split. Only T2S images were utilized among the available.

The following configurations were used in all experiments:

- Adam Optimizer with a learning rate of 5×10^{-5}.
- Combined loss function using parameters $\alpha = 0.3$, $\beta = 0.7$, and $\gamma = 3$ for Focal Tversky Loss; Class-weighted Categorical Cross-Entropy Loss with weights of 0.1 for background and 5 for CMBs.
- The losses were weighted as 1 for Focal Tversky Loss and 0.1 for Class-weighted Categorical Cross-Entropy Loss.
- Preprocessed VALDO dataset with isotropic voxels of 0.5 mm was used, after padding and cropping to image size (400, 400, 400).
- No whole-MRI level transforms or data augmentation applied. Same for patch-level
- Segmentation masks were generated using argmax approach from the obtained logits.

We chose such values for α and β, along with the balanced cross-entropy, to accelerate the model's learning of the minority CMB class by penalizing the under-segmentation of CMB voxels. Furthermore, setting γ to 3 is aimed at directing the model's attention towards smaller regions, accounting for the tpyical diminute size of CMBs. Additionally, we opted for a voxel size of 0.5mm to fully leverage the high resolution of our dataset and to avoid potential anomalies that might arise from downsampling one of the cohorts.

TABLE IV
ARCHITECTURE DETAILS OF THE 3D U-NET MODEL USED IN EXPERIMENTS.

Parameter	Value
Input Shape	$patch\_size$
Number of Channels	1
Filters per Depth	[32, 64, 128, 256]
Kernel Size	[3, 3, 3]
Number of Classes	2
Pool Size	[2, 2, 2]
Convolutional Parameters	Padding = "same", Strides = [1, 1, 1]
Convolution Layers per Depth	2
Hidden Activation Function	"elu"
Output Activation Function	"softmax"
Up Sampling Interpolation	"nearest"
Merge Layer	"Concatenate"
Normalization	"BatchNormalization"
Kernel Regularizer	"L2" with $l2 = 0.0001$

We employed the same patch sampling strategy across all experiments. An overlap of $overlap\_frac$=0.2 was chosen for evaluation patches, and a total of $num\_patches$=16 were collected from each subject in each epoch, comprising 14 patches containing at least one CMB voxel and 2 patches with only background, set by $class\_props$ of 0.1 and 0.9 respectively. This aims to have enough CMB-like patches given the imbalance. The architecture of the 3D U-Net model employed in our experiments is detailed in Table IV, which has a total of 26334178 parameters.

We conducted **three distinct experiments**, which yielded 3 different trained models. In Experiment 1, we utilized a patch size of [64, 64, 64] and the lower resolution version of the dataset, applying a Tversky loss with parameters ($\alpha = 0.5$, $\beta = 0.5$). Experiment 2 increased the patch size to [80, 80, 80] and employed the higher resolution dataset, maintaining the same Tversky loss parameters. Experiment 3 also used a patch size of [80, 80, 80] and the higher resolution VALDO dataset, but with adjusted Tversky loss parameters ($\alpha = 0.5$, $\beta = 0.5$). Each experiment was conducted over 5000 epochs, involving 100 training steps and 50 validation steps per epoch, with a batch size of 4. Each experiment involved 5000 epochs of training, 100 training steps, and 50 validation steps per epoch, on a batch size of 4. We utilized TensorFlow on CEREBRIU's server GPUs, and ClearML as MLOps framework. GPU was NVIDIA A100-SXM4-40GB, with CUDA version 11.4. Python version was 3.11.3 with tensorflow 2.14.0.

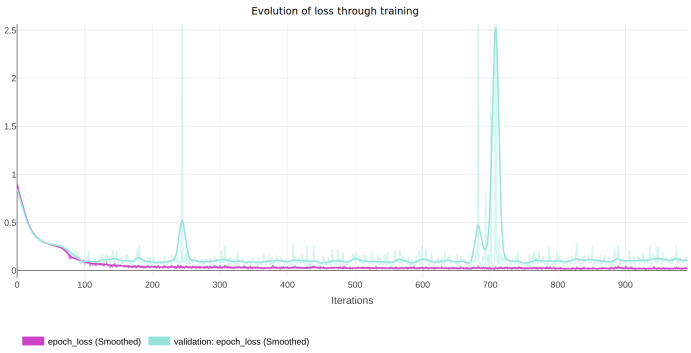

Fig. 7. **Training and validation loss**. Loss value for training and validation sets for first 1000 iterations (or gradient steps) of training. Smoothing has been applied with running average of window size 10

