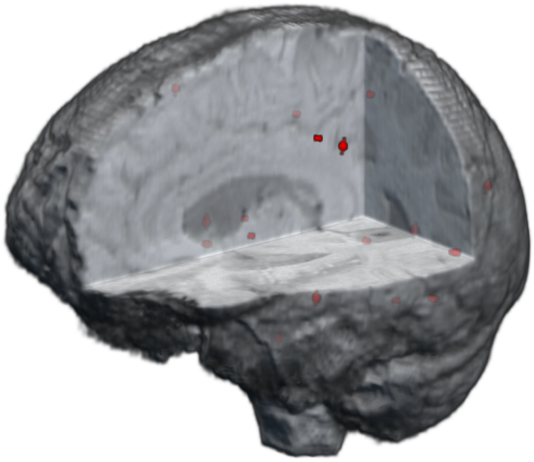
IT UNIVERSITY OF COPENHAGEN

# Automatic Detection of Cerebral Microbleeds: A Clinically Robust Deep Learning Framework

*Jorge del Pozo Lérida*



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

CEREBRIU supervisor:
Mathias Perslev (mp@cerebriu.com)

CEREBRIU co-supervisors:
Silvia Ingala (si@cerebriu.com)
Akshay Pai (ap@cerebriu.com)

**Master Thesis Project**
*KISPECI1SE*
MSc in Data Science
*June 3rd, 2024*

**Jorge del Pozo Lérida**
jord@itu.dk

*Abstract*

Cerebral Microbleeds (CMBs) are neuroimaging biomarkers visible as small round hypointensities on magnetic resonance images (MRI) in T2*-weighted (T2S) or susceptibility-weighted imaging (SWI). Associated with over 30 medical conditions, the accurate quantification and localization of CMBs are crucial for diagnostic and prognostic assessments. However, their manual detection by radiologists is labor-intensive and error-prone, particularly when numerous CMBs present, positioning them as prime candidates for automated detection. Despite this, automation remains challenging due to CMBs' small size, the scarcity of publicly available annotated data, and their similarity to various other biological mimics. The interpretation of the performance of existing methods is compounded by a lack of standardized metrics and task definitions, incomplete metrics reporting, imperfect evaluations, and the absence of a robust benchmark for comparison. Current methods often exhibit suboptimal performance, characterized by a high rate of false positives, and are not trained or evaluated on data representative of the variations found in clinical settings, which typically include a broad range of demographic, pathological, and MRI data variation. In this study, we develop a sequence-agnostic model applicable to SWI or T2S that is robust against the data variation commonly found in real-life clinical settings. We enhance our model's robustness through the curation of a large collection of public and private data, supplemented by advanced data augmentation and an initial pretraining phase with a large set of negative samples and synthetic CMBs. In parallel, we investigate the benefit of using transfer learning from a bigger source segmentation task with a larger and more representative dataset but find no tangible improvement. We conduct a rigorous evaluation first on a public dataset for benchmarking our model, achieving 69% recall and 84% precision across the entire test set with an average of 0.1 false positives per scan. Next, we evaluate the model on an in-house annotated dataset, crafted to simulate challenging real-life conditions. Performance metrics show a drop to 30% recall and 70% precision, with 0.13 false positives per scan. We find that the reason is the model struggling with insufficient slice thickness, which makes CMBs that look elongated like veins become false negatives. Additionally, the high number of CMBs per scan poses a significant challenge and leads us to raise concerns about the reliability of inter-rater agreement assessments for existing rating methods.

# Acknowledgements

On the CEREBRIU side, my sincere thanks to Mathias Perslev for his invaluable guidance and support, exceptional human virtues, and for introducing me to the fascinating world of medical image segmentation. I also want to express my gratitude to Silvia Ingala for her insightful clinical perspectives and for patiently enduring my constant demands for more microbleeds. Finally, I am deeply grateful to Akshay Pai for placing his trust in me and providing me with this incredible opportunity.

On the ITU side, I want to thank Veronika Cheplygina, who has helped me bring rigor and structure to my research and provided me with invaluable feedback. I also extend my gratitude to the rest of the PURRlab team, whose insights have been incredibly useful on numerous occasions.

Most importantly, I am deeply grateful for the unwavering support and love from my parents and family, who have helped me navigate through periods of stress and have generously shared their valuable expertise. I also extend heartfelt thanks to my partner, Maria Rebassa, who has consistently been there to lend an ear to my concerns.

And of course, a special thank you to everyone with whom I have had discussions about those tiny fellows known as microbleeds—your conversations have enriched my journey.

# List of Common Abbreviations

**AI** Artificial Intelligence

**BOMBS** Brain Observer MicroBleed Scale

**CMB** Cerebral Microbleed

**CNN** Convolutional Neural Network

**DL** Deep Learning

**DSC** Dice Similarity Coefficient

**FN** False Negative

**FP** False Positive

**GT** Ground Truth

**MARS** Microbleed Anatomic Rating Scale

**MIA** Medical Image Analysis

**MIS** Medical Image Segmentation

**ML** Machine Learning

**MRI** Magnetic Resonance Imaging

**SWI** Susceptibility Weighted Imaging

**T** Tesla

**T2S** Gradient Recalled Echo T2*-weighted Imaging

**TE** Echo Time

**TL** Transfer Learning

**TN** True Negative

**TP** True Positive

**TR** Repetition Time

# Contents

# 1 Introduction

Cerebral Microbleeds (CMBs) are neuroimaging biomarkers detectable as round hypointensity foci on magnetic resonance images (MRI) in T2*-weighted or susceptibility-weighted sequences with less than 10 mm in diameter [1]. They arise from the accumulation of hemosiderin, a byproduct of blood degradation, within the brain parenchyma following microscopic hemorrhages [1, 2]. CMBs have been associated with more than 30 medical conditions [3], making them critically important for diagnostic and prognostic assessment and treatment planning. The ability to accurately quantify and localize these lesions holds significant clinical utility. For instance, the presence of >2 CMBs increases the risk of parenchymal haemorrhage on the day following intravenous thrombolysis and predicts a poor clinical outcome [4]. Furthermore, they can assist in predicting the likelihood of cerebral amyloid angiopathy (CAA) in cases where more than one strictly lobar CMB is detected. Furthermore, the presence of >4 microbleeds is associated with cognitive decline [5]. Currently, CMBs are manually labelled by trained radiologists, but this procedure is laborious, time-consuming, and prone to error, particularly when numerous CMBs are present [6, 7].

This situation underscores the need for automated detection systems. However, this task is highly challenging due to their small size, variations in MRI contrast, sparse distribution, and the presence of imaging artifacts. Additionally, the existence of various mimics, such as calcifications and pial vessels, makes accurate detection very difficult [8]. Various semi-automated and automated approaches have been proposed. Typically, these methods incorporate a multi-step process that includes an essential step for reducing false positives. This is crucial as the proposed methods often suffer from very low precision, with average rates of up to hundreds of false positives per scan [9, 10]. Although there has been a notable improvement in the accuracy of CMB detection in recent years through the adoption of deep learning-based approaches, it remains insufficient, and a clinically integrated automated solution for CMB detection is yet to be realized.

However, interpreting published approaches to CMB detection is highly challenging due to significant gaps in evaluation methodologies. Often, papers report an insufficient range of evaluation metrics, typically omitting metrics that provide complementary insights—such as reporting recall but not precision. Moreover, the computation of these metrics is frequently flawed, overall obscuring potential issues in CMB detection. Additionally, the metrics employed often correspond to disparate and unclearly defined tasks, such as binary classification at the image level (determining the presence of more than X microbleeds), classification at the patch level, and detection of specific CMB objects. Moreover, many researchers use test sets that are merely separate splits of the same cohort used for training, rather than distinct datasets specifically reserved for unbiased testing, which leads to inherently biased evaluations.

Another major obstacle in advancing CMB detection is the scarcity of public annotated CMB data and the lack of a clear benchmark dataset. Yet, some datasets are available [11, 12], and initiatives like the VALDO challenge [13] have emerged to propel advancements in automated CMB detection by emulating the successful dynamics seen in challenges like BRATS [14]. However, the datasets commonly used are too small and do not adequately represent the complexity of data encountered in real clinical practice, failing to capture essential variations in MRI acquisition parameters, demographics, and co-occurring pathologies. Thus, it is crucial to assess and ensure robustness against domain shifts, which occur with particular frequency in the context of MRI.

This project is performed as a continuation of previous work [15] and in collaboration with CEREBRIU. GitHub repository for this project can be found at https://github.com/jorgedelpozolerida/MicrobleedNet.git

## Project Aim

We aim to develop an automated system capable of accurately detecting CMBs on MRI images of high clinical and technical variability, with the ultimate goal of providing significant clinical value and overcoming the gaps identified in previously published methods. To achieve this, we do the following:

To this end, we have designed a sequence-agnostic deep learning model, capable of working on the two most commonly employed MRI sequences for CMB detection: T2*-weighted Gradient Recalled Echo and Susceptibility Weighted Imaging. To enhance the robustness of our model, we have implemented a comprehensive strategy that includes the creation of a large collection of both public and private CMB datasets. These datasets have been meticulously processed and enhanced with data augmentation techniques to reflect the variety of data variations typically encountered in clinical environments. Additionally, we utilize synthetic datasets and a significant amount of internally available negative data during an initial pre-training phase. In parallel, we explore the potential benefits of utilizing a pre-trained model for closely related tasks, aiming to leverage extensive data variations present in its training dataset.

Addressing the challenges found in the literature concerning the evaluation CMB detection methods, we conduct an in-depth performance evaluation that adheres to best practices in medical image analysis. This evaluation is meticulously performed on completely independent test sets—one publicly available for benchmarking our method and another privately collected to accurately reflect the diverse and challenging data variations encountered in clinical practice. Through this rigorous evaluation process, we aim to uncover and address potential flaws and challenges within our solution, the data, and the inherent complexities of the task itself.

# 2 Background

This chapter offers a detailed background to enhance understanding of the clinical importance of cerebral microbleeds, their detection methods in clinical settings, and the technological strategies for automating these processes. Initially, it introduces magnetic ressonance imaging, the medical imaging modality employed for detecting CMBs. Subsequently, the chapter delves into a detailed discussion of CMBs, including their clinical relevance, detection techniques, and the challenges associated with their identification. Finally, it explores the specific machine learning task of medical image segmentation applied to tackle the problem, highlighting the principal challenges and methods involved.

## 2.1 Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (MRI) is an advanced imaging technique that can generate detailed images of internal body structures. MRI utilises two primary energy sources: strong magnetic fields and radiofrequency (RF) waves. The method leverages the magnetic properties of specific nuclei within the body, predominantly hydrogen, due to its abundance from water molecules. Routine clinical MRI scanners use magnetic fields at field strengths between 0.2 to 3.0 Tesla (T).

### 2.1.1 Image acquisition

The MRI imaging process begins with the application of an initial magnetic field (B0) to align the protons. This is followed by the perturbation of the protons by a radio frequency (RF) pulse, which is perpendicular to the initial field. This RF pulse shifts the net magnetization from the longitudinal plane (z-direction) to the transverse plane (x-y), causing the magnetization to process transversely. This change generates an electric field that is picked up by a coil as these nuclei return to their baseline states. Data acquisition involves sampling these emitted signals, which, thanks to the use of spatial encoding gradients, are sampled directly in the frequency domain, known as k-space. These signals are subsequently transformed into anatomical images through the inverse Fourier transform. Three primary types of relaxation times exist:

- **T1 (Longitudinal Relaxation Time):** This is the period required for protons to realign with the longitudinal (z) axis after magnetic excitation. T1 values vary among tissues, providing valuable structural information.
- **T2 (Transversal Relaxation Time):** T2 measures the time constant for the decay of transverse magnetization due to natural interactions at the atomic or molecular levels.
- **T2* (Observed Transverse Relaxation Time):** similar to T2 but including the susceptibility effect, caused by magnetic field inhomogeneities that lead to phase coherence loss and rapid signal decay, visible as dark spots on MRI images.

The contrast in MRI images is shaped by how proton density, T1 and T2 are weighted. Manipulating specific scanning parameters allows for the adjustment of image contrast. The two most critical parameters are:

- **Repetition Time (TR):** TR is the interval between consecutive RF pulses.
- **Echo Time (TE):** TE is the duration between the application of the RF pulse and the echo signal measurement.

### 2.1.2 MRI sequences

Adjusting parameters such as TR and TE allows for the creation of various MRI sequence types, each tailored to highlight specific tissue characteristics. Regardless of the sequence used, the objectives remain consistent: to enhance the signal from a particular tissue to improve contrast, maximize efficiency by reducing scan time, minimize artifacts, and maintain an optimal signal-to-noise ratio [16].

One essential MRI sequence type is the Gradient Recalled Echo (GRE), characterized by its use of lower flip angles, typically less than 90 degrees, and the absence of a 180-degree radiofrequency rephasing pulse. This configuration facilitates quicker image acquisition and is particularly effective in enhancing features related to magnetic susceptibility, allowing for T2* measurements. Magnetic susceptibility refers to how much certain materials, such as iron-containing compounds found in blood, become magnetized when placed in an external magnetic field. Within this family lie the two most common sequences used to detect CMBs.

- **T2*-weighted GRE (T2S)**: Developed through GRE pulse sequences with echo times carefully chosen to enhance contrast sensitivity. Typically acquired as two-dimensional axial images with a slice thickness of 3–5 mm at 1.5 Tesla, this sequence is particularly sensitive to magnetic susceptibility effects due to parameters such as low flip angle, long echo time, and long repetition time. The effectiveness of T2S increases with higher magnetic field strengths (e.g., 3 T or 7 T), enhancing susceptibility effects and signal-to-noise ratio.
- **Susceptibility-weighted Imaging (SWI)**: Utilizes three-dimensional GRE imaging enhanced by advanced post-processing techniques to emphasize susceptibility contrasts, thus improving the visibility of hemosiderin deposits. SWI sequences are generally captured with magnetic field strengths of 1.5T or 3.0T and incorporate flow compensation in all three planes to reduce artifacts. This sequence is designed to maximize the susceptibility effect by employing long echo times and both magnitude and filtered phase information, albeit with longer acquisition times compared to T2S.

Over a hundred unique MRI sequence types exist, and their naming conventions vary depending on the MRI machine's manufacturer [17]. For example, the term 'susceptibility-weighted imaging' is designated differently across vendors: Siemens uses the trademark 'SWI,' GE Healthcare refers to it as 'SWAN,' and Philips Healthcare labels it 'SWIp.' The availability of these sequences can also differ due to licensing and patent constraints [18].

## 2.2 Cerebral Microbleeds (CMBs)

Cerebral microbleeds (CMBs), also known as microhemorrhages, are a radiological construct characterized as hypointense foci detected on T2*-weighted and susceptibility-weighted MRI sequences [17]. Different studies have employed various size cut-off points to classify microbleeds, typically setting a maximum diameter between 5 and 10 mm and a minimum diameter of 2 mm in some cases [8, 19]. Histopathologically, they arise from the accumulation of hemosiderin, a byproduct of blood degradation, within the brain parenchyma following microscopic hemorrhages [1, 2]. Consequently, they exhibit paramagnetic properties, which result in signal loss due to susceptibility effects.

Over the past two decades, CMBs have gained prominence as both diagnostic and prognostic imaging biomarkers [20], being associated with more than 30 pathologies. Their detection requires trained radiologists, and knowing the number and location of CMBs in a patient can provide invaluable clinical insights. This section will review some key concepts for detecting CMBs on MRI, explore factors that influence variability in detection, discuss associated pathologies, and evaluate the current rating scales used in clinical practice

### 2.2.1 Detection

The prevalence of CMBs varies widely, influenced by factors such as age, patient health conditions, and neuroimaging techniques [20]. In the elderly, prevalence ranges from 5% to 35% according to population studies. In patients with ischemic strokes or spontaneous intracerebral hemorrhage, around 60% exhibit CMBs, with 40% developing new CMBs within about 27 months post-stroke. Additionally, CMB prevalence is notably high in genetic small vessel diseases and Alzheimer's disease, where one in five patients has CMBs.

The visibility of CMBs, characterized by a blooming effect, is affected by various factors such as resolution, signal-to-noise ratio, echo time, field strength, and magnetic susceptibility [21]. Although CMBs can be detected on computed tomography (CT) scans, they are most visible shortly after occurrence, typically fading within 7 to 10 days. Therefore, MRI is considered the superior imaging modality for CMB detection due to its ability to maintain prolonged visibility of these lesions.

The detection of CMBs is significantly enhanced by higher magnetic field strengths, achieving the highest sensitivity with ultra-high fields such as 7 T MRI. At clinical field strengths of 1.5 T and 3.0 T, only about 50% of CMBs are detectable. While T2S has been a longstanding standard for detecting CMBs [8], SWI has shown greater reliability and sensitivity [22, 23]. For instance, in the case of mild cognitive impairment, the detection rate increases from approximately 20% on routine two-dimensional T2S images to approximately 40% on SWI images at 3.0 T [24]. However, increases the visibility of other structures that mimic CMBs and exaggerates the blooming effect, even causing CMBs to appear more irregular in shape. Despite this, SWI can reveal up to six times more CMBs than T2S [25]. Typically, SWI is performed using magnetic field strengths of 1.5T or 3.0T, although it can also be employed at 7.0T. Recent advances such as Quantitative Susceptibility Mapping (QSM) provide a more precise quantitative assessment of tissue susceptibility changes across different MRI parameters and over time. Nonetheless, T2S and SWI remain the primary methods for CMB detection in both automated systems and routine clinical practice.

### 2.2.2 Mimics

CMBs are often misidentified due to several **mimics**, which can complicate their accurate detection by introducing false positives, and include:

- **Deposits of Calcium and Iron:** These appear as small, low signal intensity spots. Can be distinguished using phase information.
- **Flow Voids from Pial Blood Vessels:** These are seen as linear structures in cross-sectional views of cortical sulci, maintaining a consistent appearance across various MRI sequences, unlike the more spherical CMBs.
- **Paramagnetic Deoxyhemoglobin in Cerebral Venules:** This produces a blooming effect similar to CMBs but can be distinguished by its tubular structure.
- **Partial Volume Artifacts:** Particularly from bone close to sinuses, these artifacts can either mimic or obscure CMBs, especially near the temporal and frontal lobes.
- **Cavernous Malformations:** Sometimes confused with CMBs, these are characterized by stagnant blood within their sinusoidal lumen and a hemosiderin rim on MRI scans.
- **Metastatic Melanoma:** Presents as hypointense areas due to melanin and hemorrhage, differentiated from CMBs by additional T1 hyperintensity or surrounding edema.
- **Diffuse Axonal Injury (DAI):** Resulting from head trauma, DAI can lead to secondary CMB-like appearances, identifiable through clinical history and imaging findings.

These mimics occur in approximately 11% to 24% of evaluations and complicate the clear identification of CMBs [26]. Figure 2.1 illustrate some of these mimics-

### 2.2.3 Associated pathology

CMBs are linked to more than thirty medical conditions [3]. In general, the quantity and location-specific distribution of CMBs (deep vs. lobar) are crucial for associating them with various pathologies and appropriately adapting clinical management strategies.

Particularly in cases of intracerebral hemorrhage (ICH), CMBs serve as a substantial independent predictor of initial ICH occurrences [3]. They are crucial in understanding stroke pathology, capable of forecasting both hemorrhagic and ischemic strokes in asymptomatic individuals. Additionally, CMBs help differentiate

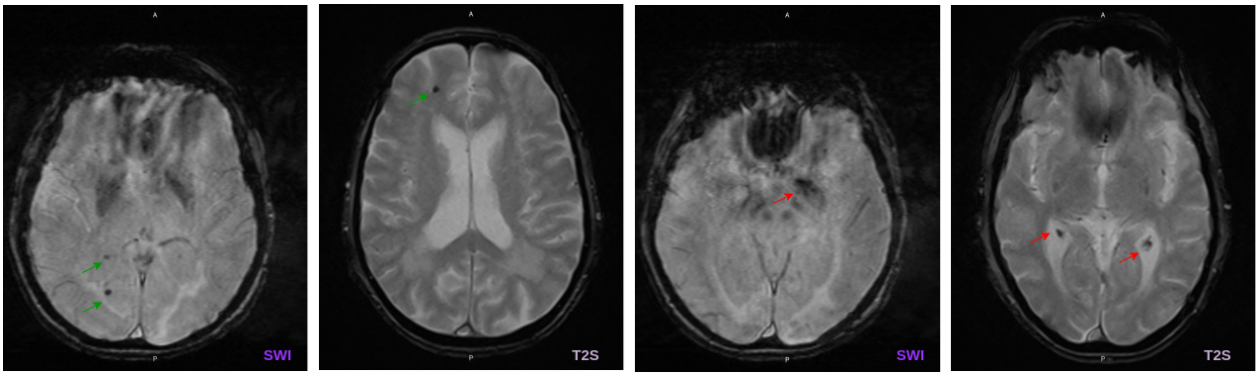



Figure 2.1: **Example of a CMB and mimics on both T2S and SWI**. From left to right: CMB on SWI, CMB on T2S, Basal ganglia calcification on SWI, calcified choroid plexus on T2S. All images are from CRB dataset from 2 different scans

between stroke types—often impacting deeper brain areas—and degenerative diseases, which are typically lobar [27, 28]. Having **more than 2 CMBs** has been demonstrated to elevate the risk of parenchymal haemorrhage 24 hours after intravenous thrombolysis and to predict an unfavorable clinical outcome independently [4]. Also, a high CMB burden of **more than 10 CMBs** significantly impacts the outcomes of reperfusion therapies like intravenous thrombolysis (IVT) for acute ischemic stroke (IS).

In patients with atrial fibrillation (AF) who are undergoing antithrombotic therapy, having **more than five CMBs** markedly raises the risk of ICH and associated mortality [29]. Furthermore, CMBs serve as markers for cerebral amyloid angiopathy (CAA), typically manifesting in peripheral brain regions, as opposed to their occurrence in the basal ganglia or infratentorial areas, which suggests hypertensive arteriopathy [30]. This distinction is crucial for assessing hemorrhage risks when administering antiplatelet, antithrombotic, or thrombolytic treatments.

In traumatic brain injury (TBI), the number and severity of cerebral microbleeds (CMBs) are linked to the extent of the injury and can worsen a week after the event [31]. Additionally, in radiation therapy, the presence of CMBs often increases after treatment, which correlates with higher doses of radiation [32]

In Alzheimer's disease (AD) and other cognitive disorders, there is a noted association between the number of CMBs and declining cognitive function. The presence of **more than 4 microbleeds** is associated with cognitive decline [5, 33]. Additionally, CMBs are linked with poorer outcomes in conditions such as coronary artery disease (CAD), infective endocarditis (IE), chronic obstructive pulmonary disease (COPD), and chronic renal disease [3].

### 2.2.4 Visual Rating Scales

CMBs are predominantly identified and quantified through visual assessment. Once a potential CMB is detected, accurate classification requires an understanding of pathologic correlations and potential CMB mimics, whose interpretation varies across the possible different co-occurring clinical conditions.

Visual rating scales significantly enhance the consistency of ratings among observers concerning both the presence and anatomical locations of CMBs [17]. Two widely recognized and validated scales are the **Microbleed Anatomic Rating Scale (MARS)** [34] and the **Brain Observer MicroBleed Scale (BOMBS)** [24], detailed in Table 2.1. These scales systematically categorize CMB locations into three primary areas: lobar, deep, and infratentorial regions. The design of these scales is crucial for determining the diagnostic importance of the number and location of CMBs, which is essential for deriving clinical insights from CMB findings.

### 2.2.5 Inter-rater agreement

Before the introduction of standardized visual rating scales, inconsistencies plagued inter-rater agreement on detecting CMBs. Although the inter-rater correlation coefficients for the total count of CMBs were generally

Table 2.1: **CMBs Evaluation Categories According to BOMBS and MARS Rating Scales**. Differences in the categorization of CMBs between BOMBS and MARS rating scales is shown. Table taken from [35]

| Category | BOMBS | MARS |
|---|---|---|
| 1. Certainty/Appearance | (a) certain<br>(b) uncertain | (a) definite<br>(b) possible |
| 2. Size | (a) <5 mm<br>(b) 5–10 mm | -<br>- |
| 3. Side of the Brain | (a) left<br>(b) right | (a) left<br>(b) right |
| 4. Location | **Lobar:**<br>i. cortex/gray-white junction<br>ii. subcortical white matter<br><br><br><br>**Deep:**<br>i. basal ganglia<br>ii. internal and external capsules<br>iii. thalamus<br><br><br><br><br>**Posterior Fossa:**<br>i. brain stem<br>ii. cerebellum | **Lobar:**<br>i. frontal<br>ii. parietal<br>iii. temporal<br>iv. occipital<br>v. insula<br>**Deep:**<br>i. basal ganglia<br>ii. internal capsule<br>iii. external capsule<br>iv. thalamus<br>v. corpus callosum<br>vi. deep and periventricular white matter<br>**Infratentorial:**<br>i. brain stem<br>ii. cerebellum |

strong (around 0.8), the agreement on whether at least one CMB was present showed notable variability in studies utilizing T2S, with Cohen's kappa ($\kappa$) values ranging from 0.33 to 0.88 [8]. Therefore, it is crucial to distinguish between agreement on total counts of CMBs and agreement on whether any CMBs are present at all.

To address these issues, efforts were made to improve inter-rater reliability through the development of rating scales. One such effort was the creation of BOMBS [24], which quickly proved to be useful. In a study involving 264 adults with stroke or transient ischemic attack (TIA), where two doctors used a rating scale to describe the number and distribution of CMBs. The initial inter-rater agreement before using the scale on the presence of ≥1 CMB was moderate overall ($\kappa$ = 0.44; 95% CI, 0.32–0.56), similar in lobar locations ($\kappa$ = 0.44; 95% CI, 0.30–0.58), and higher in deep ($\kappa$ = 0.62; 95% CI, 0.48–0.76) and posterior fossa locations ($\kappa$ = 0.66; 95% CI, 0.47–0.84). Using BOMBS, the agreement improved for any location ($\kappa$ = 0.68; 95% CI, 0.49–0.86) and was notably better in lobar locations ($\kappa$ = 0.78; 95% CI, 0.60–0.97). Another significant effort was the development of the MARS [34]. In a study of 301 patients, the agreement for microbleed presence in any brain location was good to very good, with intrarater reliability at $\kappa$ = 0.85 and interrater reliability at $\kappa$ = 0.68. Excellent reliability was also observed for the number of microbleeds, measured through the intraclass correlation coefficient (ICC), with a value of 0.98 and across different MRI sequences.

Following these developments, agreement levels in studies improved significantly. For instance, Goos et al. (2011) [36] reported excellent interrater agreement with weighted Cohen's kappa values of at least 0.82 for GRE and 0.87 for SWI. Similarly, Shams et al. (2015) [23] found excellent interrater agreement across different MRI sequences, with an intraclass correlation value of 0.897. Another study by Purrucker et al. (2018) [37] reported good interrater agreement for the dichotomization of <5 versus ≥5 CMBs, with a kappa value of 0.74.

Despite these improvements, skepticism remains regarding the true accuracy of these agreement metrics. Kuijf et al. [38] argue that the **current measures used to validate the reliability of microbleed**

**ratings are outdated**. If the interest is confined to the presence or absence of microbleeds, the inter-rater agreement can be adequately assessed using the kappa ($\kappa$) statistic. However, this measure falls short when dealing with multiple microbleeds in an individual subject, as it does not consider the number and location of the microbleeds. This means that raters who agree on the presence of microbleeds in an individual subject might still disagree on their count or distribution. For that reason, more studies have reported the intraclass correlation coefficient (ICC) as a measure for inter-rater agreement. While ICC addresses the number of microbleeds, it still does not account for their location. Two raters might agree on the microbleed count but have counted different microbleeds. Additionally, ICC is data-dependent and can be significantly influenced by outliers. For example, a subject with an exceptionally high number of microbleeds compared to others can skew the ICC, compromising the reliability of the inter-rater agreement. To overcome these issues, Kuijf et al. [38] propose using the Dice similarity coefficient (DSC) as a more reliable measure of inter-rater agreement, suggesting that DSC should be reported in studies rating microbleeds or other pathologies to complement existing measures like $\kappa$ or ICC.

Another limitation is the dependency of the rating scales on specific MRI sequences and field strengths. The MARS and BOMBS scales were developed using T2S at 1.5 T, but their reliability using other sequences, such as SWI or at higher field strengths, is less understood. Puy et al. [20] noted that while these scales provide structured approaches for assessing CMBs, their effectiveness and reliability need further validation across different MRI modalities and settings.

## 2.3 Medical Image Segmentation

**Image segmentation** is the process of dividing an image into separate segments, particularly crucial in medical imaging for delineating distinct anatomical structures or areas of interest. This segmentation supports clinicians in making diagnostic decisions and planning treatments or surgeries. Techniques for segmentation vary widely and fall into six main categories: (a) Thresholding, (b) Region growing, (c) Region merging-splitting, (d) Clustering, (e) Edge detection, and (f) Model-based methods [39]. Certain segmentation methods categorize pixels into segments by grouping those that share similarities (known as the similarity approach), whereas others distinguish segments based on contrasting differences (referred to as the discontinuity approach) [40]. Additionally, segmentation can be further classified into: (a) Semantic segmentation, which labels each pixel into a category without differentiating individual objects within the same class, and (b) Instance segmentation, which not only classifies each pixel into a category but also identifies distinct objects within that category [40].

Medical images present significant challenges in image segmentation due to extensive **class imbalance**. Typically, a medical image contains a single region of interest (ROI) that occupies only a small percentage of pixels, while the rest is annotated as background. This imbalance implies that the model classifier must be trained on data with a rare ROI class and a background class with a prevalence often over 90%. This extreme inequality affects all aspects of a computer vision pipeline for Medical Image Segmentation (MIS), from preprocessing to model architecture, training strategy, and performance evaluation [41].

### 2.3.1 Deep Learning

In recent years, Machine Learning (ML) and its subset, Deep Learning (DL), have significantly impacted various industries [42]. This influence extends to Medical Image Analysis (MIA), where Convolutional Neural Networks (CNNs), a type of DL algorithm, have become the dominant approach [42]. DL in MIA is applied in multiple tasks such as classification, where they categorize images or detect specific entities; detection, which involves identifying anatomical features and abnormalities; and segmentation, aimed at outlining detailed anatomical structures. Beyond these, DL contributes to registration, content-based image retrieval, image generation and enhancement, and the integration of image data with textual reports [43–45].

DL relies on substantial datasets for training, yet the data from medical institutions often fails to meet these requirements. Although healthcare institutions possess extensive medical data, utilizing it for research, especially commercial projects, is challenging. The data may be poorly labeled or irrelevant due to differences in patient groups, demographics, pathologies, or technical factors like varying imaging modalities and sequences. These issues can lead to overfitting, hindering the development of precise models [43]. However,

the community has developed strategies to maximize the use of these imperfect datasets for medical image segmentation, as outlined by Tajbakhsh et al. [46]. In MIA, the integration of professional expertise is crucial and can significantly improve model performance beyond just enhancing the complexity of CNNs [47]. Despite these challenges, the field of DL in medical imaging continues to progress, with ongoing advancements in research and development [44]. For medical image segmentation (MIS), numerous studies confirm that deep learning approaches produce more precise segmentations compared to other methods [48, 49].

#### 2.3.1.1 Fully convolutional networks (FCNs)

In particular, fully convolutional networks (FCNs) have emerged as potent models [50]. FCNs utilize layers that are exclusively locally connected, such as convolution, pooling, and upsampling. The two most critical layers are:

- **Convolutional Layers**: these layers use a weight tensor (kernel) to convolve with the input image, generating a feature map. These layers can vary in shape (nx × ny × #filters) based on the architectural design. Convolutional Layers have three main properties:
  - **Sparse Connections**: by using kernels smaller to image size, they can detect localized features such as edges or corners.
  - **Parameter Sharing**: each individual is applied across the entire image, enhancing learning efficiency by reducing the number of parameters of the model.
  - **Equivariance**: a transformation applied to the input image results in a corresponding transformation in the output feature maps
- **Pooling Layer**: these layers reduce the dimensions of the input filters. By diminishing model parameters, these layers also help in preventing overfitting and increasing tolerance to distortions. Examples include MaxPooling and AveragePooling.

After the input passes through a convolution layer, an activation function—typically ReLU—is applied elementwise to introduce nonlinearity to the feature maps. These activated feature maps are then processed by a pooling layer. This sequence of operations forms what is known as a **convolution block**. By excluding dense layers, FCNs have fewer parameters, which speeds up the training process. Typically, FCNs are trained in a supervised manner by minimizing a loss function that quantifies the discrepancies between the predicted segmentations and the actual ground truth across a set of labeled training images, which in the case of medical segmentation, requires annotation of large 3D images by domain experts.

#### 2.3.1.2 U-Net

A very important FCN architecture is the **U-Net** [51], which has shown an outstanding overall performance in segmenting medical images. At a high level, the U-Net comprises two main sub-networks: an encoder sub-network and a decoder sub-network, complemented by a final pixel-wise classifier. Figure 2.2 illustrates the original U-Net architecture, which consists of the following parts:

- **Contracting Path:**
  - Starts with convolutional layers to extract initial low-level features from images.
  - Employs downsampling, typically through max pooling, to reduce the spatial dimensions of the feature maps while often increasing the depth (number of filters) in subsequent layers.
  - The designs aim to capture different features at every level of depth:
    - Shallow Layers: Capture basic localized features such as edges and textures.
    - Intermediate Layers: Capture more complex features, including shapes and structures.
    - Deep Layers: Capture high-level, abstract features such as the presence of a tumor or other significant patterns.
- **Expansive Path:**

– Uses upsampling to reconstruct the image from the compressed feature maps.
– During upsampling, **skip connections** are used to reintegrate feature maps from the downsampling phase, critical for restoring detailed spatial information lost during pooling and avoiding the problem vanishing gradients.
– Ends with a 1x1 convolution that translates deep feature representations back into segmentation classes.

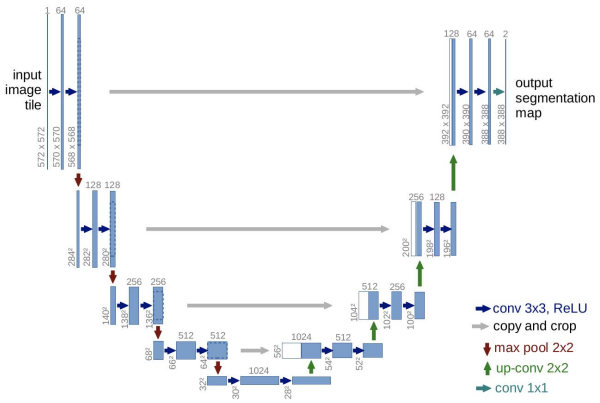



Figure 2.2: **Original U-net architecture**. In this 2D architecture, an example for 32x32 pixels at the lowest resolution is shown. Taken from the original U-Net paper [51].

### 2.3.2 Domain shift

ML methods traditionally operate under the assumption that training and test datasets, or source and target domains, share identical data distributions [52]. However, this assumption is frequently invalidated in real-world applications, particularly leading to the issue known as **domain shift** [53, 54]. Domain shift refers to variations in data distribution between training and testing datasets that can significantly increase test errors.

This challenge is more pronounced in MIA. Inlike in general image processing with extensive labeled datasets like ImageNet, in MIA there is a significant scarcity of labeled medical images. Labeling such images demands intensive efforts and expertise from medical professionals, making the process both costly and time-consuming. Further complicating matters are the inherent variations in medical imaging data due to differences in scanners, scanning parameters, and patient cohorts. These variations exacerbate the domain shift problem, leading very often to models that show excellent performance on available data to fail when applied to different distributions from real-life clinical settings [55]. MRI neuroimaging is highly sensitive to variations in soft tissue appearances and contrasts across different protocols and settings. This sensitivity underscores the need for developing domain-agnostic models, which is crucial given the high costs and limited availability of MRI studies. Yet, currently, there are no large-scale, domain-agnostic solutions for this type of data and most published studies rely on small datasets, often just a few tens or hundreds of images, which are particularly susceptible to domain shift problems [56].

### 2.3.3 Transfer Learning

Transfer learning (TL) is a method widely used in MIA, involving the application of knowledge acquired from solving one problem to solve separately after another distinct problem. According to Pan and Yang (2010) [57], TL is formally defined through the concepts of **domains and tasks**. A domain $D$ consists of a feature space $X$ and a probability distribution $P(X)$ over $X$, while a task $T$ includes a label space $Y$ and a predictive function $f(x) = P(y|x)$ for $x \in X$ and $y \in Y$. Transfer learning typically involves transferring from a source domain and task $(D_S, T_S)$ to a target domain and task $(D_T, T_T)$, where the domains, tasks,

or both may differ. The objective is to improve learning in the target task $T_T$ by leveraging the model $f_S$ learned in the source task $T_S$. The practical implementation of transfer learning can vary greatly, depending on the type of information transferred and how it is utilized in the new learning context $f_T$ [58].

In MIA, the advance of DL has popularised in particular the transfer of information across different tasks and domains [59], which can thus help reduce the domain shift problem. The goal is to develop a robust feature representation, typically by training a deep network on a large and diverse source dataset, which then serves as the basis for further training or fine-tuning to a target task. Based on published results, the effectiveness of this approach appears to depend on a trade-off between the similarity, size and diversity of the source data [59]. Source and target data can be very different, and the use of non-medical images for 2D networks is a common practice. However, opinions are divided as to whether medical or non-medical data sources are more effective for TL, and it appears that the correct answer is very context-dependent [60].

In the specific case of MIS, TL has been demonstrated to reduce training time significantly. However, **the tangible benefits for segmentation accuracy of TL are highly task- and data-dependent** [58]. In particular, significant improvements in accuracy are only observed for more complex segmentation tasks where the available target training data is limited. Furthermore, addressing the challenge of domain shift in MRI data, Ghafoorian et al. [55] suggest that optimal results are achieved by fine-tuning only the last dense layers when few target training cases are available, to avoid overfitting due to a large number of parameters relative to the training sample size. As more training data become available, it becomes advantageous to fine-tune shallower representations.

### 2.3.4 Data Augmentation

When training a DL model, the primary dataset is divided into training and test sets. The training set is utilized to optimize model parameters against a loss function, and then the test dataset is used to assess the final model's performance. If data used is not enough, overfitting can occur, a problem that arises when a model trained on a limited dataset fails to generalize to new data. To mitigate this, data augmentation is employed to expand and diversify the training set, serving as a regularization strategy to lower the model's generalization error and improve robustness [61]. Data augmentation is used to increase the diversity of data available for training models without actually collecting new data. It involves generating new data instances or modifying existing data instances to create a larger training set. According to a review by Chlap et al. [62], the primary data augmentation techniques for deep learning in medical imaging include:

- **Basic Augmentations:** This category involves simple transformations such as repositioning or altering the intensity of image pixels. A single image is modified and then returned to the dataset to increase its volume. Basic augmentations are not necessarily aimed at creating realistic images but are easy to apply and can substantially boost the efficacy of the trained model . Common examples include geometric transformations, cropping, noise injection, filtering, and combinations thereof.
- **Deformable Augmentations:** These techniques are constrained by user-defined parameters to maintain clinical plausibility. They replicate realistic variations such as tissue deformations and motion artifacts commonly found in clinical scans. Examples include randomized displacement fields, spline interpolation, deformable image registration, and statistical shape models.
- **Deep Learning-based Augmentations:** Utilizes generative models, such as Generative Adversarial Networks (GANs), to synthesize realistic images

# 3 Related Work

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. [35], outlining 67 past approaches and intending to establish guidelines. Jiang et al. [63] conducted an extensive review of automated methods for CSVD biomarkers, identifying 36 relevant papers for CMB automatic detection up to November 2021. Then, Matsoukas et al. [64] focused on the performance of AI systems in detecting Intracerebral Hemorrhage (ICH) and CMBs, reviewing 18 ML-based approach papers until early 2021. They all recommend strategies like transfer learning, 3D spatial information integration, and diverse dataset utilization for robustness. Other articles or books have also reviewed vaguely some approaches to the task [65], many times in relation to a specific pathology like cerebral small vessel disease (CSVD) [66], [67] and TBI or DAI [68], [69].

Previous approaches for automated CMB detection typically involve a two-stage process: initial detection of CMB candidates followed by post-processing to eliminate false positives (FPs). In the detection stage, deep learning strategies include creating custom neural networks using architectures like Artificial Neural Networks (ANNs), Backpropagation Neural Networks (BPNNs), Stacked AutoEncoders (SAEs), and Convolutional Neural Networks (CNNs), with some studies also using Extreme Learning Machines (ELMs) for efficiency. Alternatively, transfer learning has been used to employ pre-trained models such as AlexNet, ResNet50, and U-Net. For CMB verification, methods use custom algorithms (e.g. based on predefined CMB features), classical ML classifiers (such as Support Vector Machines (SVM), Linear Discriminant Classification (LDC), Quadratic Discriminant Classification (QDC), Parzen window classifiers, and Random Forest Classifiers (RFC)). Many times post-processing thorugh brain masks and morphological operations are used to further enhance predictions.

## 3.1 Most Recent Approaches

Here, we summarize the most recent methods for automatic CMB detection not included in the comprehensive review by Ferlin et al. [35], covering developments from June 2023 to 2024:

- Ali et al. [70] introduced a streamlined approach by employing a variant of the U-Net model within an Internet of Medical Things (IoMT) framework. With an end-to-end design, they eliminate the need for any pre-processing or post-processing steps.
- Ferrer et al. [71] proposed a 3D deep learning framework that leverages multi-domain data. They adapted a 3D structure from 2D CNN, utilizing MultiResUNet models trained on three orthogonal views of MRI scans. This method involves reshaping volumes into a standardized cube and using thickened 2D slices as input, enriching the model with 3D spatial information.
- Sundaresan et al. [72] adopted a multi-stage detection strategy that begins with initial candidate detection to ensure high sensitivity, followed by a candidate discrimination step using a knowledge distillation framework. A multi-tasking teacher network guides a student network, and a subsequent morphological cleanup step reduces false positives by employing anatomical constraints.
- Wu et al. [73] focused on a deep learning model tailored for SWI sequences to detect CMBs and classify them into having CSVD or not, aiming to assist neurologists in optimizing CSVD management. Their model integrates a Mask R-CNN for detection and a multi-instance learning network for classification.
- Fang et al. [74] introduced a novel architecture that moves away from local patches-based approaches. Their method leverages whole-brain distribution information as prior knowledge and employs a 2.5D convolutional network that assesses morphological differences, effectively balancing computational efficiency and spatial information retention.

- Jun et al. [75] proposed a 3D deep learning framework that detects CMBs and also provides their anatomical locations within the brain. They utilized a U-Net with a Region Proposal Network and integrated a Feature Fusion Module and Hard Sample Prototype Learning to reduce false positives. This method utilizes both SWI and phase images to capture detailed 3D information.
- Xia et al. [76] developed a two-stage pipeline for identifying CMBs in QSM images. Initially, a 2.5D fast radial symmetry transform algorithm coupled with a convolutional network pinpoints candidate regions, followed by a V-Net that distinguishes between true CMBs and mimics, thereby reducing false positives.
- The most recent development by Xia et al. [77] in 2024 addresses the challenges of anisotropy by implementing an encoding-enhanced network. This novel method transfers knowledge from high-resolution models to models working lower resolution scans

# 4 Datasets

Our study incorporates **seven datasets from five distinct sources**. The proprietary datasets ***CRB*** and ***CRBneg*** at CEREBRIU include 18 and 742 scans respectively, offering a broad spectrum of clinically representative data with T2S and SWI sequences. ***MOMENI*** is a publicly available dataset featuring 370 scans from 118 patients, and a derivative dataset ***sMOMENI*** contains synthetically augmented CMBs with ten times the number of scans. ***VALDO***, sourced from a public challenge, includes 72 T2S scans from three different cohorts. ***DOU*** contains 20 SWI images publicly available. Finally, ***RODEJA*** is a private dataset comprising 103 SWI scans and patients. We want to highlight that, despite the assertion in [35] that 20 CMB datasets are available upon request, no author was willing or able to share their data with us following our initial contact.

## 4.1 CRB

The *CRB* dataset comprises **18 patients and 18 scans —11 T2S and 7 SWI— with a total of 127 CMBs** weakly annotated. We created this dataset internally at CEREBRIU according to 4.1.1 below. The data is sourced from various data providers and originates from hospitals across three continents outside Europe (Brazil, India, USA). Owing to confidentiality regulations, the data is anonymized, and no patient-level metadata is available. The creation of this dataset involved pixel-level annotations of CMB conducted as part of this project. These annotations were done on a pre-selected number of studies that had been case-level annotated previously outside the scope of this project.

### 4.1.1 Selection of studies

We carefully curated our selection from a collection of internally annotated case-level studies, which was originally designed to generate training data for DL models to detect infarcts, tumors, and macro-hemorrhages. We then selected studies that exhibited CMB as additional findings. These selected cases were then incorporated into our annotation process, facilitated by our experienced internal rater, SI, to enrich the dataset as time permitted.

This dataset is designed to reflect diverse clinical settings, with variations in slice thickness (6 to 12 mm), magnetic field strengths (1.5-3T), TE (19-48ms), TR (27-936ms), flip angles (12-20), and multiple sequence types. It includes patients with co-occurring pathologies — four with various types of infarcts and five with hemorrhages. Specifically, two patients have intra-axial acute hemorrhages, one has chronic infarcts, and another exhibits bleeding sequelae. Additionally, one patient has both acute SDH/EDH and bleeding sequelae, another has a subacute infarct, one shows both hyperacute/acute and chronic infarcts, and another has an intra-axial chronic hemorrhage.

### 4.1.2 Image-level Annotation

This section outlines the case-level annotations **conducted outside of this project**, that we re-annotated at the pixel level. Each case in the dataset includes a radiological report provided by a radiologist who did not participate in the image-level annotations. These radiological reports offer detailed insights into findings across various MRI sequences combined with patient history. The annotation process included the following steps:

- **Main/Additional Finding Identification**: Image-level annotations were performed to identify various subtypes of infarcts, tumors, and macro-hemorrhages, along with other findings, including the 'Microhemorrhages <10mm' category.

- **Quality Control**: A thorough quality assessment ensured that all images met the required standards for radiological examination.
- **Exclusion Criteria**: In adherence to CEREBRIU's internal requirements, specific exclusion criteria were applied, disqualifying individuals under 18 or over 90, those with a history of neurosurgery, non-brain cancer, demyelinating or inflammatory diseases, or inadequate MRI sequence quality.

The annotation process also involved double readings and final arbitration to resolve any discrepancies, carried out by two certified radiologists. These annotations helped us pre-select the CMB-containing studies, and also provide extra metadata to enrich our dataset.

### 4.1.3 Voxel-level Annotation

To obtain CMB segment maps of our data, we need to localize in the 3D volume where each cMB is. Thus, **voxel-level annotations were meticulously designed and performed as part of this project**. The annotations protocol is as follows:

- **Rating Scale**: The annotation process adhered to the visual rating scale BOMBs [24].
- **Scribble Annotations**: weak annotations were recorded in the form of scribbles, capturing the vague outlines of individual CMBs. Each annotated CMB scribble was stored on a separate 3D segmentation map. In cases flagged as 'multiple', an isolation post-processing step was required to differentiate different CMBs within each scribble, otherwise each scribble mapped to one CMB.
- **Metadata Collection**: For each CMB, metadata detailing certainty (certain or uncertain), potential causes (e.g., trauma, hypertension, DAI, trauma-DAI, CAA, radiotherapy, or 'Other' with free text), size (<5mm or larger), and precise anatomical location (e.g., cortex/grey-white junction, basal ganglia grey matter, subcortical white matter, internal and external capsule, thalamus, brainstem, cerebellum) was compiled.
- **Mimics Documentation**: The presence of mimics was also recorded to allow for a comprehensive analysis. This included placing landmarks on the images to indicate the approximate locations of the mimics and documenting specific metadata for each type. Common mimics included cavernomas, teleangiectasis, basal ganglia calcifications, or 'Other' with free text.
- **Recording Additional Findings**: Similarly, the process was employed to document any other relevant additional findings as free text.

The annotations were conducted by an experienced rater (SI), who not only has expertise in annotating microbleeds and is one of the annotators of the VALDO challenge dataset [13] — but also has conducted scientific research on CMBs. The annotator's extensive experience in CMB annotation enables us to treat these annotations as ground truth with sufficient confidence. All these detailed annotations are implemented into an annotation project using the RedBrick AI platform, which supports robust data collection and documentation strategies tailored for this complex task (see Figure 4.1 for an annotation example). The resulting scribbles are scattered data points in space and need further post-processing to be used as CMB segmentation masks.

## 4.2 CRBneg

The *CRBneg* dataset consists of **742 patients with 742 scans —541 T2S and 201 SWI— with no CMBs** with a high likelihood. This dataset was derived from the same extensive pool of case-level annotated cases as the *CRB* dataset. However, the selection criteria were the inverse, focusing exclusively on cases where 'Microhemorrhages <10mm' were not indicated in the additional findings. Further verification involved a thorough review of radiological reports, specifically seeking to confirm the absence of CMBs by excluding any mention of the following keywords: 'microbleeds', 'micro-bleeds', 'micro bleeds', 'microhemorrhages', 'micro-hemorrhages', 'micro hemorrhages', and 'microbleedopathy'. Equally, both quality and inclusion criteria

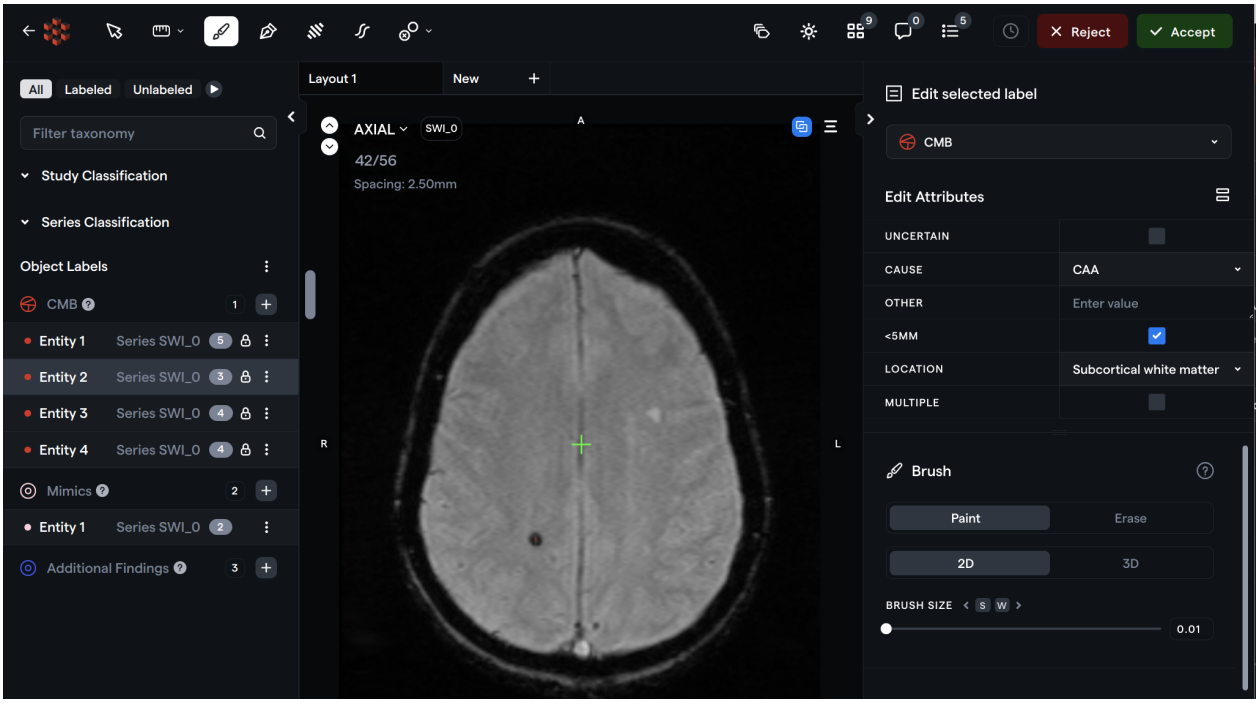



Figure 4.1: **Example of annotation on Redbrick**. For a randomly selected scan, an axial slice of the SWI sequence is shown for CMB finding number 2. It can be seen that the guess on possible cause is CAA, that it has less than 5mm diameter, and that the location of CMB is in subcortical white matter. 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 a mimic annotated somewhere in the image as a landmark, even if now shown on this slice.

were enforced in data selection. This dataset provides an even more representative sample of real-life clinical data than *CRB*, also featuring a diverse range of MRI acquisition parameters and demographics. Notably, it contains a substantially higher prevalence of pathological conditions, including different types of infarcts, hemorrhages, and tumors. A detailed list of these pathologies is available in the Appendix for reference.