B. Results

Each model took two days to train. Figure 7 shows the first 1000 of 5000 gradient steps. Notably, the validation loss shows considerable fluctuations over time, marked by several peaks. This instability likely stems from the substantial penalties imposed on false negatives (FNs). The loss evolution also reveals a characteristic pattern seen in binary segmentation problems. Initially, the model primarily learns to identify the majority class (background) before advancing to distinguish the CMB class. Following these initial phases, the loss remained relatively stable for the rest of the training. The model selection was based on achieving the best validation loss throughout the training period.

Due to time constraints, evaluation of the models was done on the validation split (30% of the data) rather than on some external unrelated dataset. The resulting metrics can be seen in table V. Metrics with an asterisk "*" belong to the classification task of detecting at least one CMB for patients with 1 or more, as explained in Section V-D.

To get a more visual idea of how the model is performing, we plotted cropped regions of the axial slices around CMBs along with their ground truth masks (green) and predictions (red) by our model in Figure 8. We provide two examples from different subjects for every type of call, that is, correct ones but also FP and FNs. It can be seen that predictions have actually smoother contours than actual ground truth, which might present these shapes due to interpolation effects.

VII. Discussion

The outcomes of our experiments demonstrate how the 3D U-net model can learn to segment and detect cerebral microbleeds. Model 3, derived from Experiment 3, emerged as the most effective, achieving a recall/sensitivity of 0.71, precision of 0.44, and an F1 score of 0.54. While Model 2 exhibited higher sensitivity, its overall recall was significantly lower compared to Model 3, likely due to the aggressive weight assigned to FNs in the Tversky loss function. This aggressive tuning led to excessive over-segmentation in Model 2 to avoid FNs, as reflected in the lower FNavg but higher FPavg and FP/cmb rates. A more balanced performance was attained in Model 3 adjusting α and β to equal values. In terms of classification accuracy — detecting the presence of at least one microbleed in patients — Model 3 again showed superior performance with an F1 score of 0.87. This indicates the model's proficiency in distinguishing patients with and without microbleeds at the higher level.

The low Dice score observed in our results does not seem to align with the detection metrics, suggesting that the model may detect CMBs accurately without precisely segmenting them. Visual inspection of the predictions reveals however well-defined areas corresponding to detected microbleeds, hinting that the issue might be present at the ground truth masks. These discrepancies might stem from contour alterations in the ground truth masks during upsampling procedures, as evidenced in Figure 8, where the model's prediction for subject 11 appears more anatomically plausible than the ground truth. Given the varying resolutions of the cohorts in our study, it would be insightful to analyze how the different original resolutions impact results.

As anticipated, our model's primary challenge lies in the high number of false positives (FPs), affecting its precision. Despite this, the range values of the FP-based metrics fall within the typical spectrum observed in previous studies, suggesting that our model has effectively learned the task. This proves the need of a two-step approach, integrating both CMB candidate detection and subsequent verification to reduce FPs. To address the latter, one initial strategy could involve excluding detections outside anatomically plausible regions for CMBs. Implementing a tool like SynthSeg [53], to segment the brain and apply a mask could effectively filter out non-cerebral FPs. Additionally, equipping the network with a broader anatomical context — like whole slices or some attention mechanism — could further enhance its ability to discriminate between actual CMBs and mimics. Also, it's important to note that our study did not incorporate mask refinement or other post-processing techniques, which might offer further improvements in this matter.

Given the clinical implications of CMB detection, over-predicting (false positives) is preferable to under-predicting (false negatives). For example, patients with more than five CMBs face a heightened risk of ICH and associated mortality when undergoing antithrombotic therapy for chronic atrial fibrillation [22]. While a cautious approach towards over-segmentation may be warranted, it's essential for an automated system to balance accuracy and efficiency, reducing the burden on radiologists without overwhelming them with false positives.

Our visual inspection of the model's results, though limited by our non-medical expertise, reveals some key insights. The model seems to be distinguishing microbleeds from similar-looking structures in certain cases. For example, in Figure 8 for subject 221, it correctly ignores two structures that visually resemble microbleeds. This suggests that the model is effectively utilizing 3D information to make these distinctions. However, there are instances where the model appears confused.