## 4.3 DOU

The *DOU* dataset includes **20 patients and features 20 SWI scans, with a total of 74 CMBs** single-point annotated at their centers. This dataset was made publicly available in [11]. SWI images are acquired from a 3T Philips Medical System at the Prince of Wales Hospital in Hong Kong. The images were captured using a 3D spoiled gradient-echo sequence with a venous blood oxygen level-dependent series. The acquisition parameters were as follows: repetition time of 17 ms, echo time of 24 ms, volume size of $512\times512\times150$, in-plane resolution of $0.45\times0.45$ mm, slice thickness of 2 mm, slice spacing of 1 mm, and a field of view of $230\times230$ mm$^2$. Subjects are either patients with stroke or individuals experiencing normal aging. Authors were contacted for more details about the exact proportions and demographics, but no extra information could be provided due to information loss in the anonymization process. The dataset was labeled by an experienced rater and verified by a neurologist using the MARS rating scale. For more specific details, refer to the original paper. CMB masks are not available, but rather, coordinates in 3D space indicating where CMB are located within the SWI image.

## 4.4 MOMENI

The *MOMENI* dataset includes **118 patients and 370 SWI scans, with a total of 146 CMBs** single-point annotated at their centers. A total of 313 scans contain no CMBs. This dataset was made publicly available in [12] and includes data from the Australian Imaging Biomarkers & Lifestyle (AIBL) study [78]. SWI scans performed on a 3.0T Siemens TRIM TRIO scanner were acquired with an in-plane resolution of 0.93x0.93mm, slice thickness of 1.75mm, repetition time/echo time of 27/20ms, and flip angle of 20°, and were reconstructed online using the scanner's VB17 software. The N4 bias field correction was applied [79]. Subjects present can belong to one of the following categories: Alzheimer's disease (AD), mild cognitive impairment (MCI), and cognitively normal. It was however not possible to know each study's specific category. SWI images were manually inspected for CMBs by two clinical experts using the MARS rating scale. There are more scans than unique patients, as one patient can have several scans acquired over time (sometimes changing its condition from healthy to unhealthy and vice-versa, will be discussed later on). CMB masks are not available, but rather, a coordinate indicating where the center of the CMB is located in the SWI image.

## 4.5 sMOMENI

The *sMOMENI* dataset is derived from *MOMENI* and includes **118 patients and 3700 SWI scans, with a total of 36812 CMBs** single-point annotated at their centers. This dataset is an augmented dataset generated from MOMENI, thus sharing acquisition, demographic and scanner characteristics described in section 4.4. Authors in [12] used a custom deterministic model to generate synthetic CMBs based on parameters observed from real CMB: shape, location, intensity, and volume. It uses a 3D Gaussian function to generate the synthetic CMBs, which are modeled to match the shape and orientation of real lesions, produced within high-resolution patches (70×70×70 pixels) that are then processed to differentiate the lesion from surrounding tissue. The volumes of these synthetic CMBs are derived from a smoothed distribution of actual CMB volumes, and their placement within the brain is strategically randomized to reflect the typical distribution and proximity to various brain tissues observed in real cases.

This dataset is also publicly available. For every scan present in MOMENI, they generated 10 new scans, each enriched with 10 extra synthetic microbleeds. Thus, all scans in this dataset contain exactly 10 microbleeds (when originated from healthy scans) or more than 10 (when originated from unhealthy ones). As in MOMENI, only the locations of the centers of the CMBs are provided.

## 4.6 RODEJA

The *RODEJA* dataset includes **103 patients and 103 SWI scans, with a total of 357 CMBs** with full voxel-level annotations. A total of 42 scans contain no CMBs. The majority of scans of this dataset were used in [71]. While it is not publicly accessible, we obtained permission to use it from the data owners at the University of Copenhagen (KU). Raw data obtained consisted of SWI sequences along with CMB masks annotated by clinical experts. The scans were acquired from multiple hospitals in the capital region of Denmark using scanners of varying magnetic strengths (1.5 T and 3 T) and resolutions ranging from 0.2x0.2x1mm$^3$ to 1x1x6mm$^3$. No scan-level or patient-level metadata was provided for this data, but as they claim in [71] it is supposed to be very diverse in terms of scanners, resolutions and sites.

## 4.7 VALDO

The *VALDO* dataset includes **72 patients and 72 T2S scans, with a total of 253 CMBs** with full voxel-level annotations. A total of 22 scans contain no CMBs. This dataset consists of the training data made available for the MICCAI Valdo Challenge in 2021[1]. From [13], we know that three subsets of population cohorts were used, which were part of retrospective studies. The SABRE cohort, a tri-ethnic group from West London, UK, was initially recruited to explore metabolic and cardiovascular diseases across ethnicities

[1] https://valdo.grand-challenge.org/

using a Philips 3T scanner, with participants aged 72 years on average, ranging from 36 to 92 [80]. The RSS focuses on chronic illnesses in the elderly using a dedicated 1.5T GE MRI scanner for individuals over 45 without dementia [81]. Lastly, the ALFA cohort, drawing from the ALFA registry of Alzheimer's Disease patients' relatives, examines a genetically predisposed group of cognitively normal individuals aged 45-74, using GE Discovery 3T MRI technology [82].

Acquisition was performed by trained radiographers according to a predefined research protocol. Different raters annotated each of the cohorts but followed very similar protocols, and had at least 3 years of professional experience in dealing with medical images. The BOMBS criteria [24] were applied for the SABRE and ALFA cohort as described in [1]. A team of trained raters under the supervision of MWV applied the protocol described in [83] for RSS. Both identification protocols are in line with the STRIVE guidelines [84], indicating that microbleeds are areas of signal void generally 2-5 mm in diameter but can be up to 10 mm. T2s scans along with 3D CMB annotations were available. We will refer to this dataset as *VALDO*.

## 4.8 Demographics and scan parameters

DL applications in MRI medical imaging face significant challenges due to data variation and domain shifts, which hinder model generalization across different clinical scenarios, imaging conditions, and demographics. These variations significantly impact model performance, and need to be well understood for any DL application [85].

Differences in age, ethnic groups, and other biological factors can substantially alter the appearance of medical images. This complexity intensifies in clinical contexts, where pathologies further contribute to variations. MRI, known for its variability, does not display uniform contrast or intensity ranges across scans due to numerous acquisition parameters. What is more, the definition of MRI sequences can vary significantly depending on the hardware model of the scanner, and even within the same MRI sequence type, clinicians often modify its appearance by tuning acquisition parameters.

In order to gain a deeper understanding of the factors of variation present in our datasets, we have carefully compiled them in Table 4.1. As can be seen, the range of datasets selected for this study exhibits substantial diversity. Geographically, they originate from 9 countries in 6 different continents. Demographically, they cover a wide spectrum of ages and pathological conditions, with both sexes equally represented. Moreover, the datasets exhibit variation in terms of scanner models across different vendors (GR, Siemens, and Phillips), sequence types utilized for CMB detection, and specific MRI acquisition parameters that influence the appearance of tissues in MRI sequences, and thus the 3D structures that will be seen by the model. Finally, both existing CMB rating scales are also represented.

## 4.9 Analysis of CMBs

This section provides further information on the CMBs present in all our datasets. Firstly, it describes how they are distributed across scans and datasets. Secondly, it outlines the characteristics of individual CMBs. CMBs are uniform in terms of intensity, that is, they are signal voids in the MRI image with very low-intensity values compared to the surrounding area. Therefore, no special attention has been paid to specific intensity variation within the CMBs. Rather than focusing on intensity, size, shape, and location were the primary characteristics examined. All features computed in this section were applied to pre-processed data.

### 4.9.1 Size and counts

The upper row in Figure 4.2 displays the number of CMBs per scan (excluding healthy scans) and estimates their radii by assuming each CMB forms a perfect 3D sphere, with a volume equivalent to the voxel space occupied by the CMB mask. This method of radius estimation is quite simplistic, presuming spherical symmetry of CMB masks which is often not the case. CMBs may undergo various deformations during preprocessing that alter their shape, even in datasets where spheres are artificially created around the CMB center. Additionally, this assumption could be flawed even with human annotations, as it assumes that annotators perfectly captured every voxel within the spherical intensity void. Our observations indicate that CMBs often do not appear fully spherical, particularly in scans with poor contrast and resolution. Despite its

Table 4.1: **Summary of Scanner specifications, Acquisition Parameters, Demographics, and rating scales used for all datasets**. Note that some datasets have been subdivided into subgroups for better cohort granularity. sMOMENI is not presented here as its characteristics are exactly the same as in MOMENI (data is derived from it, see Section 4.5)

| **Dataset** | **Location** | **Demographics** | **Scanner Model** | **MRI Seq.** | **Field (T)** | **TR (ms)** | **TE (ms)** | **Flip $\alpha$** | **Slice Thick. (mm)** | **Rating Scale** |
|---|---|---|---|---|---|---|---|---|---|---|
| CRB | Brazil | Ages between 40-57, mean 48, mostly female | GE: Optima MR450w, Genesis Signa | T2S | 1.5 | 75-317 | 23-48 | 15-20 | 4.6-5.5 | BOMBS |
| CRB | India | Ages between 25-80, mean 59, 31% female | Siemens: Spectra, Magnetom Sempra, Sempra, Symphony | T2S, SWI | 1.5/3 | 48-936 | 19-40 | 12-20 | 2.5-5 | BOMBS |
| CRB | U.S.A | Ages between 73-90, mean 79, 33% female | Siemens: Magnetom Vida | SWI | 3 | 27 | 20 | 15 | 1.9 | BOMBS |
| CRBneg | Brazil | Ages between 18-90, mean 50, 53% female | GE: Brivo MR355, Optima MR450w, Genesis Signa | T2S | 1.5 | 74-717 | 19-50 | 15-20 | 3-6 | BOMBS |
| CRBneg | India | Ages between 18-87, mean 52, 45% female | Siemens: Spectra, Magnetom Essenza, Sempra, Symphony | T2S, SWI | 1.5/3 | 48-968 | 19-40 | 12-20 | 2-5 | BOMBS |
| CRBneg | U.S.A | Ages between 18-90, mean 55, 54% female | Siemens: Magnetom Vida, Prismafit, Magnetom Sola, TrioTim | SWI, T2S | 1.5/3 | 27-782 | 20-40 | 15-20 | 1.75-4 | BOMBS |
| DOU | China | Stroke and normal aging | Philips Medical System | SWI | 3 | 17 | 24 | 20 | 2 | MARS |
| MOMENI | Australia | Alzheimer's disease, mild cognitive impairment and cognitively normal | Siemens: TrioTim | SWI | 3 | 27 | 20 | 20 | 1.75 | MARS |
| RODEJA | Denmark | - | - | SWI | 1.5/3 | - | - | - | - | - |
| VALDO | UK | SABRE: Tri-ethnic, high cardiovascular risk, 36-92 years old, mean age: 72 | Philips | T2S | 3 | 1288 | 21 | 18 | 3 | BOMBS |
| VALDO | Netherlands | RSS: Aging population $>45$ without dementia | GE: MRI | T2S | 1.5 | 45 | 31 | 13 | 0.8 | [83] |
| VALDO | Spain | ALFA: Enriched for APOE4, family risk of Alzheimer's. Cognitively normal participants aged 45-74 | GE: Discovery | T2S | 3 | 1300 | 23 | 15 | 3 | BOMBS |

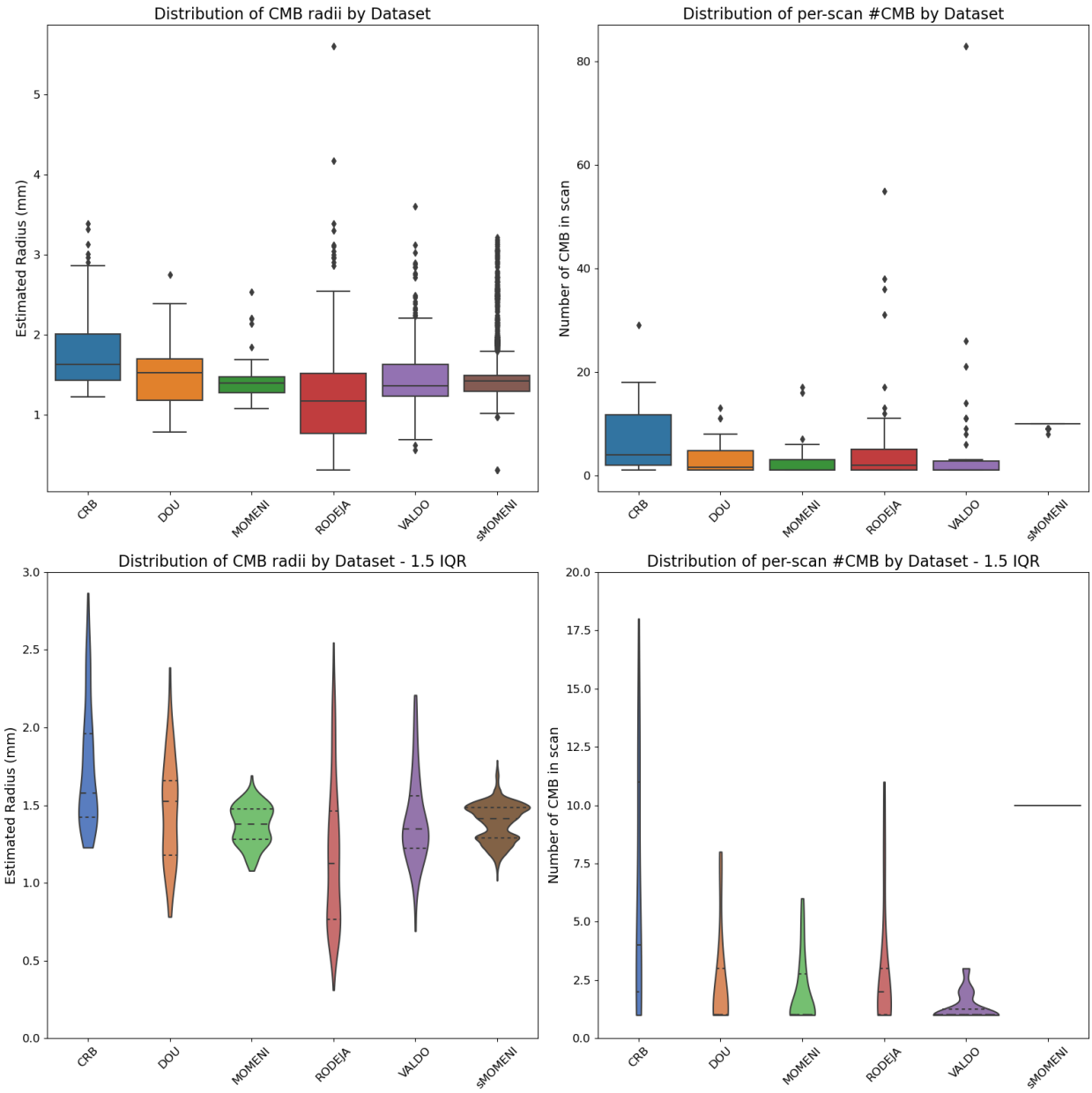



Figure 4.2: **Distribution of CMB Radii and Counts Across Datasets** *Upper Panels:* Boxplots displaying the distribution of CMB radii and counts per scan across datasets for all scans. Radii are estimated by modelling individual CMB masks as perfect spheres wit equal volume. *Lower Panels:* Violin plots showing the same distributions after filtering outliers beyond 1.5 times the interquartile range (IQR). For boxplots, the central line represents the median; the box's lower and upper edges mark the lower and upper quartiles, respectively; whiskers extend to the highest and lowest values within 1.5 times the IQR; datapoints outside of this range are considered outliers and plotted as separate markers. In violin plots, each violin shape represents the density estimation of the underlying distribution using a kernel density estimate (KDS), with the thick dashed lines in the center representing the median of the data, and the thin dash lines indicating the first and third quartiles. *CRBneg* is not presented as it contains no microbleeds at all. These figures are obtained from the pre-processed data.

limitations, this approach still provides valuable information about the size of microbleeds, which is essential for differentiating them from macrohemorrhages. A more sophisticated analysis of shapes and alternative radius estimations are further discussed in Section 4.9.2.

By examining the radii data, we observe for all datasets that all CMBs have a radius less than 5mm, which aligns with the definition that microbleeds should be less than 10mm in diameter. The only exception is the RODEJA dataset, which includes a very extreme outlier that could be considered a macro-hemorrhage (and was actually learned to be ignored by all models trained). More concerning is the observation that many microbleeds in the MOMENI dataset have an estimated radius below 1mm, whereas the minimum clinical size for microbleeds is 2mm in diameter. This discrepancy raises doubts about the quality of annotations or potentially about the adequacy of mask expansion during preprocessing, or possibly both.

When examining the number of CMBs present per scan, we observe that most cases contain between 1 and 5 CMBs, except in the CRB dataset, where the range is between 1-10. RODEJA and VALDO datasets show a higher number of scans with a high number of CMBs, which might be attributed to the inclusion of a more aged population, although this cannot be definitively stated for RODEJA due to the lack of demographic information. Notably, the boxplot for sMOMENI appears collapsed because most of the data is derived from healthy scans with exactly 10 synthetic CMBs added. Furthermore, when filtering values within the 1.5 IQR, we observe that the CRB dataset exhibits the most variance in both the number of CMBs per scan and the size of CMBs. Understanding these characteristics can aid in assessing whether datasets are suitable for training models or for evaluating their capacity to differentiate between groups containing few versus multiple CMBs, which is clinically significant.

### 4.9.2 Shape features

Many early approaches to automatic CMB detection exploited the spherical shape features of CMBs to enhance classifier performance or to differentiate them from more cylindrical structures like veins [10,86]. In line with these, we computed several shape features of the microbleeds using pyradiomics package

One key feature calculated is **sphericity**, which measures the roundness of the shape of the CMB region relative to a sphere. Sphericity is a dimensionless measure, independent of scale and orientation. The value of sphericity ranges from 0 to 1, where a value of 1 indicates a perfect sphere—a sphere has the smallest possible surface area for a given volume compared to other solids. The formula for sphericity is given by:

$$\text{Sphericity} = \frac{\pi^{\frac{1}{3}} \times (6 \times \text{Volume})^{\frac{2}{3}}}{\text{Surface Area}} \tag{4.1}$$

Additionally, we computed the **Maximum 3D diameter**, defined as the largest pairwise Euclidean distance between the CMB surface mesh vertices (points in the surface), also known as Feret Diameter. Additionally, we computed the **Flatness**, which shows the relationship between the largest and smallest principal components in the CMB shape. We compute the inverse of the true flatness as:

$$\text{Inverse Flatness} = \frac{\lambda_{\text{least}}}{\lambda_{\text{major}}} \tag{4.2}$$

where $\lambda_{\text{least}}$ and $\lambda_{\text{major}}$ are the lengths of the largest and smallest principal component axes, respectively. The values range from 1, indicating a non-flat, sphere-like shape, to 0, which represents a flat object or a segmentation confined to a single slice. Finally, we examined **Elongation**, which illustrates the relationship between the two largest principal components in the ROI shape. We compute the Inverse Elongation:

$$\text{Inverse Elongation} = \frac{\lambda_{\text{minor}}}{\lambda_{\text{major}}} \tag{4.3}$$

where $\lambda_{\text{major}}$ and $\lambda_{\text{minor}}$ are the lengths of the largest and second largest principal component axes, respectively. The values of the inverse elongation range between 1 (indicating a cross-section through the first and second largest principal moments, so circle-like and non-elongated) and 0 (indicating maximal elongation: i.e., a 1-dimensional line). The principal component analysis is performed using the physical coordinates of the voxel centers defining the CMB. The distribution of shape feature values across datasets is illustrated in

Figure 4.3, and for more information on metrics computation please visit full documentation in pyradiomics - Shape Features (3D).

Notably, datasets subjected to sphere creation processes (detailed in Chapter 5) exhibit increased sphericity, with the exception of the CRB dataset. This anomaly in the CRB dataset may be attributed to its lower resolution and slice thickness, which potentially distort the spherical shape upon resampling. The RODEJA and VALDO datasets, despite featuring full 3D annotations by clinical experts, did not display markedly spherical CMBs. This suggests that CMBs might not always appear perfectly spherical, possibly due to the limitations in resolution that prevent a clear spherical representation in MRI signals. It should be noted that the maximum observed 3D diameter does not match the previously computed spherical radius approximations, indicating the presence of elongated shapes. Diameters exceeding 10mm were observed in all VALDO, RODEJA, and CRB datasets.

Regarding elongation, the CMBs in the CRB and RODEJA datasets span the entire 0-1 range, indicating a uniform representation across all levels of elongation, with VALDO showing a very close behaviour for a shorter less alongated range. This broad distribution poses a challenge, as cortical vessels—common mimics of CMBs—are typically differentiated by their tubular shapes. If CMBs in the ground truth data also exhibit elongated forms, it may complicate the model's ability to distinguish them from such mimics. In contrast, the MOMENI and DOU datasets display almost no elongation, with DOU in particular having the least elongated CMBs. Observations regarding flatness are consistent with those on elongation. Overall, the metrics of sphericity, elongation, and flatness appear to correlate, as evidenced by similar patterns across datasets in the violin plots for these three metrics

### 4.9.3 Location in the brain

To better understand the distribution of microbleeds across brain locations, we utilized SynthSeg [87], a robust deep-learning tool optimized for the automated segmentation of MRI brain scans across any contrast and resolution. SynthSeg assigns anatomical labels to each voxel, detailing specific cortical and subcortical structures with a total of 33 distinct labels (see full list here). We selected SynthSeg for several reasons. Firstly, its domain-agnostic training allows it to be directly applied to our T2S and SWI scans indistinctly. Additionally, it is designed to handle the imaging variations commonly present in clinical data, such as scans with low signal-to-noise ratios, varying resolutions, and poor tissue contrast; which is the case in our clinically diverse set of data. Finally, SynthSeg operates "out-of-the-box" with minimal setup required, minimal post-processing (resampling needed), and no re-training/fine-tuning required.

| **BOMBS Anatomical Region** | **BOMBS High-Level Location** | **SynthSeg Labels mapped** |
|---|---|---|
| Cortex / Grey-White Junction | Lobar | Cerebral Cortex, Hippocampus, Amygdala |
| Subcortical White Matter | Lobar | Cerebral White Matter |
| Basal Ganglia Grey Matter | Deep | Caudate, Putamen, Pallidum, Accumbens Area |
| Thalamus | Deep | Thalamus |
| Brainstem | Deep | Brainstem |
| Cerebellum | Deep | Cerebellum Cortex, Cerebellum White Matter |

Table 4.2: **Mapping of SynthSeg labels to BOMBS anatomical locations**. Note that the BOMBS anatomical location 'internal and external capsule' is not represented by SynthSeg labels, and that SynthSeg has different labels for left and right versions of regions, which are shown here as one. Also note that cerebellum and brainstem have been grouped together with 'deep' to simplify the analysis, although they should belong to 'posterior fossa'.

We mapped every label to broader anatomical areas as defined in the BOMBS rating scale, as detailed in Table 4.2. SynthSeg was applied across all our data sets. For each cerebral CMB in every scan, we quantified the number of voxels overlapping with each SynthSeg outputted label. Surprisingly, we found significant overlaps with non-brain regions such as background, cerebrospinal fluid (CSF), ventricles, and the ventral

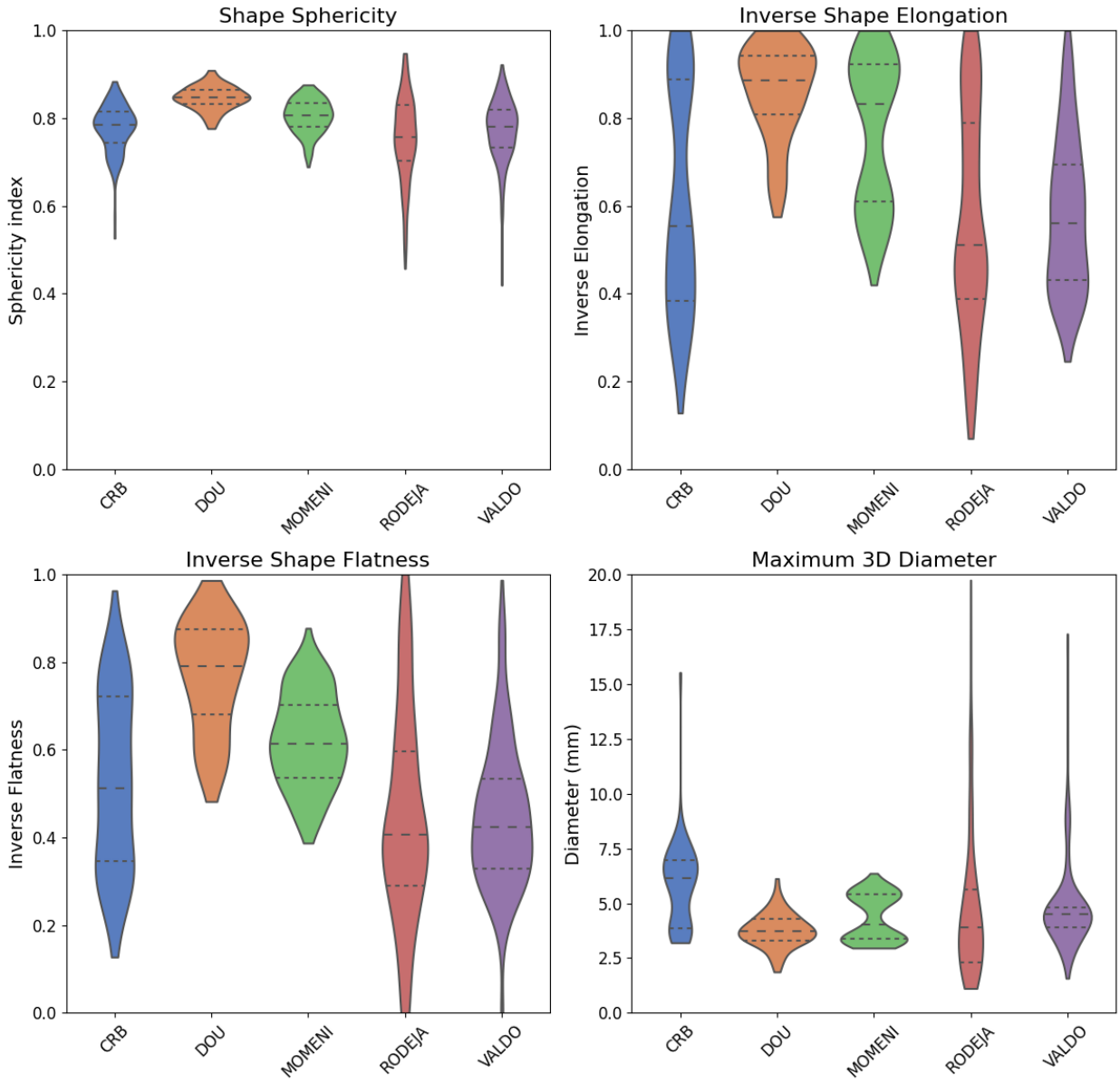



Figure 4.3: **CMB Shape Characteristics by Dataset**. Displays the variation in shape features of cerebral microbleeds across different datasets in the form of violin plots *Note:* The *CRBneg* dataset is excluded as it contains no microbleeds

diencephalon, which is not possible because CMBs cannot be present outside of brain regions according to the BOMBS rating scales. This indicates that the predictions made by SynthSeg may be erroneous, particularly at the boundaries between different anatomical regions, based on our observations. Thus, we used the tool with caution. To assign a location to each CMB in our ground truth (GT) annotations, we followed this process: each CMB volume was assigned the SynthSeg label with the greatest overlap, ignoring non-brain labels when brain labels were also present. Specifically, in cases where the predictions were particularly flawed between cortex and white matter labels, and both were present as majority labels, we opted for white matter after manual examination by an experienced CMB rater confirmed this approach (SI). Subsequently, we grouped the labels into BOMBS anatomical regions, which allowed us to categorize CMBs into relevant regions for analysis, emulating the process a radiologist might use when annotating CMBs according to BOMBS criteria (in the CRB dataset we have this metadata from SI annotations). Figure 4.4 displays the counts of each location group for all CMBs across the datasets. The majority of CMBs in all datasets are found in the cortex/white matter junction, followed by subcortical white matter, with other locations evenly distributed in most datasets. This is consistent with the literature, as the prevalence of microbleeds has been found to be approximately two thirds for lobar CMBs and one third for deep and posterior fossa [88].

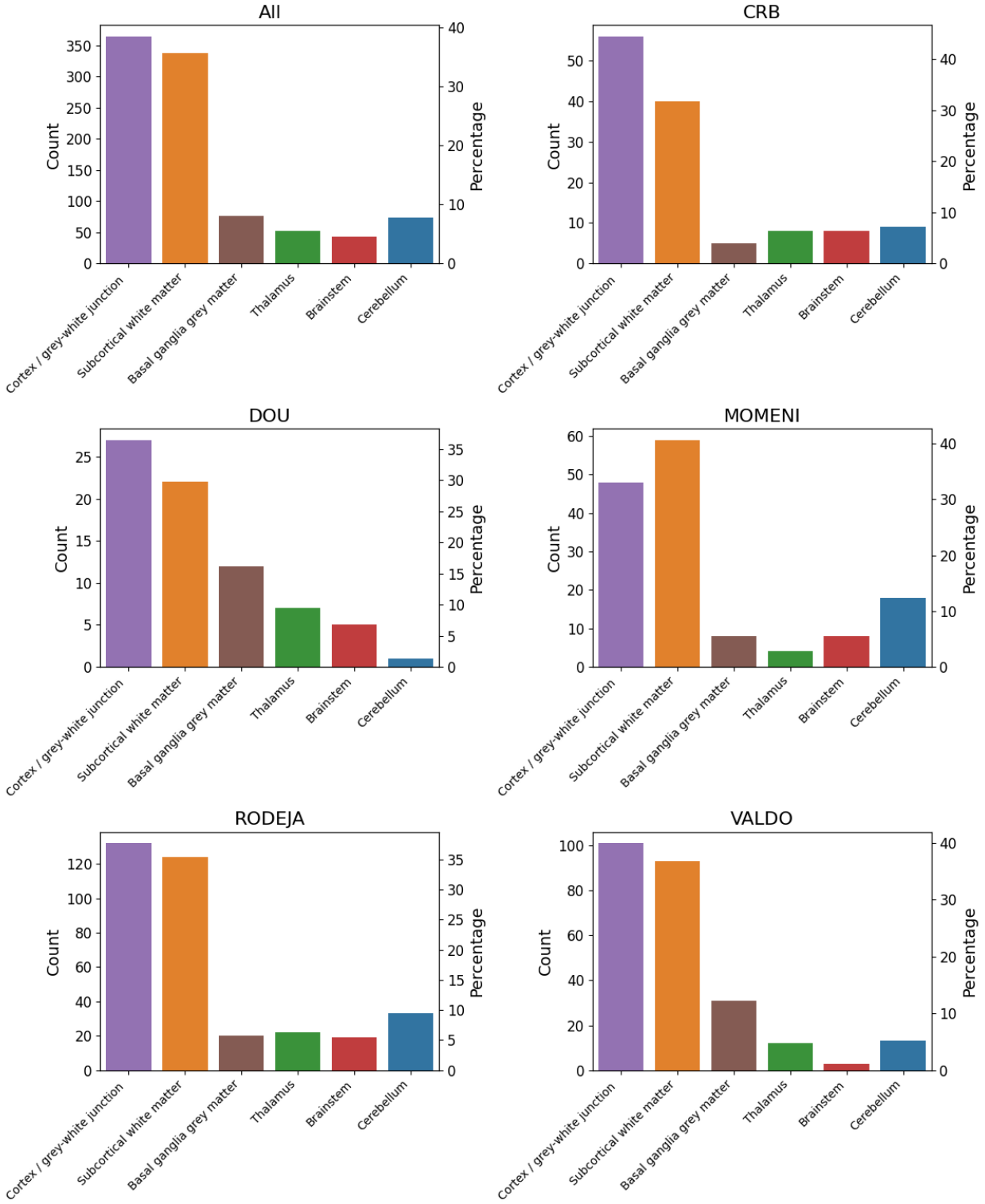



Figure 4.4: **Counts and percentages of CMBs for each location and dataset**. Each subplot represents the total number (left y-axis) and percentage (right y-axis) of CMBs in each of the BOMBS anatomical regions for a given dataset (or all combined), with location names ordered as in the original BOMBS paper.

# 5 Data Pre-processing

We received (or generated) CMB annotations from the different datasets in different forms, so a label refinement process was necessary custom to every dataset. Additionally, data was in different resolutions and sizes, so data needed pre-processing to fit to a standardized space for both segmentation masks and MRI images. This section explains the different steps that we performed to prepare the masks and pre-process the data before being used for either training or evaluation.

At every step of the process, processing metadata was saved into log files that are made available to allow traceability and metadata was saved to allow for a posterior analysis. Throughout the process, the NIfTI1 file format was utilized [1], and was processed in Python employing libraries such as NiBabel and Nilearn, among others. Whenever necessary, corresponding headers were modified to accurately reflect operations that could affect header information (e.g. affine or voxel dimensions). Given the critical importance of accurately counting CMBs, care was taken to ensure that their count remained consistent from the beginning to the end of the process (e.g. plots for individual microbleeds were created at different steps of the processing and manually inspected). This was to avoid potential discrepancies caused by their small size, which would lead to errors during various processing stages. Also, note that pre-processing was an iterative process that required adapting code and debugging on the fly based on observed behavior.

The total size of the data was 220GB on disk after pre-preprocessing. Therefore, efficient data loading and processing was a must and all pre-processing computations were done leveraging Python's multiprocessing library to parallelize operations on up to 40 CPUs.

### Step 0: Data Cleaning

Each dataset presented unique characteristics, from data storage formats and folder structures to subtler aspects like variations in coordinate indexing. Generally, for MRI scans, any NaN values detected were replaced with the median background value to maintain consistency. The process of annotating CMBs varied across datasets, necessitating tailored approaches to achieve a single clean binary mask. In the CRB dataset, this involved isolating seeds corresponding to different microbleeds within the same mask (several masks were obtained for same study, each containing sometimes more than one CMB). In contrast, the RODEJA dataset featured distinct labels for each CMB in the masks which needed to be collapsed into binary.

## 5.1 Label Refinement

### Step 1: Region Growing

In certain datasets, CMB annotations were provided solely as the coordinates of their centers or as scribbles (group of sparse points). Consequently, we needed to generate proper segmentation masks covering the microbleed intensity profile observed in the MRI image. To achieve this, we employed a custom implementation of a Region Growing algorithm as shown in Algorithm 1. This algorithm initializes using the center coordinate as initial seed point — or all points from a scribble — and expands to adjacent voxels based on intensity similarity, using either a 6-connectivity or 26-connectivity. The process is managed through a breadth-first search strategy, utilizing a FIFO (First In, First Out) queue that starts populated with the seed points and and initial mask containing them. The algorithm checks each element's neighbors within the image boundaries, adding those with an intensity difference below a set tolerance threshold to both the queue and the mask. The region-growing algorithm employs two possible methods for assessing intensity differences: (a) "Parent-son," comparing the intensity of each voxel with its neighboring 'parent' voxel; and (b) "Running Average," where a voxel's intensity is compared to that of the average of the grown mask (which contain

[1] https://nifti.nimh.nih.gov/nifti-1

intensities of seed points and newly added voxels during growing). Intensities can be compared either in absolute mode or relative mode.

**Algorithm 1:** Region Growing implementation in 3D Volume

**Input:**

- $V$ - Volume data
- $S$ - Seed points
- $T$ - Tolerance for intensity difference
- $C$ - Connectivity type (6 or 26)
- $Mt$ - Maximum size threshold for region growth
- $Md$ - Maximum distance threshold for growth
- $modeI$ - Intensity mode for reference intensity calculation
- $diffM$ - Difference mode for intensity comparison

**Output:** Region mask $R$ and metadata

```
Initialize queue Q with S;
Initialize region mask R with S;
Set total points totalVox = 0;
while queue Q is not empty do
    Dequeue a point p from Q;
    Compute refI based on modeI;
    foreach neighbor n of point p do
        if n is within bounds of V and not visited and distance to S is less than Md then
            Calculate intensity difference diff between n and refI (using diffM);
            if diff < T then
                Add n to R;
                Enqueue n in Q;
                Update refI based on modeI;
                Increment totalVox;
                if totalVox > Ms then
                    break;
    Mark p as visited;
return R and metadata;
```

Therefore, four parameters need to be optimized: threshold tolerance, connectivity type and intensity method. To set the tolerance hyperparameter, we implemented an automatic method for tuning it for every individual scan. This process iterates through possible tolerance values, observing changes in the region's size. Due to CMBs' distinct intensity contrast, a significant increase in region size is expected when the tolerance is large enough to include surrounding pixels not in the microbleed, which will make total size of grown mask considerably larger due to small size of CMB in comparison. To prevent uncontrolled growth, we imposed a maximum size limit on the mask based on the known maximum theoretical diameter of microbleeds (10mm) and image resolution (exact values in voxels dependant on voxel specs of the image). For example, with 1mm isotropic voxels, the volume of a 5mm radius sphere is approximately 105 voxels, serving as a conservative maximum size. We also set maximum distance of 10mm between any point added to mask and the original seed points. We then identify the optimal tolerance for each image as the point just before this marked size increase, identifying the elbow of change. This method not only ensures precise segmentation of CMBs but also enhances computational efficiency by avoiding exploring unnecessary tolerance levels. All combinations of the rest of input parameters to algorithm were explored on a per-study basis and the

combination that yielded the biggest size of the CMB was chosen (to avoid under-segmentation, which was commonly observed). That is, every study of every dataset was individually optimized for connectivity type, reference intensity mode and intensity difference mode. Then, the estimated radius was computed modelling grown mask as a 3D sphere and saved along with other metadata from processing. It is important to note that these computed masks were only used to estimate the radius of the microbleed.

### Step 2: Sphere creation

Whenever 3D segmentation masks (all voxels within a CMB annotated) were available in the raw dataset, this step was omitted. For the other cases, this processing step was implemented: CRB, MOMENI, sMOMENI, DOU. In this phase, for every CMB a sphere with a radius equal to that estimated by the prior region growing step was created around its center seed. This approach aims to mitigate any unintended expansions observed during the region growing phase (even if limited thanks to size and distance constraints) and the non-smooth contours resulting from these expansions. Furthermore, since the primary objective of this task is to accurately detect and count CMBs, precise segmentation masks are not strictly necessary, but rather precise quantification and location of the CMBs. Additionally, we hypothesized that introducing a inductive bias towards sphericity would benefit the model's learning capacity, as CMBs are spherical by definition.

## 5.2 Data pre-processing

### Step 3: Skull Stripping & Cropping

To minimize potential noise during CMB segmentation, we performed skull stripping on MRI images using SynthStrip [89], a deep learning-based brain extraction tool[2]. This process generated the skull-stripped MRI sequence along with the binary mask that distinguished brain from non-brain regions. We then cropped both the MRI image and the annotation using the brain masks to determine the cropping bounds by locating the minimum and maximum indices where the mask is True in each dimension (x, y, z).

### Step 4: Fix Orientation & Resampling

The MRI and annotations were initially reoriented to align with the RAS convention, which in terms of the subject goes like: left to Right; posterior to Anterior; and inferior to Superior. Subsequently, the MRI sequence was resampled to an isotropic resolution using 0.5mm voxel cubes, employing linear interpolation; for voxels outside the input volume, the fill value was set to the minimum intensity observed in the image. The CMB mask was then resampled to match the MRI sequence dimensions, using nearest neighbor interpolation with a fill value of 0.

### Step 5: CMB Pruning and Cleaning

An additional step is implemented for the RODEJA, CRB, and DOU datasets to address potential CMB artifacts resulting from the resampling process. Often, due to the inadequate slice thickness, resampling may cause CMBs to stretch excessively, leading to fragmentation into multiple parts. To refine the dataset, we employ a pruning process where only valid CMBs are retained. This involves mapping the center of mass coordinates into the resampled image and preserving only those connected components that encompass these centers. Following pruning, each CMB connected component undergoes a cleaning process using morphological operations. Initially, holes within the components are filled. Subsequently, a connected component analysis is conducted, where each component is subjected to two dilations followed by two erosions (closing operation). After these operations, centers of mass are recalculated, and new radii are estimated based on a perfect sphere approximation (those are the radius depicted in Figure 4.2).

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

Table 5.1: **Overview of Dataset Processing Steps for every Dataset**. Each dataset undergoes different processing steps, indicated by ✓if done or ×if skipped. Raw annotations vary among datasets, impacting the applicability of each step.

| Dataset | Raw CMB Annotation | Step 0 | Step 1 | Step 2 | Steps 3-6 |
|---|---|---|---|---|---|
| CRB | scribble | ✓ | ✓ | ✓ | ✓ |
| CRBneg | n/a | ✓ | n/a | n/a | ✓ |
| DOU | center of CMB | ✓ | ✓ | ✓ | ✓ |
| MOMENI | center of CMB | ✓ | ✓ | ✓ | ✓ |
| RODEJA | full 3d mask | ✓ | × | × | ✓ |
| VALDO | full 3d mask | ✓ | × | × | ✓ |
| sMOMENI | center of CMB | ✓ | ✓ | ✓ | ✓ |

### Step 6: Padding and Standardizing

Prior to training, the images were padded centrally with zeros to a uniform size of 400x400x400 voxels, a dimension selected to exceed the maximum size across all datasets while allowing for an additional margin. Subsequently, the images were normalized to zero mean and unit variance. This normalization step was also repeated after data augmentations (refer to Section 6.3 for details on augmentations).

## 5.3 Overview

Table 5.1 outlines the processing steps applied to each dataset. To provide a visual perspective of the outcomes, Figure 5.1 displays 3D renders of final CMB masks for a random selection of 10 CMBs from validation and test sets. This visualization helps assess the morphology of the CMBs, highlighting both the smallest and largest CMBs at the left and right sides respectively, with intermediate sizes uniformly sampled between them. Most CMBs exhibit the expected spherical shape. However, the resampling process has distorted some, leading to tubular rather than spherical shapes (see for instance the VALDO CMB with a 2.25 mm radius). The visualization also clearly shows edge-cases, such as the 5.6 mm radius CMB in RODEJA, which appears more like a macrohemorrhage. In contrast, CMBs from the DOU dataset are perfectly spherical, likely benefiting from superior resolution and slice thickness that mitigated resampling effects. Similarly, the MOMENI dataset shows well-defined, smaller CMBs.

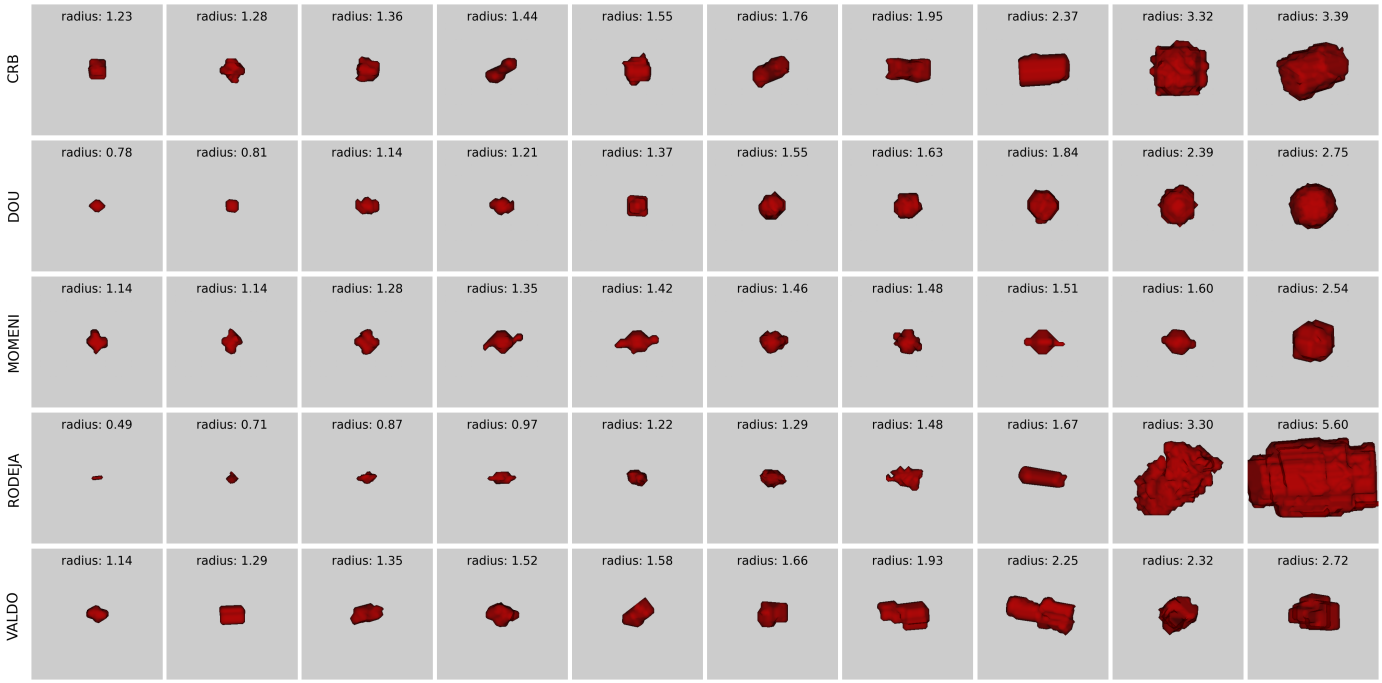



Figure 5.1: **Selection of CMB 3D Shapes Across Datasets** Each row contains a uniformly distributed random selection of 10 CMBs between the smallest and largest CMBs (included) based on estimated radius, as derived from modelling CMBs as perfect spheres, arranged by increasing size. 3D rendering has been performed using the VTK Python library. sMOMENI has not been included as all samples have the same Gaussian spherical shape due to the synthetic generation process.

# 6 Methods

Our goal is to develop an automated system capable of accurately detecting CMBs on MRI sequences employed in clinical practice, and that is robust against the data variability typically found in clinical settings. The motivation behind this development is to enhance the clinical value provided by CMB-based diagnostic and prognostic assessments, which require accurate localization and count of CMBs in the brain. This effort aims to address the gaps identified in previously published methods. The following sections detail how we have framed this task and the specific methodological approach we have adopted to achieve these objectives.

## 6.1 Task definition

Clinically, both the count and specific locations of CMBs are critical for accurate diagnosis and effective treatment planning. The precise location of CMBs holds significant implications for understanding the underlying pathology, as different brain regions affected by microbleeds can influence the clinical approach. Thus, simply classifying the presence of CMBs is insufficient for medical use, as physicians require detailed information on the localization and context of each microbleed (e.g. to differentiate them from possible mimics in the imaging or understand possible cause).

Object detection might seem suitable in this scenario, however, it does not provide the exact boundaries or the extent of each microbleed. Instead, semantic segmentation offers information on every voxel belonging to each CMB, allowing for the extraction of quantitative measures such as size and shape, which we foresee could be valuable in future clinical applications using CMBs.

The problem with the latter approach is in the case of CMB overlapping between them and being predicted as one. Thus, Instance Segmentation seems to be the perfect match for this task. However, as the risk of CMBs overlapping is minimal and because we wanted to leverage some available pre-trained model weights we decided to define our task as **Semantic Segmentation**. We believe that we do not really use the detection part of the problem for this specific application. For instance, evidence from the VALDO challenge indicate that segmentation performance closely correlates with detection outcomes, even though detection was often a secondary task for many participants [13]. The task then becomes to generate a segmentation masks that classifies each voxel as either background or a CMB. Subsequent connected component analysis within these masks enables the determination of the count and precise location (object detection) of each microbleed, facilitating a more comprehensive clinical assessment.

### 6.1.1 Input MRI sequence

SWI evolved from traditional two-dimensional T2S sequences into three-dimensional sequences with enhanced spatial resolution and improved susceptibility contrast, which enhances the visualization of lesions previously identified with standard T2S — including cerebral microbleeds — and allows the detection of new ones previously undetected [17]. This evolution increased the clinical adoption of SWI and is now the gold standard for detecting CMBs with higher sensitivity [23]. Despite these advantages, the differences in CMB prevalence observed between T2S and SWI sequences do not significantly impact their association with clinical parameters, indicating that results from both types are clinically comparable [23]. Moreover, a notable challenge with SWI is the increased likelihood of FPs due to its increased sensitivity to calcifications, which can mimic CMBs. Additionally, the requirement for post-processing in SWI, which may not be integrated into older MRI scanners, means that T2S remains widely used.

Our ideal system should be able to operate on typical MRI sequences currently used in clinical practice for CMB detection. Therefore, it is essential for our model to be **T2S/SWI sequence-agnostic**, capable of detecting CMBs on either of them, depending on their availability. This approach ensures broad applicability and utility across different clinical settings.

## 6.2 Model

In the task of segmenting CMBs, it's crucial to leverage three-dimensional data to distinguish them from their primary mimics — notably from blood vessels, with their characteristic tubular structure — and literature shows that DL-based systems have consistently outperformed other methods in this area. In face of this, we have selected as backbone a model architecture based on a 3D U-Net [51].