TABLE V
COMBINED PERFORMANCE METRICS FOR EACH EXPERIMENT.[1]

Model	TPR	PPV	F1	TPavg	FPavg	FPmedian	FP/cmb	FNavg	Dice Score	TPR*	PPV*	F1*	TNR*	ACC*
1	0.61	0.22	0.32	1.64	5.91	4.0	2.20	1.05	0.18	1.00	0.68	0.81	0.00	0.68
2	0.79	0.14	0.24	1.36	8.23	5.5	4.76	0.36	0.22	0.87	0.72	0.79	0.29	0.68
3	**0.71**	**0.44**	**0.54**	**1.23**	**1.59**	**1.0**	**0.92**	**0.50**	**0.37**	**0.87**	**0.87**	**0.87**	**0.71**	**0.82**

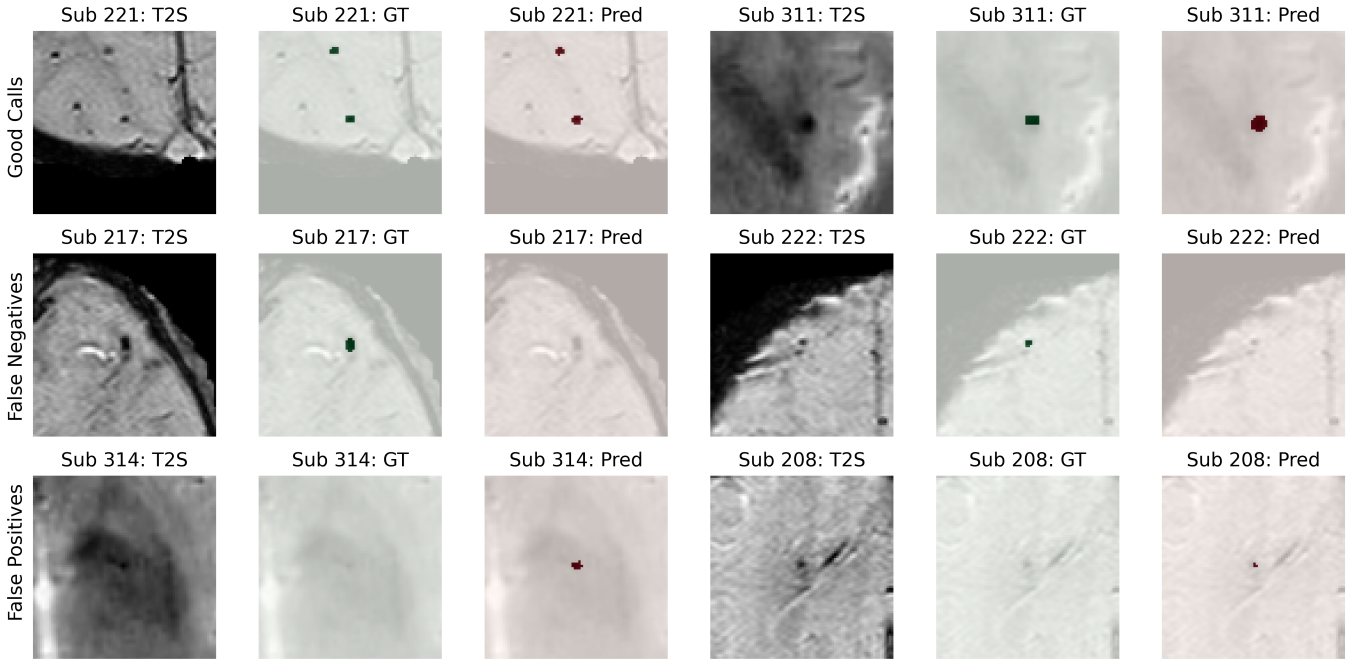

Fig. 8. **Illustrative examples of CMB detection using our best model** For every subject and CMB selected, the raw T2S sequences is presented, as well as its corresponding Ground Truth (GT) and model prediction (Pred). The microbleeds have been cropped around in the axial slice with squares of side 80 voxels, in an attempt to imitate the models field of view in a patch. The figure is organized into three rows representing different types of model predictions examples. *Top Row*: correct CMB detections where the model correctly identified/omitted CMBs. *Middle Row*: model failed to detect existing CMBs (False Negatives). *Bottom Row*: model incorrectly identified CMBs (False Positives).