### 6.2.1 Apollo architecture

Apollo is the ML backbone of CEREBRIU's commercialized software for detecting infarcts, hemorrhages, tumors and edemas on different brain MRI sequences. We employed this model's architecture (slightly modified) to be able to leverage its pre-trained model weights for transfer learning. This architecture is a convolutional neural network (CNN) of the encoder-decoder type, inspired by the U-Net [51] and very similar to 3D U-Net [90], with several significant modifications tailored for 3D image segmentation. Key features differing from the original U-Net implementation are:

- **Dimensionality** The model processes 3D inputs and outputs 3D segmentation maps, with both the input and output represented as $96 \times 96 \times 96$ voxel cubes. Each voxel reflects a $0.5 \times 0.5 \times 0.5$ mm$^3$ segment of physical space, amounting to a total volume of $4.8 \times 4.8 \times 4.8$ cm$^3$.
- **Mapping** Functionally, the model maps an input tensor $x$ to an output tensor $\hat{y}$ through a transformation defined as $f(x;\theta) \rightarrow \hat{y}$, where $\theta$ denotes the model's learnable parameters. Specifically, the input tensor $x$ is in $\mathbb{R}^{96\times96\times96\times1}$, indicating a single-channel input that represents intensity values for either SWI or T2S MRI sequences. The output tensor $\hat{y}$ is in $\mathbb{R}^{96\times96\times96\times2}$, where the two channels encode normalized probability values (from softmax function) for each voxel, categorizing them as either 'normal/no pathology/background' or 'micro-hemorrhage'.
- **Model Depth:** The depth of the model has been increased to five encoder and decoder blocks (compared to four in the original U-Net). Each block includes three convolutional layers, utilizing strided convolutions for downsampling instead of max-pooling.
- **Filter Configuration:** The number of filters starts at 32 (instead of 64) at the first depth and doubles at each subsequent level, capping at 320 after the fourth depth.
- **Activation Function:** The convolution layers utilize a leaky ReLU activation function with a slope parameter $\alpha = 0.001$, a modification from the standard ReLU used in the original U-Net.
- **Normalization:** Instance normalization layers are introduced to normalize the outputs at each convolutional block, enhancing model stability and performance.

The model architecture employed in this study can be seen in Figure 6.1. The final number of parameters of this architecture is 36302116. The next sections explain how the original model was trained and how the pre-trained weights were leveraged for the modified architecture.

### 6.2.2 Apollo training

The Apollo model utilized in this study was pre-trained on a diverse mix of internal proprietary and external public datasets, comprising studies with a range of conditions including normal scans, tumors, hemorrhages, and infarcts. A very rich and diverse set of data augmentations was also used.

The internal datasets, which were strictly curated to include cases with comprehensive radiological reports and a variety of MR sequences, excluded pediatric cases, post-operative conditions, and patients with multiple micro-metastases. The demographic profile for these datasets spanned a broad age range and included subjects from Brazil, India, Ukraine, and the USA, with a balanced gender distribution. Public datasets were also used in training to enhance the model's robustness by introducing a variety of slightly out-of-distribution data. These included datasets from several well-known sources like BraTS19, BraTS23-Meningioma, BraTS23-Metastases, ISLES22, and Meningioma-SEG-CLASS.

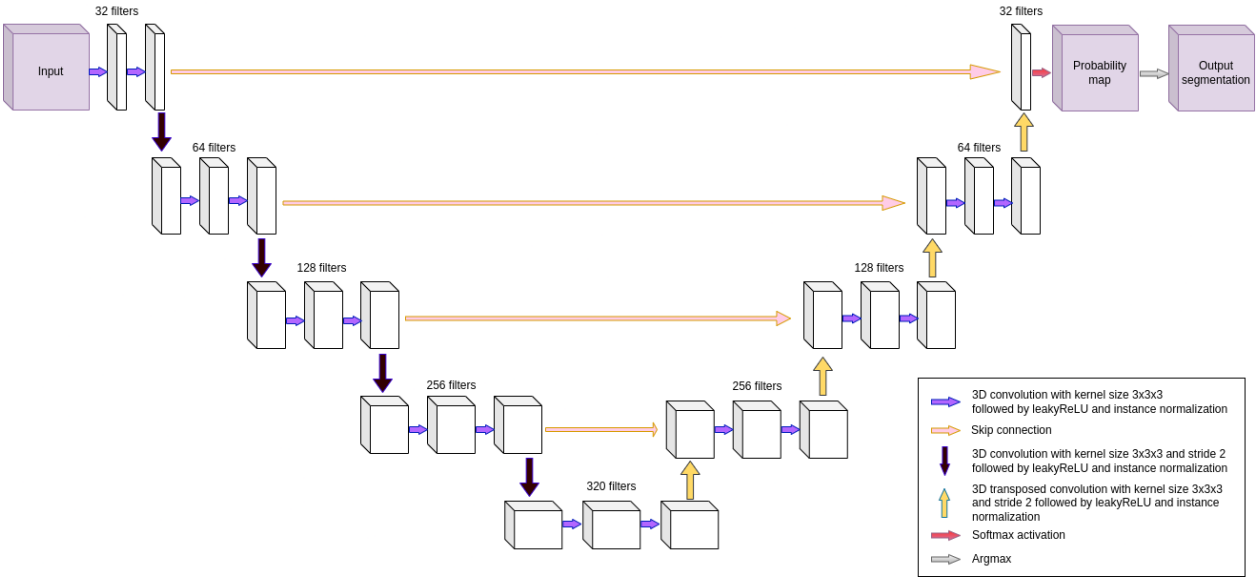



Figure 6.1: **3D U-Net Apollo architecture**. White boxes represent multi-channel feature maps, with the number of channels indicated on top. The width represents qualitatively the number of channels, which is equal to the number of convolutions 3D filters applied in the previous convolution. These are actually 4D, so the dimensions shown do not correspond with the real shape of the matrices. Purple boxes represent a 3D volume, being either input MRI patch, voxel-level probabilities for predicted class, or predicted segmentation.

Annotation within the internal datasets was meticulously conducted by certified radiologists who performed detailed voxel-level segmentation for major pathology sub-types, especially focusing on different progression stages of hemorrhages. Image pre-processing steps standardized the orientation, resolution, and intensity across scans, and involved techniques like creating brain masks, resampling, and padding smaller images for consistency.

While exact numbers cannot be disclosed due to privacy concerns, it is important to note that the data pool utilized in this study was significantly larger than those typically used in most published deep learning-based CMB detection methods. This extensive dataset effectively represents the clinical context across both high-resource and low-resource settings.

### 6.2.3 Transfer of weights

It is important to understand that the input and output channel configurations are a simplification of the original Apollo architecture, which included 3 input channels for three sequence types (FLAIR, SWI/T2S, and DWI), and five output channels for 5 categories (four pathology types plus normal/absent category). Consequently, weights must be carefully selected to leverage the pre-trained model.

Firstly, our model processes inputs using only one channel. To accommodate this, we needed to reduce the dimensions of the weight matrix in the first layer of the encoder block from (3, 3, 3, 3, 32) to (3, 3, 3, 1, 32). Secondly, while the original model outputs five different categories, our interest is focused solely on two outcomes: the presence or absence of CMBs. Consequently, the weight matrix in the output convolutional block should be reduced from size (1, 1, 1, 32, 5) to (1, 1, 1, 32, 2). This also implies reducing the number of sigmoid activation parameters, and changing the weight dimensions from (5,) to (2,).

Given that one of the input channels corresponds precisely to the SWI/T2S MRI sequence, we opted to retain only the filters associated with that channel in the encoder block. Furthermore, because hemorrhages are among the original output categories, we preserved the weights related to generating logits for this category in the output convolutional block and the subsequent softmax function.

## 6.3 Data Augmentations

To increase the model's robustness against typical MRI data variations, we implemented an extensive data augmentation pipeline. This approach introduces either invariance or equivariance properties to the model against the set of transformations (including potential composite transformations) generated by the augmentation pipeline. If appropriately selected, these augmentations enhance model robustness by mimicking real-world variations such as noise patterns and other image transformations that are likely to occur in clinical settings but should not affect the model's output. This ensures that the model remains reliable under varying imaging conditions. These were applied on the whole MRI (and segmentation maps if spatial) and during the data loading phase for training, before performing the patch sampling. The data augmentation process consists of three types of random transformations applied sequentially: spatial deformations, intensity modifications, and Gaussian blurring. Each transformation is applied with specific probabilities—0.4, 0.6 and 0.25 respectively— and specific parameters based on empirical evidence gathered from previous projects at CEREBRIU. Additionally, employing the sMOMENI dataset serves also as a method of data augmentation aimed at increasing the quantity of annotations available for training. However, it is important to note that this does not diversify scanner parameters or imaging conditions, as the MRI images are the same as in MOMENI but with CMB intensity profiles added.

### 6.3.1 Random Spatial Deformations

The augmentations are based on the SynthSeg RandomSpatialDeformation class. In this case, both the MRI image and the label map are applied same spatial transformation. Two types of transformations are applied sequentially with probabilities of 0.6 and 0.3:

**Affine Transformation**: An affine transformation is applied by first sampling parameters for rotation, shearing, scaling, and translation. These parameters are combined to form a transformation matrix that is then sequentially applied to the image. The parameters for each transformation are derived as follows:

- Scaling: Each dimension's scaling factor is independently sampled from a uniform distribution within the range [0.95, 1.05].
- Rotation: The rotation angle for each dimension is independently sampled from a uniform distribution within the range [-0.2, 0.2] radians.
- Shearing: The shearing factor for each dimension is independently sampled from a uniform distribution within the range [0.99, 1.01].

**Elastic Transformations**: Images and segmentation maps were deformed by applying delta offsets on a voxel meshgrid across all spatial axes. The deformed images and maps were then resampled using linear interpolation for images and nearest-neighbor interpolation for segmentation maps. Default parameters from the class were used.

### 6.3.2 Random Intensity Transformations

This transformations modify the value of individual voxels' intensities:

**Gaussian Noise:** noise values are drawn from a Gaussian distribution and applied to the original intensities with a 25% probability.

**Salt and Pepper Noise:** This noise, representing dead pixels or sensor faults, is introduced into the images with a 25% probability.

**Gamma Augmentation:** With a 50% likelihood, gamma corrections are used to adjust brightness and contrast levels.

**Contrast Inversion:** Applied with a 25% probability, this technique inverts the contrast of the image.

### 6.3.3 Random Gaussian Blur

Gaussian blurring is applied to 25% of images to simulate the variability in scanner resolutions or focus levels encountered in clinical MRI settings. This process uses a sigma range from 0.5 to 2.5, effectively blurring the

image to mimic different qualities of imaging equipment. Specifically, the blurring is achieved by convolving an image channel with the first derivative of a Gaussian kernel, where the standard deviation is randomly chosen from a uniform distribution between 0.5 and 2.5.

## 6.4 Patch Sampling Strategy

To address the memory constraints of GPUs and effectively process MRI volumes, instead of inputting the whole MRI volume to the model, we predict on smaller, localized sub-volumes or patches extracted from the larger original image. As can be seen in Figure 6.2, the strategy in which patches are extracted or sampled from the full volume varies between training and inference time.

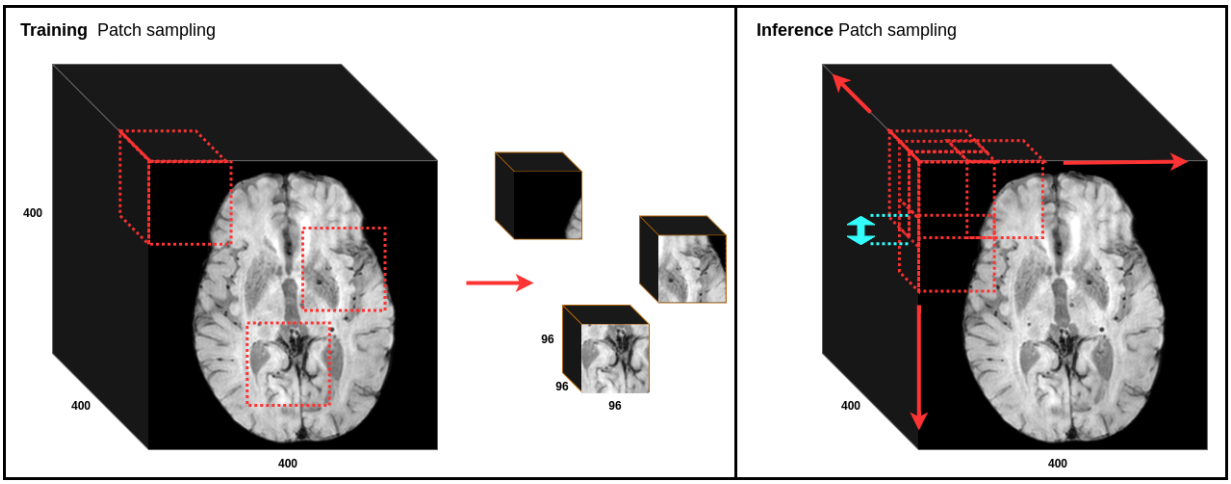



Figure 6.2: **Patch Sampling Strategy During Training and Inference**. *During training (Left)*, a fixed number of patches are randomly sampled from the 3D volume at various locations. Half of these patches are guaranteed to contain a CMB, with its center randomly shifted within the patch. *During inference (Right)*, patches are deterministically sampled with a predefined overlap in each dimension, as indicated by the blue arrow in the image. If the overlap is 0, the next patch begins immediately after the previous one. With shown dimensions for images and patches, and an overlap of 0.4, this would results in 343 patches instead of 125 when no overlap present.

### 6.4.1 During Training

During training, we loop over the training set scans and sample a fixed number of 3D patches from each scan whose size matches the input size of the model. Because we iterate frequently over the scans, all areas of all images are likely visited over several epochs. Whenever a scan is sampled, the patches are added to a queue, whose size is determined by the memory capacity of the system used for training. These patches in the queue are then randomly sampled and sequentially selected in mini-batches (as many as can fit in the GPU being used) to be used for backpropagation optimization (see section 6.5 for more details).

These patches are extracted by randomly selecting starting points within the permissible boundaries of the image. Due to the substantial class imbalance inherent to the data and task at hand, we facilitate faster learning by setting a minimum proportion of patches containing a CMB. These CMB-containing patches are generated by identifying the CMB locations, placing the CMB in the middle of the patch, and then randomly shifting its position within the patch. The reason for the shifting is to prevent the model from learning that CMBs are always in the middle and to enhance the model's robustness to various CMB positioning in the patch. Key parameters at this stage include the size of the patch, *patch_size*; the class-specific proportions of patches, *class_props*; and the total number of patches sampled per training iteration, *num_patches*.

### 6.4.2 During Inference

During model evaluation or inference on full images (happening at the end of every epoch on validation set or on the test sets after model's training), patch extraction is deterministic and involves a systematic extraction that covers the full volume with a defined degree of overlap between patches. This overlap is crucial for reducing boundary effects between patches. Special attention must be paid to dimension matching, often requiring cropping of patches on the edges. In the post-prediction reconstruction phase, logits from overlapping regions are aggregated by summation. The parameter *overlap_frac* governs the extent of overlap between patches during evaluation.

## 6.5 Optimization

Optimization in deep learning specifically involves finding parameters $\theta$ that minimize a cost or loss function $L(\theta)$, which typically includes a performance measure evaluated on the training set as well as additional regularization terms. The objective is to adjust the model's weights to reduce prediction errors through iterative training, which is usually done using **backpropagation**. This method starts with a forward pass, where the model computes outputs based on given inputs for every parameter. Then, in the backward pass, the model calculates the gradient of the loss function for each parameter using the chain rule of calculus. This gradient represents how much a small change in each parameter will increase or decrease the loss and are propagated backward through the network. One way of updating these parameters is through **Gradient Descent (GD)**, a first-order iterative algorithm for unconstrained optimization that updates the parameters values by taking repeated steps in the opposite direction of the (approximate) gradient of the loss function at the current point (which has been back-propagated), that is, in the direction of steepest descent. The simplest update rule for GD is $\theta^{(t+1)} = \theta^{(t)} - \eta \nabla L(\theta^{(t)})$ where $\theta^{(t)}$ denotes the parameter values at iteration $t$, $\eta$ is the learning rate, and $\nabla L(\theta^{(t)})$ is the gradient of the loss function at these parameter values. Subsequent sections will further explore specific optimization strategies we used in training our model.

### 6.5.1 Loss Function

Class imbalance is already a significant challenge in medical image analysis, and is particularly pronounced when dealing with CMBs, which are tiny lesions that represent a minuscule fraction of the total volume in an MRI scan. For example, in the VALDO dataset, CMB voxels constitute only about 0.0007% of the total voxels. In this scenario, a naive implementation with a non-weighted loss function could lead to suboptimal training, where the model predicts all voxels as the background or neglects small foreground classes in favor of segmenting larger ones. To address these challenges, several strategies have been adopted in the literature, including the use of weighted loss functions [50, 51], step-wise training approaches [91, 92], and resampling techniques to balance class distributions [91]. Notably, the generalized Dice loss [93] and Tversky loss [94] have been specifically formulated for semantic segmentation tasks to mitigate class imbalance by allowing adjustment in the sensitivity towards FPs and FNs. In our approach, we employ a balanced combination of two advanced loss functions.

**Class-weighted Categorical Cross-Entropy Loss** — We employ a weighted cross-entropy loss to handle class imbalance effectively. The loss for each class is weighted differently to prioritize rare classes such as CMBs. The formula for the weighted cross-entropy loss is given by:

$$L_{\text{CE}} = -\sum_{c=1}^{C} w_c \cdot y_c \log(p_c) \tag{6.1}$$

where $y_c$ is the binary indicator (0 or 1) if class label $c$ is the correct classification for observation $o$, $p_c$ is the predicted probability of observation $o$ being of class $c$, and $w_c$ are the weights associated with each class. Thus, the loss function becomes effectively a Weighted Binary Cross-Entropy Loss, as $C$=2.

**Focal Tversky Loss (FTL)** — This is a variant of focal loss that incorporates the Tversky index to enhance the precision-recall balance, which is crucial for segmenting small lesions [95]. The Tversky index

($TI_c$) and the FTL are defined as follows:

$$TI_c = \frac{TP}{TP + \alpha \cdot FP + \beta \cdot FN} \tag{6.2}$$

$$FTL = \sum_{c=1}^{C} (1 - TI_c)^{\gamma} \tag{6.3}$$

where $TI_c$ assesses class-specific segmentation accuracy. The parameters $\alpha$ and $\beta$ are used to balance the emphasis on precision or recall, thus accommodating different degrees of class imbalance. Notably, when $\alpha = \beta = 0.5$, $TI_c$ resembles the Dice Coefficient. For cases where $\alpha + \beta = 1$, it corresponds to a specific $F_\beta$ score. $FTL$ becomes useful for class imbalance when $\gamma > 1$, as this setting increases the loss gradient for cases where $TI_c < 0.5$. This intensification forces the model to focus more on challenging cases, notably in small-scale segmentations that often receive lower $TI_c$ scores. Consequently, the model is driven to prioritize enhancements in accuracy within these difficult-to-segment areas.

The final loss function used in our training combines the two previous functions and weights them with $\lambda_1$ and $\lambda_2$ weighting factors as follows:

$$L_{\text{combined}} = \lambda_1 \cdot FTL + \lambda_2 \cdot L_{\text{CE}} \tag{6.4}$$

### 6.5.2 Adam Optimizer

We use the Adam optimizer [96] optimization algorithm for backpropagation. Adam is analogous to an earlier optimization algorithm, RMSProp [1], but incorporates a momentum term, enhancing its efficiency and convergence rate. Adam applies adaptive learning rates to all parameters in the vector $\theta$, adjusting these rates based on the first and second moments of the gradients. These moments are computed as exponentially moving averages, with the decay rates controlled by the hyperparameters $\beta_1$ and $\beta_2$ as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

where $g_t$ represents the gradient of the loss function with respect to the parameter $\theta$ at time step $t$, and $m_t$ and $v_t$ are estimates of the first and second moments of the gradients respectively. However, the algorithm must manage the zero-initialization of these moment vectors, which introduces a bias toward zero in initial updates. To address this, Adam includes a bias-correction step before each parameter update — ensuring more accurate adjustments in the early phases of training — as follows:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

The final update rule is then given by:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{6.5}$$

This adaptive approach helps the optimizer to adjust based on the topology of the error surface, lessening sensitivity to initial learning rate configurations.

### 6.5.3 Training Procedure

Once we have computed the loss between predictions and ground CMB binary masks, we are able to update our model parameters through backpropagation and gradient descent. Ideally, the gradient is calculated across all training samples (batch learning) and the weights are updated according to the specific optimization algorithm being used. However, as we are working with very big 3D images, computing gradients on every

[1] RMSprop is not published and was proposed by Geoff Hinton in an online course: Course Slides.

training sample for each update becomes computationally prohibitive due to high run-time and memory demands. In contrast, we use mini-batch learning or **Stochastic Gradient Descent (SGD)**, which involves selecting a random subset $D_m$ from the entire dataset $D$ and estimating the gradient from this subset. This approach is feasible because most loss functions in machine learning can be expressed as a sum of individual losses over training examples, allowing the gradient computed from a subset to serve as an unbiased estimate of the gradient over the full dataset.

Our training procedure continuously cycles through the dataset, sampling a fixed number of patches per scan that are added to a queue (see Section 6.4). In this setup, a mini-batch consists of a group of patches sequentially sampled from the shuffled queue. The size of these mini-batches varies, depending on GPU availability for each experiment, with batch sizes typically ranging from 3 to 5 patches. The batch size is determined by what memory is left on GPU after including the model's weights, the activations and gradients for backpropagation, and the parameters of the optimizer. Each mini-batch consumption for backpropagation and GD constitutes a single **gradient step**. The queue is continuously replenished with new patches sampled from the scans until all scans have been sampled, after which the sampling cycle restarts without interruption. In this context, the traditional concept of an 'epoch'—a stochastic gradient pass over the entire dataset—does not apply. Instead, we define an **epoch** as completing 5000 gradient steps and is used merely as a reference for tracking and knowing when to perform evaluation on the validation set. It is important to note that this definition does not correspond to a complete pass through all data, but rather to an iteration through 30 sampled patches from every scan in the dataset.

### 6.5.4 Validation Procedure

At the end of every epoch, we evaluate the model's performance on a randomly selected subset comprising 50% of the validation set. This subset always includes an equal distribution of healthy scans and scans containing CMBs, to facilitate a balanced assessment of the model's capabilities in detecting both normal and pathological tissues. The decision to use only half of the validation set for routine evaluations is driven by computational efficiency, as inferring on one image takes substantial time, even if the batch size used can be bigger. The batch size is determined by what memory is left on the GPU after including the model's weights and is typically 6-10 depending on the specific hardware used. The reason is that when evaluating one scan and using overlap the number of patches is very high, for instance, with a 40% overlap of 96x96x96 patches to span all voxels in a 400x400x400 volume, this yields 343 patches instead of 125 with 0 overlap.

At the end of every epoch, evaluation metrics (same as explained in Section 6.7) and plots of predictions are generated and saved to monitor the model's progress both analytically and visually on the validation set. Five models are saved and overwritten whenever any of the following metrics is improved: loss, count-based F1 and PPV detection scores, and global dice score.

## 6.6 Data Post-processing

We introduced a second stage of CMB verification by incorporating an additional post-processing step to refine the predicted masks. Using SynthSeg [87] (robust version), we segmented the MRI volume into distinct anatomical brain labels (for details on SynthSeg, see Section 4.9.3). Unlike CMBs in the ground truth, which we can assume to reside within brain tissue (even if SynthSeg indicates differently), we cannot make this assumption for predicted CMBs, as it may be the model predicting a FP in the wrong region. Given that SynthSeg's segmentation accuracy on SWI and T2S has shown limitations in our project, we opted to filter out only those predicted CMBs that showed no overlap whatsoever with any brain anatomical regions. This included full overlaps with non-brain structures such as the background, ventricles, CSF, or ventral DC. This approach aims to mitigate the model's limited global contextual understanding, which is confined to local contexts within small patches.

## 6.7 Validation metrics

The validation of ML algorithms in Medical Image Analysis (MIA) often faces deficiencies, particularly in aligning developments with clinical practice, primarily due to the lack of validation based on relevant

metrics [97,98]. Addressing this, we used the *Metrics Reloaded framework* [99] to guide our selection process of suitable detection- and segmentation-based validation metrics for this particular task.

### 6.7.1 Detection Performance

Object detection is significantly different from segmentation in terms of the metrics used for evaluation. One difference is the need to identify individual instances within the same class, which needs a step to locate and correctly assign predicted objects to their respective reference entities. More importantly, in object detection **there are no true negatives** (absence of prediction is background, which is everywhere on the image), which makes several common classification metrics like accuracy, specificity, and AUROC inapplicable. In our case we are dealing with a binary classification problem (whether I detected a CMB GT object or not), where we can only rely on three components of the confusion matrix, excluding true negatives.

We considered every individual 3D connected component in the predicted CMB mask as one predicted object, and every connected component in the GT CMB mask as a true object. We followed recommendations from Metrics Reloaded [99] and followed a three-step process for evaluating our predictions:

1. Localization Criterion: we evaluated two criteria: Point Inside Mask and Center Distance. In the first we considered a match whenever any point of the predicted CMB overlaps with a GT CMB. In the second we calculate the Euclidean distance between the center of mass (equivalent to the geometric center since all voxels in a CMB have equal value) of the predicted and GT CMBs, considering it a hit if the distance is less than 5 mm. After experimentation and acknowledging that GT CMBs were not always perfectly annotated or processed, thereby not fully covering the CMB intensity profile, we found the Point Inside Mask criterion could miss true predictions. Consequently, we opted for the **Center Distance** criterion.