Understanding why the model makes specific predictions, particularly in differentiating between true CMBs and their mimics, requires expertise beyond our reach. Collaborating with radiologists for a detailed assessment would greatly aid in interpreting these results. For this reason, during the design of the new dataset creation in this same project we included such metadata for better understanding and debugging false positives, among other things.

The evaluation methodology in this project did not fully adhere to best practices due to time constraints. Our approach of considering a single voxel overlap as a match could be overly optimistic. A better approach would be to use some distance metric for matching predictions to true CMBs, for instance the distance between the centers of mass. Additionally, our evaluation used the validation set instead of an unrelated test set, not reflecting real-world performance. To enhance comparability and robustness, future evaluations should include the VALDO dataset's online test set and/or real clinical data from the CEREBRIU dataset, which is meant to better represent clinical context. Assessing the model on datasets with varied acquisition parameters and demographics is crucial for a comprehensive performance evaluation. For instance, it's important to consider scenarios with lower resolution MRI scanners, as not all clinical settings have access to high-resolution equipment.

Datasets play a crucial role in deep learning, particularly in medical imaging. In our study, we primarily utilized the VALDO dataset, which, despite its variability, does not fully capture clinical reality. Exploring the use of the CEREBRIU dataset, developed as part of this study, along with other datasets, could pave the way for more robust and realistic training and evaluation of models. Significantly, Ferlin et al. [8] have identified over ten datasets used in previous studies with either SWI or T2S sequences, which could potentially be accessed for further research. Merging these datasets could lead to more comprehensive training and validation. Another interesting line of work would be to develop a model capable of processing both T2S and SWI sequences. This would be highly beneficial in clinical settings, as typically, either one or the other is used in standard hospital protocols.

Finally, it must be noted that only three different sets of hyperparameters were tested. This limited exploration, while in line with the project's initial goals to provide a baseline using a common architecture like 3D U-Net, suggests that broader experimentation could potentially lead to enhanced outcomes. Definitely, more complex architecture and ensembles could help in this challenging task. Approaches such as multi-task learning, transfer learning, and 2.5D CNNs appear promising. It might also be worth having a look at attention mechanisms — which originate in Natural Language Processing (NLP) and are now gaining traction in computer vision [54] — to provide global context to the model. All in all, our research successfully demonstrates that conventional architectures can perform competently in detecting CMBs, provided they are tuned in accordance with the latest methodological insights from literature.

VIII. Conclusion

In this work, we reviewed the literature as well as technical and medical aspects of the automated segmentation of cerebral microbleeds. Utilizing a 3D U-Net architecture, we addressed the segmentation and detection of CMBs within using VALDO challenge dataset. Our approach yielded a recall of 0.71, a precision of 0.44, and an F1 score of 0.54. Notably, the model predicted an average of 1.5 false positives per subject and a rate of 0.9 FPs for each true CMB. These outcomes, better than expected, demonstrates that conventional architectures can perform competently in detecting CMBs, provided they are tuned in accordance with the latest methodological insights from related literature. This motivates further exploration using more advanced architectures and techniques, as well as using more and more clinically representative dataset. On this line is the second contribution of this study, through the creation of a novel dataset. This dataset encompasses a diverse array of pathologies and features a broad spectrum of acquisition parameters and demographics, while containing metadata specifically useful for the task of CMB detection.

References

[1] S. Ingala, L. Mazzai, C. H. Sudre, G. Salvadó, A. Brugulat-Serrat, V. Wottschel, C. Falcon, G. Operto, B. Tijms, J. D. Gispert, J. L. Molinuevo, and F. Barkhof, "The relation between apoe genotype and cerebral microbleeds in cognitively unimpaired middle- and old-aged individuals," *Neurobiology of Aging*, vol. 95, pp. 104–114, 11 2020.

[2] S. Haller, M. W. Vernooij, J. P. Kuijer, E. M. Larsson, H. R. Jäger, and F. Barkhof, "Cerebral microbleeds: Imaging and clinical significance," pp. 11–28, 4 2018.

[3] W. H. O. Geneva, "Global health estimates 2020: Deaths by cause, age, sex, by country and by region, 2000-2019," 2020. [Online]. Available: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

[4] O. Colliot, *Machine Learning for Brain Disorders*. Springer Nature, 2023. [Online]. Available: http://www.springer.com/series/7657

[5] C. H. Sudre, B. G. Anson, S. Ingala, C. D. Lane, D. Jimenez, L. Haider, T. Varsavsky, R. Tanno, L. Smith, S. Ourselin, R. H. Jäger, and M. J. Cardoso, "Let's agree to disagree: learning highly debatable multirater labelling," 9 2019. [Online]. Available: http://arxiv.org/abs/1909.01891

[6] C. H. Sudre, K. V. Wijnen, F. Dubost, H. Adams, D. Atkinson, F. Barkhof, M. A. Birhanu, E. E. Bron, R. Camarasa, N. Chaturvedi, Y. Chen, Z. Chen, S. Chen, Q. Dou, T. Evans, I. Ezhov, H. Gao, M. G. Sanguesa, J. D. Gispert, B. G. Anson, A. D. Hughes, M. A. Ikram, S. Ingala, H. R. Jaeger, F. Kofler, H. J. Kuijf, D. Kutnar, M. Lee, B. Li, L. Lorenzini, B. Menze, J. L. Molinuevo, Y. Pan, E. Puybareau, R. Rehwald, R. Su, P. Shi, L. Smith, T. Tillin, G. Tochon, H. Urien, B. H. M. van der Velden, I. F. van der Velpen, B. Wiestler, F. J. Wolters, P. Yilmaz, M. de Groot, M. W. Vernooij, and M. de Bruijne, "Where is valdo? vascular lesions detection and segmentation challenge at miccai 2021," 8 2022. [Online]. Available: http://arxiv.org/abs/2208.07167

[7] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Çağatay Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, 10 2015.

[8] M. Ferlin, Z. Klawikowska, M. Grochowski, M. Grzywińska, and E. Szurowska, "Exploring the landscape of automatic cerebral microbleed detection: A comprehensive review of algorithms, current trends, and future challenges," 12 2023.

[9] F. Fazekas, R. Kleinert, G. Roob, G. Kleinert, P. Kapeller, R. Schmidt, and H.-P. Hartung, "Histopathologic analysis of foci of signal loss on gradient-echo t2*-weighted mr images in patients with spontaneous intracerebral hemorrhage: Evidence of microangiopathy-related microbleeds," pp. 637–642, 1999.

[10] S. M. Greenberg, M. W. Vernooij, C. Cordonnier, A. Viswanathan, R. A.-S. Salman, S. Warach, L. J. Launer, M. A. V. Buchem, and M. M. Breteler, "Cerebral microbleeds: a guide to detection and interpretation," pp. 165–174, 2 2009.

[11] A. L. Cheng, S. Batool, C. R. McCreary, M. L. Lauzon, R. Frayne, M. Goyal, and E. E. Smith, "Susceptibility-weighted imaging is more reliable than t2*-weighted gradient-recalled echo mri for detecting microbleeds," *Stroke*, vol. 44, pp. 2782–2786, 10 2013.

[12] S. Shams, J. Martola, L. Cavallin, T. Granberg, M. Shams, P. Aspelin, L. O. Wahlund, and M. Kristoffersen-Wiberg, "Swi or t2: Which mri sequence to use in the detection of cerebral microbleeds? the karolinska imaging dementia study," *American Journal of Neuroradiology*, vol. 36, pp. 1089–1095, 6 2015.

[13] S. Mittal, Z. Wu, J. Neelavalli, and E. M. Haacke, "Susceptibility-weighted imaging: Technical aspects and clinical applications, part 2," pp. 232–252, 2 2009.

[14] S. Buch, Y. C. N. Cheng, J. Hu, S. Liu, J. Beaver, R. Rajagovindan, and E. M. Haacke, "Determination of detection sensitivity for cerebral microbleeds using susceptibility-weighted imaging," *NMR in Biomedicine*, vol. 30, 4 2017.

[15] M. M. Poels, M. W. Vernooij, M. A. Ikram, A. Hofman, G. P. Krestin, A. V. D. Lugt, and M. Breteler, "Prevalence and risk factors of cerebral microbleeds: An update of the rotterdam scan study," vol. 41, 10 2010.

[16] A. A. Ibrahim, Y. A. Ibrahim, E. A. Darwish, and N. H. Khater, "Prevalence of cerebral microbleeds and other cardiovascular risk factors in elderly patients with acute ischemic stroke," *Egyptian Journal of Radiology and Nuclear Medicine*, vol. 50, 12 2019.