2. Assignment Strategy: Since localization does not always yield unique matchings, a strategy to resolve potential ambiguities is necessary. [2] We employed Greedy by Score Matching, where after calculating all localization criteria scores between predicted CMBs and GT CMB, we iteratively match each GT CMB to its highest scoring predicted counterpart, provided it meets a valid localization criterion (e.g., distance less than 5mm). In instances where multiple predictions match the same GT CMB, only the prediction with the best score is considered a hit. This leaves the following possible call counts for every study:
   - *True positives*: predicted CMBs successfully matched with a GT CMB.
   - *False Positives*: predicted CMBs left unmatched
   - *False Negatives*: GT CMBs left unmatched

3. Classification metrics: we generate a confusion matrix without TNs, and compute the following metrics:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

*Precision*, also known as Positive Predictive Value (PPV), provides information about the relevance of the predicted CMBs, indicating what fraction of the predictions are relevant, that is, the probability that a predicted CMB is actually a true CMB ($P(\text{CMB}_{\text{true}}|\text{CMB}_{\text{pred}})$). On the other hand, *Recall*, also known as True Positive Rate (TPR), provides insights into how many of the existing CMBs in the GT were successfully detected by the model. It tells us how many of the true CMBs we hit, or in other words, the probability that a true CMB was identified by the model ($P(\text{CMB}_{\text{pred}}|\text{CMB}_{\text{true}})$). An especially useful single performance measure for the overall model performance as a whole in this context is the *F1 score*. It is defined as the harmonic mean of precision and recall and has favorable characteristics especially when prevalence is

[2] For example, if a large predicted CMB overlaps with two closely positioned true CMBs, without a proper assignment strategy, this situation might mistakenly be counted as two TPs instead of one TP and one FN. Conversely, if two predicted CMBs overlap with a single true CMB, this could incorrectly result in two TPs instead of one TP and one FP.

low [100]. A high score on this metric indicates that the true CMBs in the scan were successfully detected (high recall) without the cost of too many false positives (which would lower precision).

Finally, we adhere to recommendations from [35] and add additional metrics to address the known issue of false positives more granularly: *Average False Positives per Scan* ($FP_{scan}$) and *Average False Positives per CMB* ($FP_{cmb}$), defined as follows:

$$FP_{scan} = \frac{FP_{total}}{n}, \quad FP_{cmb} = \frac{FP}{m}$$

where $n$ is the number of scans, $m$ represents the total number of CMBs in all scans and $FP_{total}$ is the total count of FPs for the whole dataset.

### 6.7.2 Segmentation Performance

The specific segmentation task significantly influences evaluation results, making it crucial to understand metric behaviors and expected scores for valid MIS performance interpretation. Depending on the ROI type, such as lesion or organ segmentation, the complexity of the task and the resulting expected score vary significantly [101]. For instance, organ segmentation is generally easier due to consistent ROI locations, resulting in higher scores, even though optimal performance metrics in organ segmentation are more likely to be achievable, they are less realistic in lesion segmentation [102, 103]. In contrast, lesion segmentation presents a greater challenge due to the high spatial and morphological variance, which makes it more difficult to achieve high scores. This challenge is particularly pronounced in the case of CMBs, where the tiny size of the lesions can easily result in a lower score due to even slight deviations, even if only a few voxels differ. Additionally, based on our observation of the human annotations, we noted that they often do not fully cover the CMB signal void, which can lead to misleading segmentation scores. Another crucial factor is the number of regions of interest (ROIs) present in an image. The presence of multiple ROIs necessitates additional attention during implementation and interpretation, as high-scoring metrics computed at the scan level can be misleading by concealing undetected smaller ROIs between well-predicted larger ROIs [101]. Consequently, it can be anticipated that segmentation scores will be lower, even if this does not necessarily align with poor detection performance.

Due to the large class imbalance often encountered in MIA, metrics with equal true positive and true negative weighting, like Accuracy or Specificity, can result in high scores even if any pixel at all is classified as ROI, significantly biasing the interpretation value. Thus, these metrics should be avoided [41]. Instead, we employed the Dice Similarity Coefficient (DSC)—also known as the F1 score or Sørensen-Dice index—to quantify the accuracy of our image segmentation task, which is the most used metric in the majority of scientific publications for MIS evaluation [103–105]. We computed the **DSC exclusively for the TP object detections**, setting all voxels not in the matched CMB connected components of both GT and prediction to 0. We do this to assess the alignment accuracy of detected CMBs, once identified by the model, with their corresponding GT representations. We denote this specific metric as $DSC_{TP}$ and compute it as follows:

$$DSC_{TP} = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

where $X$ and $Y$ represent the predicted and GT masks respectively of a single matched CMB.

### 6.7.3 Image-level Classification Performance

We also evaluate the model's ability to differentiate between groups of scans based on the number of CMBs present. We define Group A as having fewer than $th_{cmb}$ CMBs, and Group B as having $th_{cmb}$ (included) or more [3]. The number of predicted CMBs is compared to the number of CMBs present in the annotation. As a straightforward example, if we want to distinguish between healthy and unhealthy scans, we would set $th_{cmb} = 1$. In this setup, with Group A comprising scans with fewer microbleeds than $th_{cmb}$ and Group B otherwise, the classifications are defined as follows:

[3] Specific threshold values used will be later introduced

- *True Positive (TP)*: Correctly classifying a scan as belonging to Group B.
- *False Positive (FP)*: Incorrectly classifying a scan from Group A as belonging to Group B.
- *True Negative (TN)*: Correctly classifying a scan as belonging to Group A.
- *False Negative (FN)*: Incorrectly classifying a scan from Group B as belonging to Group A.

In this scenario, we compute all relevant metrics from the confusion matrix at the dataset level, since each classification type corresponds to an entire scan rather than an individual CMB. The classification metrics chosen are the same as in detection, but also including *Specificity* (or True Negative Rate, TNR), defined as follows:

$$\text{Specificity (TNR)} = \frac{TN}{TN + FP}$$

### 6.7.4 Aggregation of metrics

For segmentation metrics, a confusion matrix (CM) can be constructed for each scan/patient, which can optionally be aggregated by summing them into a single one. For image-level classification, each scan contributes a single value to the confusion matrix. We then define two distinct approaches for calculating these metrics:

- **Scan-averaging**: Metrics are computed individually per scan and then averaged across all scans. This method provides insights into the average performance across the dataset.
- **Global-averaging**: All counts are aggregated into a single confusion matrix from which the metrics are derived. This approach is analogous to micro-averaging in multi-class classification problems, offering a holistic measure of performance across all scans.

We distinguish between these two averaging methods by appending the suffixes *scan-* and *glob-* to the metrics, for example, *scan*Precision and *glob*Precision. During averaging, NaN scores are excluded from the calculation, which can occur, for instance, when the PPV calculation lacks a denominator due to no model predictions. Additionally, the F1 score is set to NaN if either Precision or Recall is NaN. The calculation of the *scan*F1 score requires special attention, as it can be computed in two different ways, similar to the macro-F1 in classification tasks:

- *scan*F1: arithmetic mean of the F1 scores calculated for each scan, the most common way to compute it.
- *scan*F1*: calculated as the harmonic mean of *scan*Precision and *scan*Recall, proposed by Sokolova and Lapalme [106]

While *scan*F1 provides a straightforward interpretation as an average of precision and recall across patients, the *glob*F1 score can be viewed as the overall probability of true positive classifications [107]. This distinction is crucial as *scan*F1 ensures that each individual scan exerts an equal influence on the final metric, helping to mitigate the impact of scans with a disproportionately large number of microbleeds, which could skew the metric significantly either positively or negatively. In contrast, *glob*F1 assigns equal importance to all CMBs in the dataset, regardless of the scan they originate from, ensuring a balanced assessment across the entire dataset.

### 6.7.5 Metrics computation

We are framing the problem as a semantic segmentation task at the voxel-level, whose output we then convert into detection at the CMB-level and classification at the scan-level as explained in sections 6.7.1- 6.7.3. To compute the metrics, **for each study** present in the test set being evaluated on, we perform the following steps:

1. For each CMB in Ground Truth (GT):

- Compute shape, location, and intensity features. Result: *scanId-cmbID* (GT) level metadata

2. For each predicted CMB:
   - Compute shape, location, and intensity features. Result: *scanId-cmbId* (pred) level metadata
   - Compute Center distance with every other CMB in GT as localization score
   - Perform Greedy by Score Matching, coupling GT CMB with predicted CMBs, and match greedily to the predicted CMB with the lowest distance, as long as <5mm
   - Determine call type: True Positive (TP), False Positive (FP)
   - If TP, compute $DSC_{TP}$ (to see how well it aligns with true CMB mask) and save
3. For each CMB in GT not matched to any predicted CMB:
   - Determine call type: False Negative (FN)
4. Compute Object Detection Metrics
   - Whenever possible, compute PPV, TPR and F1. Set to None otherwise
5. Compute Segmentation metrics
   - Get the Dice score at whole-image level for CMB class
6. Compute Image-level classification Metrics
   - For every threshold set, compute classification metrics
7. Save per-study computed metrics and CMB individual calls

After repeating the previous **for all studies in the test set**, we end up with a single confusion matrix from all call type counts for every CMB (note that TP is only counted once when there is a match), from which we can compute the micro-averaged classification metrics. Also, we average the scan-level metrics across scans and obtain the macro-averaged metrics. In this process, we can easily filter out based on metadata at the CMB level (from both GT and predictions), for instance, by removing those CMBs that are e.g. located in a specific area of the brain and recomputing the averaged metrics. Equally, we can filter out certain scans based on scan-level metadata. This helps us investigate performance in a more granular way.

# 7 Experiments

This chapter describes the various experiments conducted, their configurations, and preliminary results used to select a model for further evaluation. The overall workflow of our experiments is depicted in Figure 7.1. We ran a parallel track comparing the model initialized with random weights versus using pre-trained weights to assess the potential benefits of transfer learning.

## 7.1 Experimental Setup

### 7.1.1 Training Phases

The training process was structured into two main phases, applied to both models:

1. **Pre-training:** During this initial phase, we utilized all datasets containing real CMBs—RODEJA, MOMENI, and VALDO—for both training and validation. Additionally, we included exclusively in the training set the following datasets:
   - sMOMENI: we employed this dataset as a form of data augmentation, providing additional microbleeds. It was used only on the training set because it was not possible to know with absolute certainty that the features shared across these synthetic microbleeds were identical to those observed in real CMBs. Nevertheless, it was hypothesized that at least some features would be shared with real CMBs, thereby conferring a benefit upon the model in learning characteristics common to true CMBs. This data was however not added to the validation set to avoid selecting a model for the next phase that prefers synthetic CMBs over real ones.
   - CRBneg: we hypothesized that including a substantial number of negative scans would assist the model in distinguishing features between healthy and unhealthy scans. However, since the selection relied on case-level annotations and regular expressions applied to radiological reports, we could not be completely certain of the scans' absolute negativity. Therefore, we confined their use exclusively to the pre-training phase and only in the training set.
2. **Fine-tuning:** This phase aimed to refine the features learned during pre-training, addressing potential discrepancies arising from the synthetic nature of sMOMENI and the partial accuracy of CRBneg. Taking the best model selected based on the F1-count score during the pre-training phase, extra training was performed using the same data split but removing sMOMENI and CRBneg from the training set.

Final models were evaluated on completely independent datasets, specifically DOU and CRB, as well as on the validation set used during training. The predictions underwent post-processing, as explained in section 6.6, before being evaluated using metrics described in section 6.7.

### 7.1.2 Data Splits

The different independent datasets were utilized at different stages of the model development process, as detailed in Table 7.1.

#### 7.1.2.1 Training and Validation Sets

A combination of datasets was selected to provide the model with sufficient demographic variation (country, age, pathological condition, sex), MRI data variation (scanner models, sequence types, field strength, TE, TR, flip angle, and slice thickness) and annotation protocol variation (BOMBS and MARS). This ensures the model is robust against data variations present in real clinical settings. The datasets were split as follows:

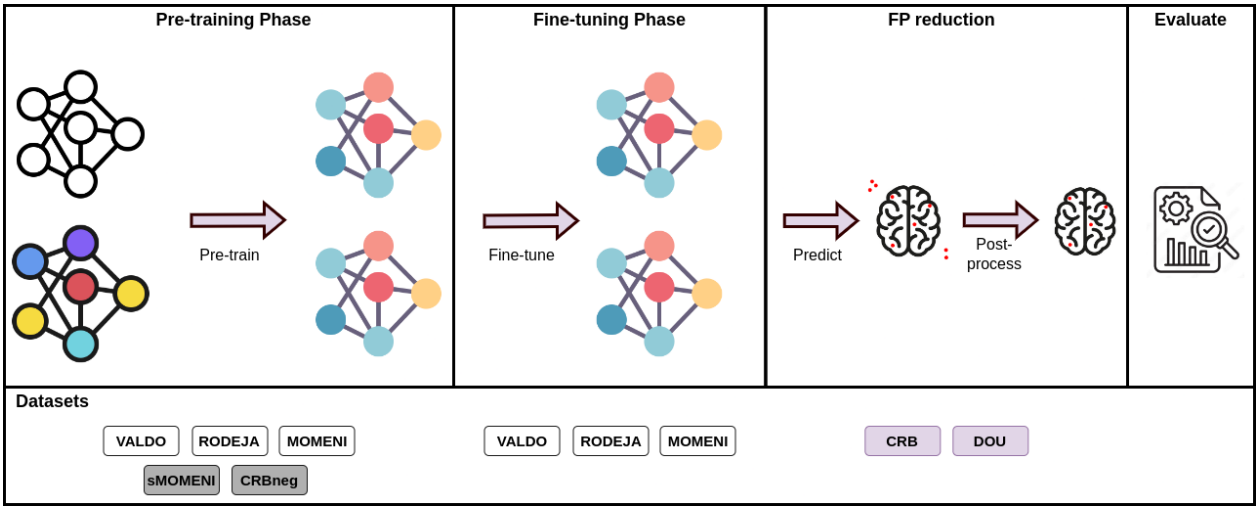



Figure 7.1: **General Overview of the Experimentation Workflow**. This figure illustrates the various models generated through distinct phases of the training process. A black-and-white graph represents a model initialized with random weights; a colored graph with black contours represents a model initialized with pre-trained weights; and colored graphs without contours indicate models that have undergone training during this project. The diagram identifies two distinct training phases of training, followed by a FP reduction and evaluation. The datasets used for each phase are indicated below.

- Training Set: The datasets MOMENI, RODEJA, VALDO and CRBneg, sMOMENI were incorporated into the training set for pre-training phase, and the first three for fine-tuning phase, as explained in section 7.1.1.
- Validation Set: Only MOMENI, RODEJA, and VALDO were also included in the validation split used for model selection. We will refer to this set of data as *VALID*

The scans present in MOMENI, RODEJA, and VALDO datasets were split with 75% for training and 25% for validation, later adding sMOMENI and CRBneg to the training set (thus decreasing the proportion of data in validation during pretraining). No patient was found in different splits, and all scans from the same patient were always included in the same split. This subject-wise splitting is crucial to prevent data leakage from the test set into the training process and ensure the integrity and reliability of our model [108]. Stratification was performed making sure both splits contained approximately equal proportions of the following:

- Field strength (1.5, 1.5/3 and 3 Tesla)
- Resolution level (low, high): scans considered having low resolution when both x and y voxel dimensions were over 0.5 mm
- Sequence type (SWI/T2S)
- Health status of scans (healthy/unhealthy): considered unhealthy if there was at least one CMB
- Number of CMBs per case ($<3$, $\geq 3$): for the unhealthy scans only
- Dataset of origin

#### 7.1.2.2 Test Sets

Two independent datasets of similar were selected exclusively for testing to ensure an unbiased evaluation:

Table 7.1: **Some Statistics and Characteristics of Datasets**. The following data is provided: (1) *# Scans (total, w/cmb, <3cmb)*—total number of scans, number with CMBs, and number with fewer than three CMBs; (2) *# Patients (total, w/cmb)*—total number of patients and number with at least one CMB; (3) *Res. level % (low, high)*—percentage of scans with low (both x and y voxel dimensions over 0.5 mm); (4) *Total CMBs*—total number of microbleeds detected; (5) *Avg # CMB per-scan*—average number of microbleeds per scan; (6) *CMB radius avg (mm)*—average radius of the microbleeds; (7) *Availability*—indicates whether the data is publicly available or not; (8) *Split*—specifies the data split used in the training and evaluation process of the model.

| **Dataset** | **Total scans** | **Scans w/cmb** ($_{1+}$, $_{<3}$) | **Patients** ($_{total}$, $_{w/cmb}$ | **Res. level**% ($_{low}$, $_{high}$) | **Total cmb** | **No. cmb per-scan** | **cmb radius** ($_{mm}$) | **Source** | **Split** |
|---|---|---|---|---|---|---|---|---|---|
| CRB | 18 | 18, 8 | 18, 18 | 44, 56 | 127 | 7.06 ± 7.53 | 1.8 ± 0.49 | Private | Test |
| CRBneg | 742 | 0, 0 | 742, 0 | 40, 60 | 0 | - | - | Private | Train |
| DOU | 20 | 20, 14 | 20, 20 | 0, 100 | 74 | 3.7 ± 3.99 | 1.48 ± 0.37 | Public | Test |
| MOMENI | 370 | 57, 46 | 118, 30 | 100, 0 | 146 | 2.56 ± 3.09 | 1.4 ± 0.2 | Public | Train-Valid |
| RODEJA | 103 | 61, 42 | 103, 61 | 15, 85 | 357 | 5.85 ± 10.08 | 1.26 ± 0.63 | Private | Train-Valid |
| VALDO | 72 | 50, 40 | 72, 50 | 38, 62 | 253 | 5.06 ± 12.41 | 1.5 ± 0.46 | Public | Train-Valid |
| sMOMENI | 3700 | 3700, 0 | 118, 118 | 100, 0 | 36812 | 9.95 ± 0.22 | 1.41 ± 0.17 | Public | Train |

- DOU: A public dataset, DOU, for benchmarking our method. It comprises high-resolution SWI scans with high field strength and decent slice thickness. The dataset is characterized by on average 2-3 CMBs per scan and has a total number of scans with a size very similar to most datasets used in automated CMB detection. It contains scans from stroke and normal patients and is also useful for testing the model's ability to differentiate between classes at the low numbers of CMBs range. The shape of these microbleeds typically exhibits sphericity and low elongation, aligning with the theoretical intrinsic characteristics of CMBs, thus making the dataset useful for verifying if the model can accurately infer these traits.

- CRB: Developed in-house, CRB offers a more representative sample of real clinical settings, characterized by high variability in demographic data, scanner models, sequence types, and MRI acquisition parameters. It is important to note that it includes scans with an extremely low slice thickness (up to 5.5), which the model has not encountered in training data containing CMBs. This dataset also features a wide diversity in terms of CMB size, shape, elongation, and maximum diameter. The CMBs are generally larger than those in most datasets, and the number of CMBs per scan is typically higher, which may challenge the model due to its unprecedented nature and lack of data augmentation implemented to account for this. Furthermore, this dataset enables us to dissect performance based on specific CMB characteristics and the presence of mimics and other metadata collected during our in-house annotations. Furthermore, half the scans present co-occurring pathological findings of hemorrhages and infarcts, enhancing the clinical representativeness.

### 7.1.3 Hyperparameter tuning

We performed some initial manual experimentation for hyperparameter tuning, yielding a final set of hyperparameters based on good convergence of validation metrics. The results were the following:

- **Patch Size**: patch_size = $[96, 96, 96]$ – This is the input size for the model, representing a volume of almost $5 \times 5 \times 5$ cm$^3$. This size should provide sufficient context of surrounding tissues for a microbleed with a theoretical maximum size of 1cm.

- **Number of Patches**: num_patches = 30 – We selected 30 patches from each scan when cycling over the training set to ensure diverse sampling.
- **Class Proportions**: class_props = $[0.1, 0.9]$ for background and CMB respectively – Set to ensure that about 50% of patches in the mini-batches contain non-background label. When applied to an also unbalanced dataset of 60% healthy scans and 40% unhealthy, should yield about 50% of patch samples having a CMB. This makes it less likely that the gradient is calculated across only background voxels, which could lead to getting trapped in a local minimum.
- **Overlap Fraction**: overlap_frac = 0.4 – Used during evaluation to deterministically extract 343 patches that overlap between them (compared to 125 with no overlap), helping to smooth the effects of patch boundaries in the model's predictions, and balanced against available computational resources.
- **Learning Rate**: learning_rate = $5 \times 10^{-5}$ – Observed to yield good convergence in reasonable time.
- **Loss Weighting**: $\lambda_1 = \lambda_2 = 1$ – Equal weighting of different losses.
- $FTL$ **Parameters**: $\alpha = \beta = 0.5$ – These settings proved to be more stable and yield better performance than having unbalanced weighting on the FPs and FNs, which made the training very unstable. Also $\gamma = 3$ – Increases the model's focus on learning from CMB voxels, which are challenging to learn as they rarely occur in the volumes even after active sampling of CMB patches due to huge class imbalance.
- $L_{CE}$ **Parameters**: $w_{background} = 0.1$ and $w_{CMB} = 5$ – Adjusts the loss function to emphasize learning from CMB voxels over the background to address the class imbalance.

## 7.2 Training Experiments

Each experiment was conducted until no further improvement was observed in the validation loss or F1-detection score. This resulted in varying completion times for each experiment. The following experiments were conducted:

- Pre-training of model from scratch: 197 epochs, taking 17 days to complete on 40GB GPUs. The training was stopped when validation loss remained static and F1-detection score started decreasing (overfitting). The model with the best F1-detection score was selected for fine-tuning. We call resulting model *Scratch-PreTrained*
- Pre-training of model with pre-trained weights: 108 epochs, taking 8 days to complete on 40GB GPUs. The training was stopped when validation metrics reached a plateau and training loss started decreasing. Model with best F1-detection score was selected for fine-tuning. We call this model *TL-PreTrained*.
- Fine-tuning of *Scratch-PreTrained*: 187 epochs, taking 17 days to complete on 40GB GPUs. The training was stopped when for 50 epochs no change was observed in either training or validation loss. We selected the model with the best validation loss and called it **Scratch-PreTrained-FineTuned**, or **Scratch** for short.
- Fine-tuning of *TL-PreTrained*: 99 epochs, taking 8 days to complete on 40GB GPUs. Training was stopped when for 100 epochs no change was observed in either training or validation loss. We selected the model with the best validation loss and called it **TL-PreTrained-FineTuned**, or **TL** for short.

An important aspect to note is that model selection was based on raw metric values over time, without smoothing. Therefore, the selection process may be susceptible to random fluctuations, although these fluctuations were not substantial. We kept track of how the model was performing through training both analytically, but also by creating plots of specific predicted or detected CMB, with a strong emphasis on false calls being made. This is illustrated in 7.2 , where we see the model caught the specific CMB first at epoch 13 and then progressively adapted the shape predicted to fully cover the CMB intensity profile at epoch 116.

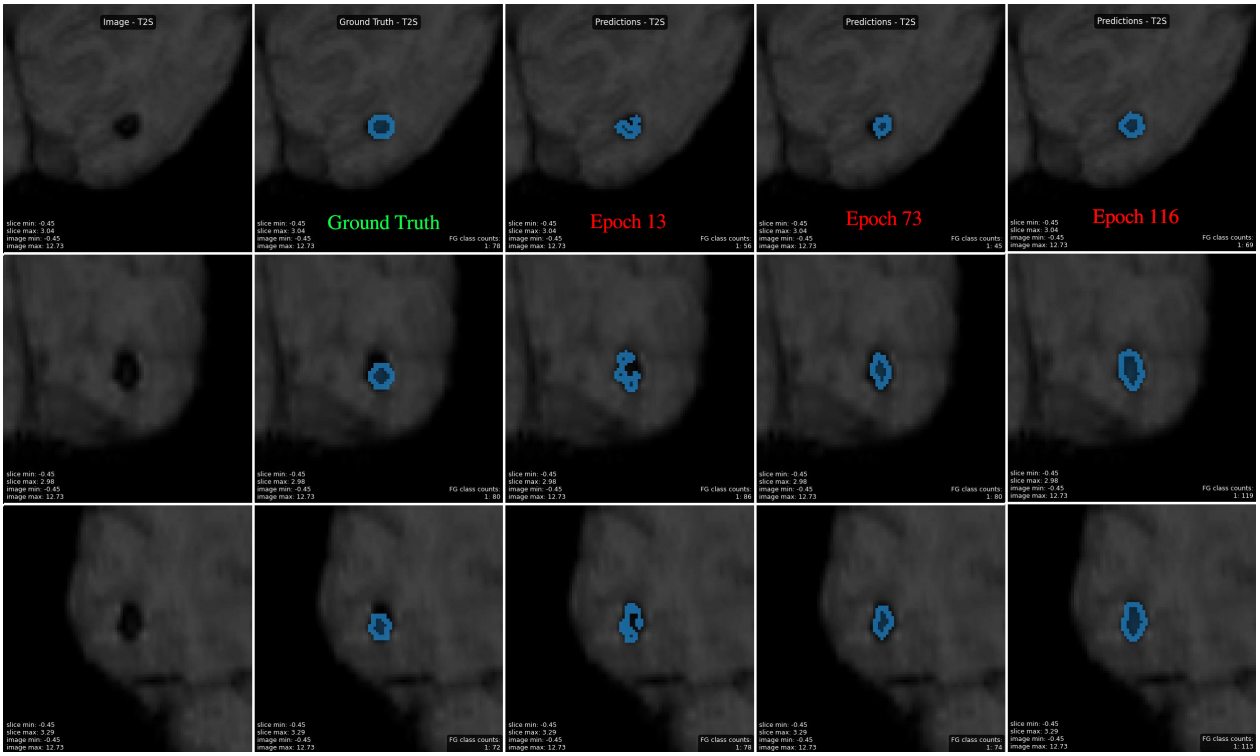



Figure 7.2: **Example of CMB Prediction Evolution Over Time**. This figure illustrates the evolution of model predictions on a selected CMB of MOMENI dataset in the validation phase during the training of the model. The rows display slices of the 3D volume in axial, coronal, and sagittal cuts in respectively. The first column features the raw image, the second column depicts the ground truth processed annotation, and the subsequent columns show predictions at various epochs. As can be seen, the model first predicts the CMB at epoch 13 and then increasingly gets better at adapting to the CMB intensity profile, even if that means deviating from the ground truth segment map, which does not fully cover the microbleed.