[17] S. M. Gregoire, U. J. Chaudhary, M. M. Brown, T. A. Yousry, C. Kallis, H. R. Jäger, and D. J. Werring, "The microbleed anatomical rating scale (mars): Reliability of a tool to map brain microbleeds," *Neurology*, vol. 73, pp. 1759–1766, 2009.

[18] C. Cordonnier, G. M. Potter, C. A. Jackson, F. Doubal, S. Keir, C. L. Sudlow, J. M. Wardlaw, and R. A.-S. Salman, "Improving interrater agreement about brain microbleeds: Development of the brain observer microbleed scale (bombs)," *Stroke*, vol. 40, pp. 94–99, 1 2009.

[19] T. J. Humphries and P. Mathew, "Cerebral microbleeds: hearing through the silence—a narrative review," pp. 359–366, 2 2019.

[20] H. Bokura, R. Saika, T. Yamaguchi, A. Nagai, H. Oguro, S. Kobayashi, and S. Yamaguchi, "Microbleeds are associated with subsequent hem-

orrhagic and ischemic stroke in healthy elderly individuals," *Stroke*, vol. 42, pp. 1867–1871, 7 2011.

[21] W. ming Lin, T. yen Yang, H. huei Weng, C. feng Chen, M. hsueh Lee, J. tsung Yang, S. N. Y. Jao, and Y. hsiung Tsai, "Brain microbleeds: Distribution and influence on hematoma and perihematomal edema in patients with primary intracerebral hemorrhage," pp. 184–190, 2013. [Online]. Available: www.centauro.it

[22] M. Fisher, "Mri screening for chronic anticoagulation in atrial fibrillation," 2013.

[23] C. Beaman, K. Kozii, S. Hilal, M. Liu, A. J. Spagnolo-Allende, G. Polanco-Serra, C. Chen, C. Y. Cheng, D. Zambrano, B. Arikan, V. J. D. Brutto, C. Wright, X. E. Flowers, S. P. Leskinen, T. Rundek, A. Mitchell, J. P. Vonsattel, E. Cortes, A. F. Teich, R. L. Sacco, M. S. Elkind, D. Roh, and J. Gutierrez, "Cerebral microbleeds, cerebral amyloid angiopathy, and their relationships to quantitative markers of neurodegeneration," *Neurology*, vol. 98, pp. E1605–E1616, 4 2022.

[24] Y.-L. Huang, Y.-S. Kuo, Y.-C. Tseng, D. Y.-T. Chen, W.-T. Chiu, and C.-J. Chen, "Susceptibility-weighted mri in mild traumatic brain injury from the department of diagnostic radiology (y," 2015.

[25] T. Tanino, Y. Kanasaki, T. Tahara, K. Michimoto, K. Kodani, S. Kakite, T. Kaminou, T. Watanabe, and T. Ogawa, "Radiation-induced microbleeds after cranial irradiation: Evaluation by phase-sensitive magnetic resonance imaging with 3.0 tesla," pp. 7–12, 2013.

[26] K. Nagata, T. Yamazaki, D. Takano, T. Maeda, Y. Ikeda, Y. Satoh, and T. Nakase, "P3-190: Cerebrovascular lesions and vascular risk factors in patients with alzheimer's disease," *Alzheimer's Dementia*, vol. 8, 7 2012.

[27] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," pp. 436–444, 5 2015.

[28] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 6 2017.

[29] M. Tsuneki, "Deep learning models in medical image analysis," pp. 312–320, 9 2022.

[30] A. Agarwal, R. Kumar, and M. Gupta, "Review on deep learning based medical image processing." IEEE, 2022.

[31] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, 7 2020.

[32] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," pp. 60–88, 12 2017.

[33] K. K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest, "A review of medical image segmentation algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, 2021.

[34] Sakshi and V. Kukreja, "Image segmentation techniques: Statistical, comprehensive, semi-automated analysis and an application perspective analysis of mathematical expressions," pp. 457–495, 1 2023.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 5 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[36] J. Jiang, D. Wang, Y. Song, P. S. Sachdev, and W. Wen, "Computer-aided extraction of select mri markers of cerebral small vessel disease: A systematic review," 11 2022.

[37] S. Matsoukas, J. Scaggiante, B. R. Schuldt, C. J. Smith, S. Chennareddy, R. Kalagara, S. Majidi, J. B. Bederson, J. T. Fifi, J. Mocco, and C. P. Kellner, "Accuracy of artificial intelligence for the detection of intracranial hemorrhage and chronic cerebral microbleeds: a systematic review and pooled analysis," *Radiologia Medica*, vol. 127, pp. 1106–1123, 10 2022.