### 7.2.1 MLOps Infrastructure

All training, predictions, and evaluations were facilitated through an MLOps framework using ClearML (version 1.13.1) at CEREBRIU. Due to proprietary restrictions, the codebase cannot be made public. This framework serves as a scaffold, enabling the integration of project-specific code with general code, and includes scripts for training, monitoring performance during training, evaluating on test sets, and general prediction tasks. The project's code was integrated within this framework to leverage the computational resources managed by the framework. Deep learning code was implemented in TensorFlow 2.14.0 with CUDA 11.4, using Python 3.11.5. The framework provides access to three different Linux servers used for experiments, each with specific GPU specifications:

- Server 1: Equipped with NVIDIA A100-SXM4 (40GB) GPUs and 128 CPU cores.
- Server 2: Equipped with NVIDIA GeForce RTX 3090 (24GB) GPUs and 60 CPU cores.
- Server 3: Equipped with Quadro RTX 5000 (16GB) GPUs and 40 CPU cores.

The framework allows for the tracking of all experiments and configurations used for every experiment, all of which were included in the projects' GitHub repository for reproducibility and replicability. Additionally, it supports a GUI that allows real-time tracking of various metrics including validation metrics, GPU and CPU usage, console output, debug sample plots, and system information. Figure 7.3 shows how it looks for a selected experiment.


Figure 7.3: **Example of a ClearML Training Task**. This shows how a training task appears in the MLOps framework. The "Scalars" tab is selected, displaying four monitored metrics over time: F1-detection, F1-segmentation, validation and training loss, and GPU performance metrics.

## 7.3 Evaluation of models

This section compares the performance of the two final models: *Scratch-Pretrained-Finetuned*, initialized with random weights; and *TL-Pretrained-Finetuned*, initialized with pre-trained weights from the Apollo model. Both models were evaluated on reserved test sets, which were not used during either the training or model selection phases, to ensure an unbiased evaluation. Testing also included an additional evaluation on the validation set. Predictions were post-processed as detailed in Section 6.6. This section identifies the best-performing model based on aggregated metrics for every test set, without splitting by subgroups. The selected model will be analyzed in more detail later on in chapter 8.

#### 7.3.0.1 Detection performance

First, we will consider the results per scan, where every scan has equal relevance but not every CMB. The distribution of detection metrics results obtained at the scan level is depicted for both models across datasets in Figure 7.4, while the *scan* values are presented in Table 7.2. By looking at the estimated distribution in the violin plots, it appears that both models have very similar performance distributions across scans for each dataset. However, *Scratch* seems to concentrate the probability mass more around 1 for all metrics compared to *TL*. Notably, when looking at the *scan*F1 score on the CRB dataset, which represents a more challenging scenario, we see that *TL* has a slightly upward-shifted probability distribution compared to *Scratch*, confirmed by a higher *scan*F1. This could indicate that *TL* is more robust than *Scratch*. However, this is only observed for the *scan*F1 score, while *scan*Precision and *scan*Recall individually appear to be worse. This suggests that the *scan*F1 distribution of scores might be flawed, and by chance, the scans where F*scan*1 could be computed (i.e., where neither Precision nor Recall were None) had better *scan*F1 scores, which may not be representative of the whole dataset. For this reason, we must also look at *scan*F1$^*$, where *Scratch* has a higher number. Overall, ***Scratch* outperforms *TL* in the detection of CMBs**. We observe a high variability in the scores of the *scan*-averages in the table 7.2 are very high, with standard deviations reaching up to 0.3 points. This indicates that performance varies considerably between scans, and should be interpreted with caution.

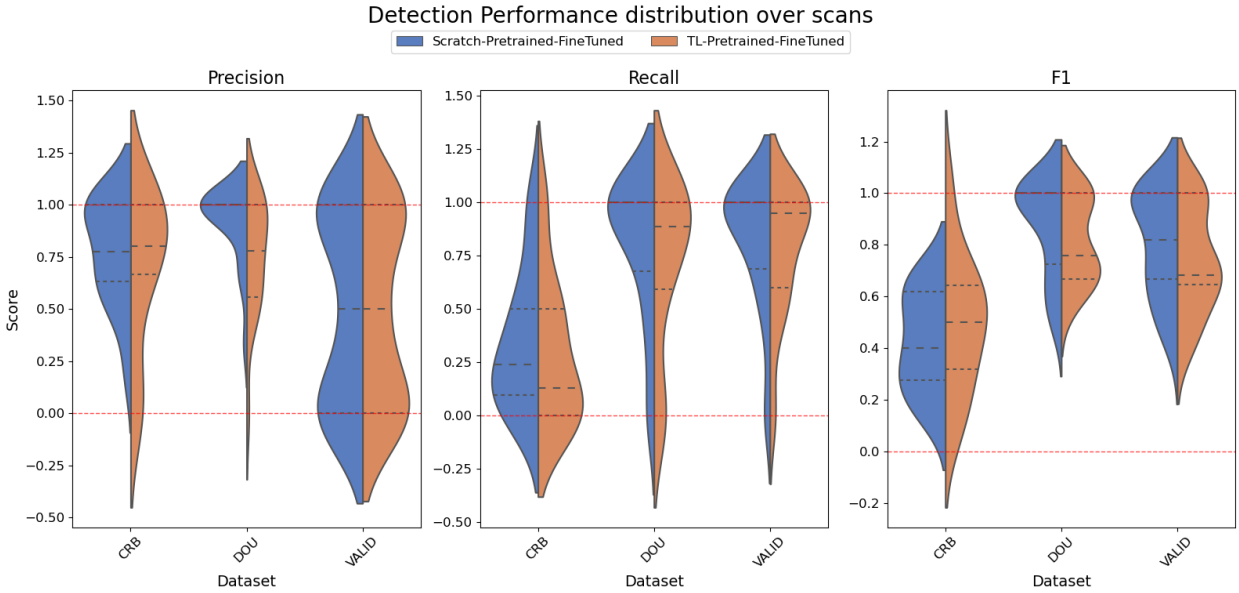



Figure 7.4: **Comparison of Performance Metrics Across Models and Test Sets.** This figure presents half-violin plots of precision, recall, and F1 scores for two models: Scratch-Pretrained-FineTuned (left half) and TL-Pretrained-FineTuned (right half), evaluated across CRB and DOU test datasets, as well as on a validation set. Upper plots are generated using a Kernel Density Estimation (KDS) with a cut-off of 2 times the observed bandwidth, extending the density curve beyond the real data range for visual purposes. A red dashed line has been drawn to clearly define the real range of the variables.

The same conclusion can be drawn when considering *glob*-averages, presented in Table 7.3. Here, each individual object or CMB has equal relevance, and the focus is on the absolute number of CMBs missed and detected, with *Scratch* emerging as the clear winner. Overall, the models seem to struggle more with recall than precision, meaning they tend to underpredict rather than overpredict, which is contrary to what is often observed in the literature. For instance, the FP rates per scan and per CMB are incredibly low compared to published methods, where disproportionate values of FPs are frequently reported.

Table 7.2: **Detection Results of Different Models *scan*-averaged**. This table shows the precision, recall, F1 score, F1* (harmonic mean of *scan*- Precision and Recall), false positives per CMB (FPcmb), and false positives per scan (FPscan) for each model evaluated on different datasets. Standard deviations are included to reflect variability within the test set. Highlighted in bold are the best metric results for every combination of dataset, model, and metric type.

| Dataset | Model | Precision | Recall | F1 | F1* | FPcmb | FPscan |
|---|---|---|---|---|---|---|---|
| CRB | Scratch | $\mathbf{0.777 \pm 0.252}$ | $\mathbf{0.343 \pm 0.322}$ | $0.426 \pm 0.193$ | **0.476** | $\mathbf{0.355 \pm 0.948}$ | $\mathbf{0.889 \pm 1.451}$ |
| CRB | TL | $0.683 \pm 0.378$ | $0.287 \pm 0.34$ | $\mathbf{0.489 \pm 0.259}$ | 0.404 | $0.383 \pm 0.957$ | $1.056 \pm 2.014$ |
| DOU | Scratch | $\mathbf{0.912 \pm 0.187}$ | $\mathbf{0.794 \pm 0.338}$ | $\mathbf{0.87 \pm 0.185}$ | **0.849** | $\mathbf{0.153 \pm 0.457}$ | $\mathbf{0.5 \pm 1.235}$ |
| DOU | TL | $0.749 \pm 0.28$ | $0.704 \pm 0.393$ | $0.814 \pm 0.162$ | 0.726 | $0.402 \pm 0.716$ | $1.15 \pm 1.872$ |
| Valid | Scratch | $\mathbf{0.501 \pm 0.45}$ | $\mathbf{0.81 \pm 0.304}$ | $\mathbf{0.803 \pm 0.204}$ | **0.619** | $\mathbf{0.429 \pm 0.748}$ | $\mathbf{1.7 \pm 2.803}$ |
| Valid | TL | $0.485 \pm 0.437$ | $0.765 \pm 0.308$ | $0.737 \pm 0.203$ | 0.594 | $0.53 \pm 0.897$ | $1.75 \pm 2.771$ |

Table 7.3: **Detection Results of Different Models *glob*-averaged**. This table shows the precision, recall, F1 score, and false positives per CMB (FPcmb) for each model evaluated on different datasets. Standard deviations cannot be provided, as information is computed once from the final confusion matrix. Also, FPscan is not reported, as its value is mathematically the same as when *scan*-averaged

| Dataset | Model | Precision | Recall | F1 | FPcmb |
|---|---|---|---|---|---|
| CRB | Scratch | **0.704** | 0.299 | 0.42 | **0.126** |
| CRB | TL | 0.672 | **0.307** | **0.422** | 0.15 |
| DOU | Scratch | **0.836** | 0.689 | **0.756** | **0.135** |
| DOU | TL | 0.693 | **0.703** | 0.698 | 0.311 |
| Valid | Scratch | **0.511** | **0.732** | **0.602** | **0.701** |
| Valid | TL | 0.496 | 0.711 | 0.585 | 0.722 |

#### 7.3.0.2 Segmentation performance

In Table 7.4, we observe the segmentation performance of both. Consistently, the ***TL* consistently outperforms the *Scratch* model in terms of** $Dice_{TP}$. It is important to note that a good Dice score indicates alignment with the ground truth segmentation when detecting the CMB. However, the ground truth has undergone a crucial preprocessing step—resampling—that can significantly affect its final shape and size. If the original resolution is not sufficiently high across some dimensions, resampling with nearest neighbor interpolation can result in geometrical shapes with sharp edges, leading to less smooth contours. Additionally, insufficient slice thickness and low z-axis resolution can produce unwanted tubular shapes in the ground truth. Consequently, the final masks may not fully correspond to the real microbleed intensity profile observed in the raw image. Thus, achieving a high Dice score does not necessarily mean having a better segmentation. Moreover, the Dice score value can vary significantly with just a few pixels' deviation (see Section 6.7).

This issue becomes evident when examining Figure 7.5, which compares predictions from both models for a selected microbleed. In this figure, we see that the TL model aligns better with the real intensity profile of the microbleeds, often better than the ground truth mask, perfectly covering the lesions but deviating from the ground truth mask, especially in the coronal and sagittal planes where the ground truth is very off. Despite this, the *TL* model achieves a relatively low Dice score of 0.65, as it deviates from the tubular and unnatural shape in the GT, where those few voxels have a significant impact on the score. The *Scratch* model under-segments in this case but also identifies the correct center of the CMB. This behavior is consistent across all scans and CMBs, with the *TL* model aligning better with the real CMB signal void (and with the GT too in most cases, as in the example provided) than both the *Scratch* model and the ground truth masks. However, the primary goal of our project is to correctly detect the CMB, a task at which *Scratch* appears to perform better than *TL*. Therefore, the superior segmentation performance of *TL* may not be as relevant for our objectives.

Table 7.4: ***scan*- and *glob*-averaged results for** $Dice_{TP}$. This table shows the *scan*- and *glob*-averaged Dice scores computed for TPs of each model evaluated on different datasets. Standard deviations are included to reflect variability within the test set. Also, highlighted in bold are the best metric results for every combination of Dataset and metric type, between both models.

| **Dataset** | **Model** | $scanDSC_{TP}$ | $globDSC_{TP}$ |
|---|---|---|---|
| CRB | Scratch | $0.41 \pm 0.20$ | 0.42 |
| CRB | TL | $\mathbf{0.49 \pm 0.18}$ | **0.53** |
| DOU | Scratch | $0.56 \pm 0.20$ | 0.58 |
| DOU | TL | $\mathbf{0.65 \pm 0.14}$ | **0.66** |
| Valid | Scratch | $0.60 \pm 0.18$ | 0.64 |
| Valid | TL | $\mathbf{0.63 \pm 0.16}$ | **0.66** |

#### 7.3.0.3 Image-level Classification Performance

We defined three different cutoffs at 3, 5, and 11 CMBs to evaluate the model's ability to provide a global count of microbleeds that can help classify scans into distinct groups. The motivation for these thresholds is clinical:

- Having >2 CMBs has been demonstrated to elevate the risk of parenchymal haemorrhage 24 hours after intravenous thrombolysis and to predict an unfavorable clinical outcome independently [4].
- The presence of >4 microbleeds is associated with cognitive decline [5].
- A high CMB burden (>10) significantly impacts the outcomes of reperfusion therapies like intravenous thrombolysis (IVT) for acute ischemic stroke (IS). However, authors warn that CMBs should be factored into a holistic individual-based risk-benefit analysis to determine treatment suitability, rather than serve as go/no-go criteria [20].

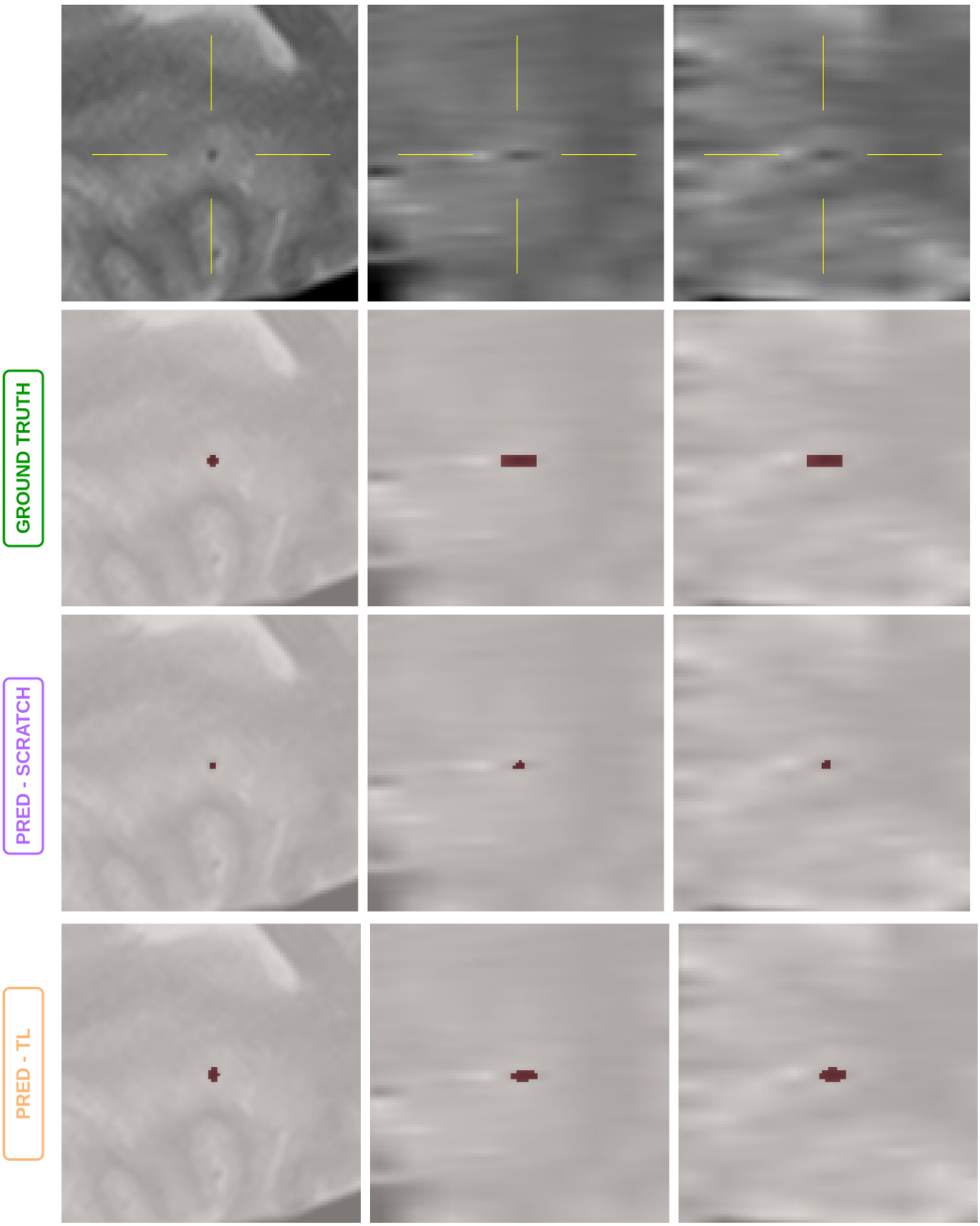



Figure 7.5: **Visual comparison of predictions for a selected CMB**. Each column represents the axial, coronal, and sagittal cuts from a 96x96x96 $mm^3$ 3D patch centered around the true CMB center of mass. Rows show the raw image, the ground truth annotation, and the predictions by the Scratch and TL models. This microbleed is 145 voxels in size, with Dice scores of 0.22 and 0.65 for the Scratch and TL models, respectively. The CMB is located in the Cortex/grey-white junction, and the sequence type is T2S. The original scan dimensions and voxel size are [272, 341, 287] and [0.43, 0.43, 6].

Results for all datasets are shown in Table 7.5. To interpret these results, it is important to note the following: CRB has eleven scans with 3 or more CMBs, eight with 5 or more, and six with 10 or more; DOU has eight, five, and three scans in each group respectively; and the validation set has eight, six, and three. Overall, performance is quite good for both models, with classification using a threshold of 3 being the most challenging, and an F1 score never going lower than 0.63.

The observed variability in performance across different thresholds can be first attributed to the sensitivity of the classification to the number of CMBs present. For lower thresholds, even a single FP or FN can significantly impact the classification outcome. For instance, in datasets like DOU and VALID, where most scans typically contain 2-3 CMBs as analyzed in Chapter 4, a single detection error could reclassify a scan, affecting precision and specificity respectively. This sensitivity is less pronounced at higher thresholds where the allowable margin for detection errors increases. Consequently, we must take this into account when analyzing results. Based on results, the ***TL* model consistently outperforms or equates the *Scratch* model across all classification metrics**.

Table 7.5: **Classification Performance Results at Different Thresholds**. This table shows the precision, recall, F1 score, and specificity for each model evaluated on different datasets and thresholds. Highlighted in bold are the best metric results for every combination of dataset, threshold, and metric type between both models. Classification is binary, with the positive group being scans with $\geq$ threshold.

| Dataset | Threshold | Model | Precision | Recall | F1 | Specificity |
|---|---|---|---|---|---|---|
| CRB | 3 | Scratch-Pretrained-FineTuned | 0.5 | 0.857 | 0.632 | 0.455 |
| CRB | 3 | TL-Pretrained-FineTuned | **0.5** | **0.857** | **0.632** | **0.455** |
| CRB | 5 | Scratch-Pretrained-FineTuned | 0.571 | 0.8 | 0.667 | 0.25 |
| CRB | 5 | TL-Pretrained-FineTuned | **0.643** | **0.9** | **0.75** | **0.375** |
| CRB | 11 | Scratch-Pretrained-FineTuned | 0.706 | 1 | 0.828 | 0.167 |
| CRB | 11 | TL-Pretrained-FineTuned | **0.706** | **1** | **0.828** | **0.167** |
| DOU | 3 | Scratch-Pretrained-FineTuned | 1 | 0.833 | 0.909 | 1 |
| DOU | 3 | TL-Pretrained-FineTuned | **1** | **0.833** | **0.909** | **1** |
| DOU | 5 | Scratch-Pretrained-FineTuned | 0.882 | 1 | 0.938 | 0.6 |
| DOU | 5 | TL-Pretrained-FineTuned | **1** | **0.933** | **0.966** | **1** |
| DOU | 11 | Scratch-Pretrained-FineTuned | 0.895 | 1 | 0.944 | 0.333 |
| DOU | 11 | TL-Pretrained-FineTuned | **0.895** | **1** | **0.944** | **0.333** |
| Valid | 3 | Scratch-Pretrained-FineTuned | 0.972 | 0.932 | 0.952 | 0.75 |
| Valid | 3 | TL-Pretrained-FineTuned | **0.972** | **0.932** | **0.952** | **0.75** |
| Valid | 5 | Scratch-Pretrained-FineTuned | 1 | 0.987 | 0.993 | 1 |
| Valid | 5 | TL-Pretrained-FineTuned | 0.987 | **0.987** | 0.987 | 0.833 |
| Valid | 11 | Scratch-Pretrained-FineTuned | 0.987 | 0.975 | 0.981 | 0.667 |
| Valid | 11 | TL-Pretrained-FineTuned | **0.987** | 0.962 | 0.974 | **0.667** |

# 8 Performance analysis

Based on the results shown in section 7.3, we selected *Scratch* as the best performing model. *Scratch* showed clearly superior performance in terms of detection, although it was not as good as *TL* in terms of real clinical value in discriminating between groups. The latter, even if consistent for all groups, was never too far from the performance of *Scratch*, while for detection the superiority of *Scratch* is evident. Moreover, even if *TL* showed a better adaptation to the CMB intensity profile than the GT mask, with a better segmentation performance, segmentation is only a proxy task for detection, which was the real goal. This chapter expands the evaluation of this model to a more granular level, analyzing predictions from the model for different subgroups. We conduct a detailed analysis of performance based on patient-, scan-, and CMB-level metadata of our test sets to better understand the challenges associated with the task and data. Finally, results are compared to state-of-the-art (SOTA) methods for this task, highlighting the limitations of such comparisons.

## 8.1 Scan characteristics

Various scan-level factors could potentially influence our predictions. While we attempt to identify some trends and draw preliminary conclusions, it is crucial to approach these interpretations with caution. The interplay of factors affecting performance is complex and challenging to disentangle. In the subsequent sections, we will explore these factors individually or in minimal combination with other impacts on the performance of the model.

### 8.1.1 In-plane Resolution and Slice Thickness

We conducted an analysis comparing low and high in-plane resolutions (ranging between 0.3mm and 1mm) and found no notable differences in performance metrics.

We identified slice thickness—defined by the voxel dimension along the z-axis or axial axis of the original image before resampling—as a potential key determinant of detection accuracy. We divided the datasets into two groups based on slice thickness: those with a thickness of 3mm or greater, and those with less. This categorization showed significant performance variability: in the CRB dataset, the *glob*F1 scores were 0.286 for slices 3mm and 0.536 for slices <3mm; in the Validation dataset, the scores were 0.800 for slices 3mm and 0.573 for slices <3mm. We hypothesize that this model failure is likely due to the elongated shapes that CMBs acquire following the resampling process when slice thickness is inadequate. This lack of information leads to poor interpolation quality between points in different slices (as the slices are distant and all information in the middle is non-existent), resulting in elongated, tubular shapes as demonstrated in Figure 7.5. Overall, it seems that **anisotropy negatively impacts the model's ability to accurately identify CMBs when the slice thickness is large**.

### 8.1.2 Number of CMB

We then decided to investigate the effect of removing scans with many CMBs from the analysis. The idea is to eliminate the potential impact of a few scans dominating the overall counts and metrics, as accurate detection when there are so many microbleeds has less clinical utility, while accuracy in detecting smaller numbers of CMBs is most important. Additionally, in [23], they identified that one of the primary reasons for disagreement among experienced CMB raters is the presence of multiple CMBs, as well as them being pale or small. This suggests that GT annotations could potentially contain errors in cases with numerous CMBs.

We excluded scans containing more than 10 CMBs from the evaluation to investigate the effect of containing many CMBs. This exclusion affected a total of 6 scans from the CRB dataset, 3 from the DOU dataset, and 4 from the validation set. Following this adjustment, we observed slight improvements in most datasets.