[38] X. Zhao and X. M. Zhao, "Deep learning of brain magnetic resonance images: A brief review," pp. 131–140, 8 2021.

[39] J. Chojdak-Łukasiewicz, E. Dziadkowiak, A. Zimny, and B. Paradowski, "Cerebral small vessel disease: A review," pp. 349–356, 3 2021.

[40] L. Zhao, A. Lee, Y. H. Fan, V. C. Mok, and L. Shi, "Magnetic resonance imaging manifestations of cerebral small vessel disease: Automated quantification and clinical application," pp. 151–160, 1 2021.

[41] A. Alberts and B. Lucke-Wold, "Updates on improving imaging modalities for traumatic brain injury," *Journal of Integrative Neuroscience*, vol. 22, p. 142, 10 2023.

[42] L. Zinnel and S. A. Bentil, "Convolutional neural networks for traumatic brain injury classification and outcome prediction," *Health Sciences Review*, vol. 9, p. 100126, 12 2023.

[43] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," 2019.

[44] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," 2021. [Online]. Available: https://github.com/jbwang1997/

[45] Z. Ali, S. Naz, S. Yasmin, M. Bukhari, and M. Kim, "Deep learning-assisted iomt framework for cerebral microbleed detection," *Heliyon*, vol. 9, p. e22879, 12 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405844023100879

[46] R. Wu, H. Liu, H. Li, L. Chen, L. Wei, X. Huang, X. Liu, X. Men, X. Li, L. Han, Z. Lu, and B. Qin, "Deep learning based on susceptibility-weighted mr sequence for detecting cerebral microbleeds and classifying cerebral small vessel disease," *BioMedical Engineering Online*, vol. 22, 12 2023.

[47] N. R. Ferrer, M. V. Sagar, K. V. Klein, C. Kruuse, M. Nielsen, and M. M. Ghazi, "Deep learning-based assessment of cerebral microbleeds in covid-19," 1 2023. [Online]. Available: http://arxiv.org/abs/2301.09322

[48] V. Sundaresan, C. Arthofer, G. Zamboni, A. G. Murchison, R. A. Dineen, P. M. Rothwell, D. P. Auer, C. Wang, K. L. Miller, B. C. Tendler, F. Alfaro-Almagro, S. N. Sotiropoulos, N. Sprigg, L. Griffanti, and M. Jenkinson, "Automated detection of cerebral microbleeds on mr images using knowledge distillation framework," *Frontiers in Neuroinformatics*, vol. 17, 2023.

[49] Z. Fang, R. Zhang, L. Guo, T. Xia, Y. Zeng, and X. Wu, "Knowledge-guided 2.5d cnn for cerebral microbleeds detection," *Biomedical Signal Processing and Control*, vol. 86, 9 2023.

[50] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, M. Wiesenfarth, A. E. Kavur, C. H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. Blaschko, M. J. Cardoso, V. Cheplygina, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, H. Kenngott, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. I. Sánchez, S. Shetty, M. van Smeden, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. V. Calster, G. Varoquaux, and P. F. Jäger, "Metrics reloaded: Pitfalls and recommendations for image analysis validation," 6 2022. [Online]. Available: http://arxiv.org/abs/2206.01653

[51] C. Horien, S. Noble, A. S. Greene, K. Lee, D. S. Barron, S. Gao, D. O'Connor, M. Salehi, J. Dadashkarimi, X. Shen, E. M. Lake, R. T. Constable, and D. Scheinost, "A hitchhiker's guide to working with large, open-source neuroimaging datasets," *Nature Human Behaviour*, vol. 5, pp. 185–193, 2 2021.

[52] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," 10 2018. [Online]. Available: http://arxiv.org/abs/1810.07842

[53] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. V. Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, 5 2023.

[54] Y. Xie, B. Yang, Q. Guan, J. Zhang, Q. Wu, and Y. Xia, "Attention mechanisms in medical image segmentation: A survey," 5 2023. [Online]. Available: http://arxiv.org/abs/2305.17937

APPENDIX A
SOURCE CODE

Most of the scripts used for this project can be found in https://github.com/jorgedelpozolerida/Segmentation_CMB.git

APPENDIX B
DATA SOURCE

The different datasets used were downloaded from the following sources:

- VALDO challenge: https://zenodo.org/record/4687995