Specifically, the new *scan*F1* scores were 0.532 ($\Delta = +0.129$) for the CRB dataset and 0.877 ($\Delta = +0.0282$) for the DOU dataset. Conversely, the VALID dataset experienced a slight reduction in its score, moving to 0.600 ($\Delta = -0.019$). These results could suggest that scans with numerous microbleeds could pose challenges for the model. Nevertheless, it is difficult to distinguish whether this is due to statistical effects, as a percentage of model and/or annotation errors could lead to larger absolute differences in counts in cases with many CMBs.

### 8.1.3 Sequence type

We aim to create a sequence-agnostic model that works well with both T2S and SWI sequences. Therefore, the train-validation splits contained equal proportions of both sequence types, and we are interested in seeing the difference in CRB test set for both types, which is shown in Table 8.1.

When analyzing the performance metrics for SWI versus T2S, we observed distinct trends across datasets. Specifically, the CRB dataset showed a globF1 score of 0.555 for SWI and 0.286 for T2S, while the validation set demonstrated an opposite pattern, with globF1 scores of 0.563 for SWI and 0.867 for T2S. Considering that this trend does not occur in the validation set, it suggests the presence of other influencing factors. Further scrutiny of the T2S studies in the CRB dataset revealed that, although the in-plane resolution is similar to that in the validation set, the slice thickness significantly differs, ranging between 5 and 6 mm compared to a maximum of 4 mm in the validation set, where most scans, both T2S and SWI, are under 3 mm.

Additionally, the T2S sequences in the CRB dataset include 4 out of 5 scans with high CMB counts: 18, 12, 12, and 13 CMBs, indicating that half the group of T2S consists of edge cases in terms of the number of CMBs. When breaking down the CRB-T2S data into scans with fewer than 3 CMBs and those with 3 or more, we find that for the latter five out of 6 scans have a total of 9 TPs and 52 FN calls. Overall, we observe 10 FN/scan and 0.8 FN/CMB for T2S, and 6 FN/scan and 0.5 FN/CMB for SWI in the CRB dataset. FPs are less abundant in all cases. This suggests that the primary issue affecting performance is the high number of FNs rather than FPs, contrary to common observations in the literature. Overall, we believe that slice thickness and the number of microbleeds are the critical factors influencing the worse performance on the T2S in CRB, rather than the sequence type.

Table 8.1: **Detection Performance by Sequence Type and Number of CMBs**. This table shows the TP, FP, FN, *scan*-averages for Precision, Recall, F1, F1* and globF1; for each dataset, split by MRI sequence type and CMB level. CMB level is categorized as high when there are 3 or more CMBs, and low otherwise. Scores are presented as mean±std, if no std shown for a scan-averaged metric, it is because sample size wwas 1

| Dataset | Seq. | cmb | TP | FP | FN | scanPrecision | scanRecall | scanF1 | scanF1* | globF1 |
|---|---|---|---|---|---|---|---|---|---|---|
| CRB | SWI | $3^+$ | 24 | 7 | 29 | $0.875 \pm 0.182$ | $0.415 \pm 0.262$ | $0.494 \pm 0.213$ | 0.563 | 0.571 |
| CRB | SWI | <3 | 1 | 0 | 4 | 1.0 | $0.167 \pm 0.236$ | 0.5 | 0.286 | 0.333 |
| CRB | T2S | $3^+$ | 9 | 3 | 52 | $0.817 \pm 0.171$ | $0.152 \pm 0.057$ | $0.25 \pm 0.079$ | 0.256 | 0.247 |
| CRB | T2S | <3 | 4 | 6 | 4 | $0.55 \pm 0.332$ | $0.5 \pm 0.447$ | $0.542 \pm 0.16$ | 0.524 | 0.444 |
| DOU | SWI | $3^+$ | 33 | 7 | 21 | $0.903 \pm 0.156$ | $0.645 \pm 0.23$ | $0.726 \pm 0.173$ | 0.753 | 0.702 |
| DOU | SWI | <3 | 18 | 3 | 2 | $0.917 \pm 0.207$ | $0.857 \pm 0.363$ | $0.942 \pm 0.151$ | 0.886 | 0.878 |
| Valid | SWI | $3^+$ | 44 | 23 | 19 | $0.723 \pm 0.242$ | $0.71 \pm 0.093$ | $0.695 \pm 0.136$ | 0.716 | 0.677 |
| Valid | SWI | <3 | 14 | 15 | 6 | $0.532 \pm 0.458$ | $0.786 \pm 0.384$ | $0.794 \pm 0.236$ | 0.635 | 0.571 |
| Valid | T2S | $3^+$ | 6 | 1 | 0 | 0.857 | 1.0 | 0.923 | 0.923 | 0.923 |
| Valid | T2S | <3 | 7 | 1 | 1 | $0.917 \pm 0.204$ | $0.917 \pm 0.204$ | $0.889 \pm 0.172$ | 0.917 | 0.875 |

## 8.2 CMB characteristics

### 8.2.1 Location

To explore whether specific brain locations posed greater challenges for the model, we conducted an analysis segmented by the anatomical locations defined in section 4.9.3. Due to insufficient numbers of CMBs in certain locations to draw reliable conclusions, we grouped all deep brain areas into one category. In the end, each group (lobar-cortical, lobar-subcortical, and deep) represents approximately one-third of the total CMBs. The results for each category are presented in Table 8.2.

Deep microbleeds in the CRB dataset exhibit slightly superior metrics, as evidenced by Table 8.2. This is supported by higher *scan*F1 scores along with consistently higher *scan*F1* and *glob*F1 compared to other locations within the dataset. A similar trend is observed in the validation set, although it is not the case for DOU dataset, where performance is robust across all locations

Table 8.2: **Performance Metrics by Dataset and Location**. This table shows the TP, FP, FN, *scan*-averages for Precision, Recall, F1, F1* and globF1. Scores are presented as mean±std.

| Dataset | Location | TP | FP | FN | scanPrecision | scanRecall | scanF1 | scanF1* | globF1 |
|---|---|---|---|---|---|---|---|---|---|
| CRB | Cortical | 16 | 11 | 40 | $0.664 \pm 0.441$ | $0.305 \pm 0.379$ | $0.555 \pm 0.325$ | 0.418 | 0.386 |
| CRB | Deep | 9 | 5 | 21 | $0.636 \pm 0.452$ | $0.392 \pm 0.456$ | $0.829 \pm 0.2$ | 0.485 | 0.409 |
| CRB | Subcortical | 8 | 6 | 32 | $0.567 \pm 0.464$ | $0.221 \pm 0.307$ | $0.591 \pm 0.248$ | 0.318 | 0.296 |
| DOU | Cortical | 21 | 7 | 6 | $0.808 \pm 0.327$ | $0.879 \pm 0.228$ | $0.859 \pm 0.2$ | 0.842 | 0.764 |
| DOU | Deep | 14 | 1 | 10 | $0.962 \pm 0.139$ | $0.678 \pm 0.386$ | $0.833 \pm 0.226$ | 0.795 | 0.718 |
| DOU | Subcortical | 14 | 4 | 8 | $0.810 \pm 0.386$ | $0.739 \pm 0.394$ | $0.860 \pm 0.223$ | 0.773 | 0.700 |
| Valid | Cortical | 16 | 30 | 7 | $0.227 \pm 0.386$ | $0.870 \pm 0.227$ | $0.817 \pm 0.214$ | 0.360 | 0.464 |
| Valid | Deep | 26 | 17 | 4 | $0.632 \pm 0.474$ | $0.871 \pm 0.312$ | $0.956 \pm 0.096$ | 0.732 | 0.712 |
| Valid | Subcortical | 25 | 25 | 19 | $0.521 \pm 0.475$ | $0.589 \pm 0.385$ | $0.796 \pm 0.207$ | 0.553 | 0.532 |

### 8.2.2 Size and shape

This section aims to explore the possible influences of shape and size CMB features on the model's performance, by calculating the same features as explained in section 4.9.2. In Table 8.3, we present the mean and standard deviation of various CMB characteristics for the CMB grouped by the call (TP/FP/FN) and type (GT/prediction) for every dataset.

For the DOU dataset, the GT CMBs have similar distributions of shape features in both the TPs and the FNs, suggesting that the shape of the CMB does not significantly affect the model's ability to detect CMBs in this data. Interestingly, the FPs predicted by the model exhibit a very flat shape (InvFlatness of 0.48) and a considerably smaller volume (2.98 $mm^3$) compared to GT CMBs (15 $mm^3$) and predicted TPs (9 $mm^3$). This observation prompted further investigation, revealing that model occasionally predicted more than one CMB in the same space as illustrated in Figure 8.1, where we see an extra FP next to a TP prediction (example 1), and two FPs in the same area (example 2). This duplication could explain the higher number of FPs and the lower average size, and motivates to do extra post-processing, as this really skews the metrics negatively when it should not.

In the case of the CRB dataset, we clearly observe that the FNs exhibit different shape characteristics compared to the TPs. The Elongation and Flatness for FNs are 0.58 and 0.50, respectively, while these values are 0.70 and 0.59 for the TPs. Additionally, the maximum diameter of the FNs is larger, measuring 6.1 mm compared to 5.38 mm for the detected CMBs. This suggests that the missed GT CMBs have a more tubular shape due to insufficient axial resolution, which causes resampling to create a sharper, elongated shape between the two slices where the CMBs are annotated during resampling. This can be seen in the previous Figure 4.3, when looking at CRB datasets and CMB with a radius of 1.95mm. This phenomenon is clearly illustrated in Figure 8.2, where models fails to predict, probably due to the shape which is very similar to that of veins in a good resolution setup. To visually inspect the impact of shape on incorrect predictions, Figure 8.3 presents pairs of GT and predicted images for the CRB dataset. It is important to note that

Table 8.3: **Shape Features Comparison for Different CMB-Call Types and Datasets**. This table shows the mean and standard deviation for various shape features across different call types: FN, FP and TPs. Column "type" indicates whether CMBs are from GT or predicted. All values are either unit-less or in mm/mm3

| Call | Type | Dataset | n | InvElongation | InvFlatness | Max Diameter | Volume | Sphericity |
|---|---|---|---|---|---|---|---|---|
| FN | GT | CRB | 89 | 0.58 ± 0.25 | 0.50 ± 0.20 | 6.10 ± 2.07 | 30.88 ± 29.68 | 0.78 ± 0.06 |
| | | DOU | 23 | 0.88 ± 0.11 | 0.78 ± 0.15 | 3.71 ± 0.97 | 15.46 ± 10.82 | 0.85 ± 0.02 |
| | | Valid | 26 | 0.58 ± 0.24 | 0.43 ± 0.25 | 5.75 ± 4.35 | 45.57 ± 144.41 | 0.74 ± 0.11 |
| FP | Pred | CRB | 16 | 0.72 ± 0.18 | 0.52 ± 0.28 | 2.77 ± 0.83 | 4.77 ± 4.29 | 0.82 ± 0.08 |
| | | DOU | 10 | 0.70 ± 0.12 | 0.48 ± 0.25 | 2.42 ± 0.60 | 2.98 ± 2.98 | 0.82 ± 0.06 |
| | | Valid | 68 | 0.68 ± 0.18 | 0.56 ± 0.18 | 3.32 ± 0.94 | 5.87 ± 4.17 | 0.82 ± 0.06 |
| TP | GT | CRB | 38 | 0.70 ± 0.25 | 0.59 ± 0.18 | 5.39 ± 1.53 | 25.72 ± 27.38 | 0.78 ± 0.05 |
| | | DOU | 51 | 0.85 ± 0.11 | 0.77 ± 0.12 | 3.79 ± 0.67 | 15.45 ± 14.13 | 0.84 ± 0.03 |
| | | Valid | 71 | 0.60 ± 0.22 | 0.49 ± 0.17 | 5.23 ± 2.09 | 16.30 ± 18.37 | 0.77 ± 0.07 |
| | Pred | CRB | 38 | 0.76 ± 0.13 | 0.64 ± 0.21 | 3.80 ± 1.33 | 14.73 ± 14.95 | 0.85 ± 0.04 |
| | | DOU | 51 | 0.83 ± 0.10 | 0.67 ± 0.15 | 3.32 ± 0.77 | 9.21 ± 6.13 | 0.86 ± 0.05 |
| | | Valid | 71 | 0.67 ± 0.19 | 0.56 ± 0.21 | 3.96 ± 1.84 | 11.75 ± 14.07 | 0.84 ± 0.06 |

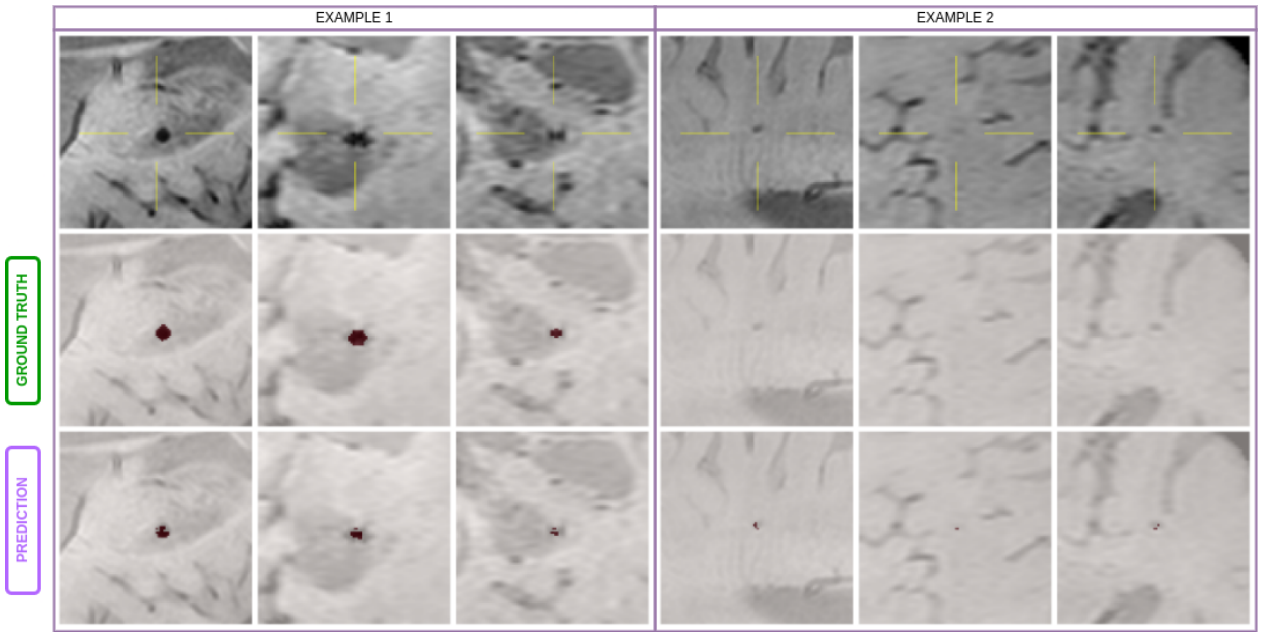



Figure 8.1: **Several FP examples in DOU dataset, SWI** For each example, each column represents the axial, coronal, and sagittal cuts from a 96x96x96 $mm^3$ 3D patch centered around the true CMB center of mass. Rows show the raw image, the ground truth annotation, and the predictions. We see for example 1 that two unique connected components are generated by the model, which due to the use of distance-based metric will cause both a TP and a FP to be present (as opposed to using overlap-based, which would actually cause 2 TPs; with one extra not desired). In example two there is no GT CMB, and two FPs are generated instead of 1. The scans are SWI, with an original 1mm slice thickness and in-plane resolution of 0.45x0.45mm2

these images are 2D snapshots of 3D renders, which can sometimes be misleading—for instance, a cylinder viewed from one end might appear spherical. Despite this limitation, we observe a notable occurrence of tubular-shaped CMBs that are incorrectly identified as FNs.

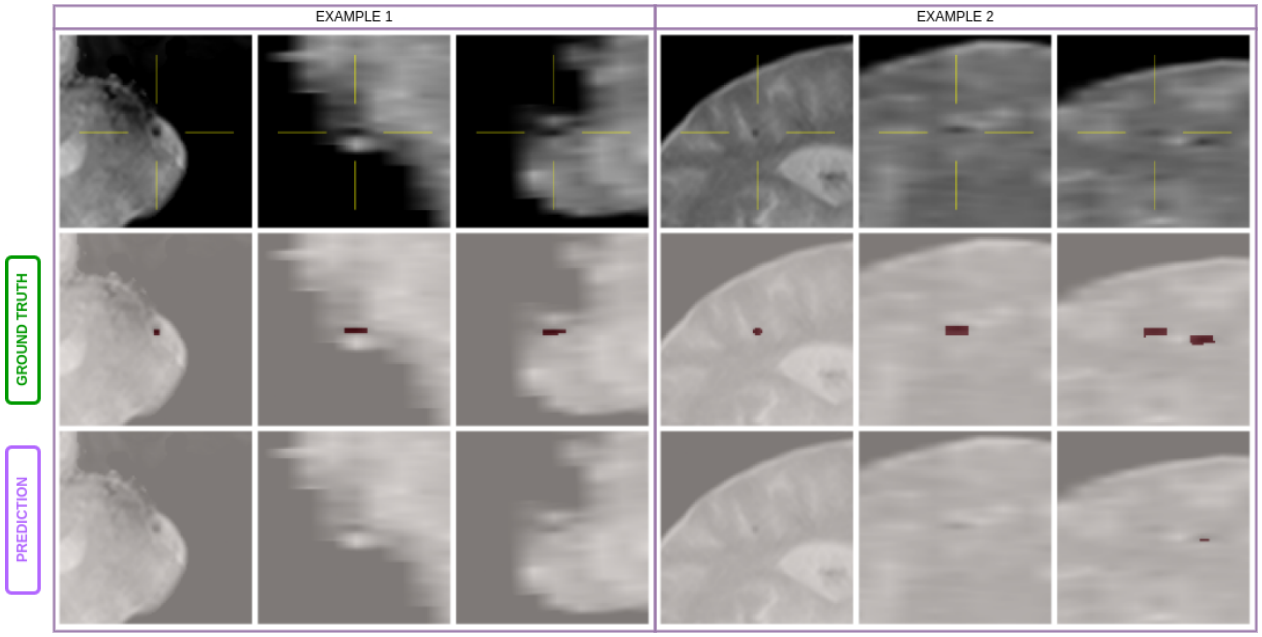



Figure 8.2: **Several FN examples in CRB dataset** For each example, each column represents the axial, coronal, and sagittal cuts from a 96x96x96 $mm^3$ 3D patch centered around the true CMB center of mass. Rows show the raw image, the ground truth annotation, and the predictions. In both examples we see that resolution in the coronal and sagittal planes is very poor, where GT CMB has an elongated shape. Both patches are from the same scan, which is T2S, with an original 6mm slice thickness and in-plane resolution of 0.45x0.45mm2. CMBs have InvElongation of 0.356 and 0.246 for example 1 and 2.

## 8.3 Comparison with published literature

Comparing results for automated CMB detection is a challenging task. There is no clear benchmark dataset, and each paper reports different metrics without any clear standardized set of metrics, lacking complementary metrics, and many times computed on private datasets. Evaluation is frequently conducted on internal test split of a certain cohort, rather than on separate cohorts, which may not reflect model performance on other cohorts or data distributions. Additionally, the specific way of calculating the metrics varies, as well as the size of the test set. Overall, the lack of a consistent and uniform evaluation framework makes it very hard to effectively compare approaches and draw meaningful conclusions. Despite these challenges, we will make an effort to place our approach in the literature.

### 8.3.1 Datasets

The data used in most studies never represents the full spectrum of medical conditions associated with CMBs, mostly focusing on only one. In terms of dataset size, a third of the published studies have a dataset of less than or equal to 20 patients. Another third has between 20 and 80 patients. For the rest, the largest reported dataset used includes a total of 320 unique patients and 1149 CMBs [11, 109, 110]. The next largest dataset ever used consists of 270 scans with 505 CMBs [111]. In our method, **we have a total of 293 unique patients, with 545 scans and 830 scans** in the train-validation splits. If we also consider the negative cases, that add 742 patients and scans (we exclude sMOMENI, as it is derived from MOMENI), no extra

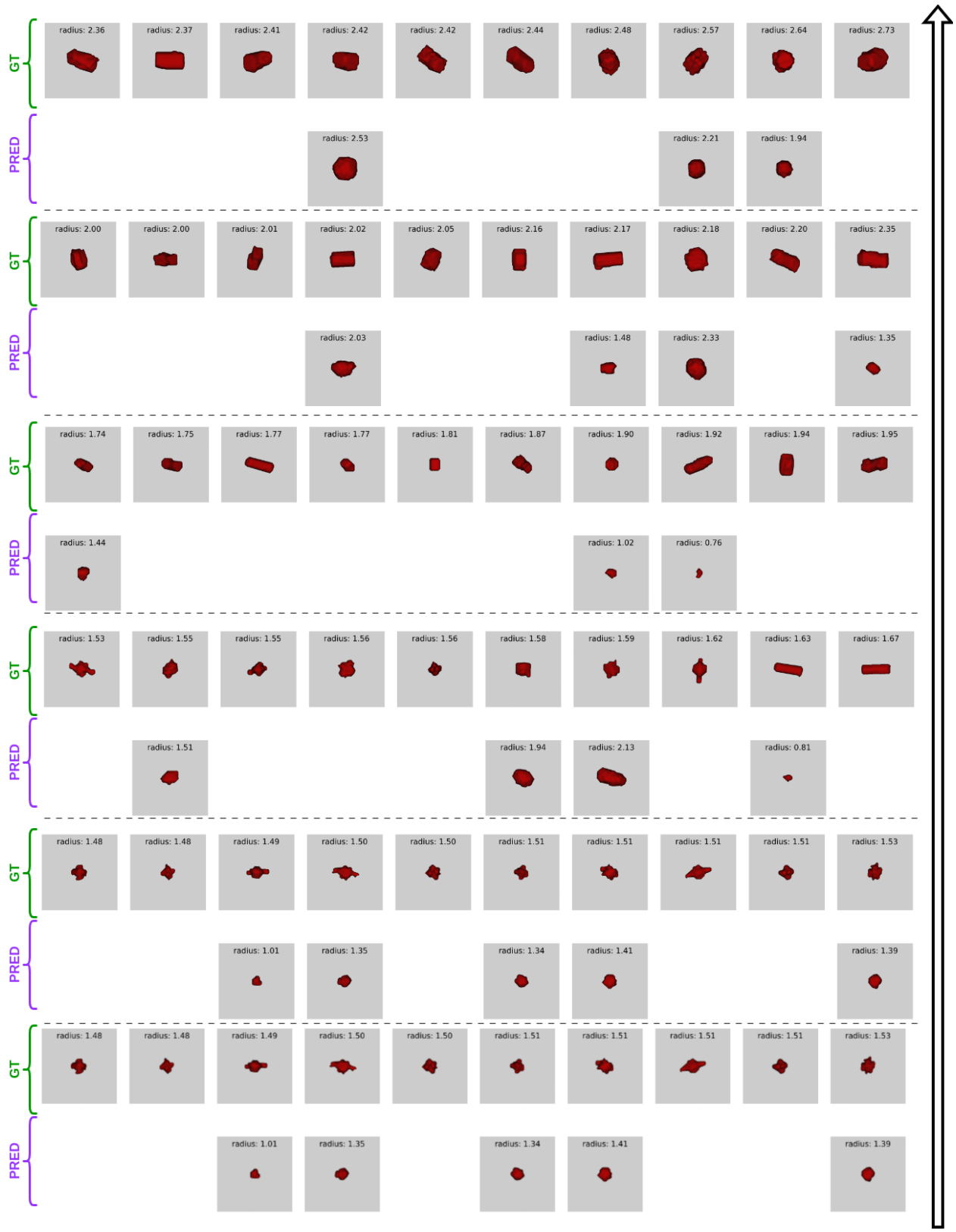



Figure 8.3: **Comparison of ground truth and predicted CMBs renders for CRB dataset**. A visual comparison between the ground truth and predicted CMBs in the CRB dataset, arranged to show increasing size and radius from bottom to top and left to right, respectively. Blank spaces highlight where the model failed to detect a CMB. Images were randomly selected to equally represent true positives and false negatives across sizes

counts are there. In the test set, we have an additional 38 patients with 38 scans and 201 CMBs. As the latter count is normally included in the published methods' total count, we have the same data size as the largest published in terms of patients.

Regarding slice thickness, only four studies reported using a slice thickness greater than 3mm [71,111–113], with the majority using thinner slices. In terms of imaging techniques, most datasets employ the MARS rating scale and SWI sequence type, with older studies opting for T2S sequences.

### 8.3.2 Test sets used

No public dataset contains all the data variations needed to fully assess the robustness of an automated method for CMB detection. We identified all existing public datasets in this project (VALDO, MOMENI, DOU), and none contained sufficient data variation to fully represent the varied MRI spectrum in which radiologists in clinical practice normally detect CMBs. Moreover, most papers test their methods on a separate data split from their data, instead of reserving separate datasets for such purposes as we did in this project. Often evaluation results are done as part of the x-fold cross-validation or leave-one-out validation.

This means that the evaluations reported for published methods are computed on very different sets of data, often private, and when not the available public datasets are insufficiently representative. There is no clear benchmark, and comparing other approaches with our methods can only be done with the few papers that were also evaluated on the DOU test set. We intended to evaluate on the VALDO challenge test set for this project, which we thought was a very good initiative to create an official benchmark, but according to the organizers, this data is no longer available.

### 8.3.3 Metrics reported

Most published papers on automatic CMB detection report an insufficient number of metrics, a single overview metric, without including complementary metrics. Most papers only report the recall/TPR of the detection, often without providing any information on the precision/PPV of the method. Only about ten papers report or allow inference of both precision and recall. Of these, only one paper reports also FPcmb [112], and no papers except [114] publishes both FPavg and FPcmb.

Only a few publications perform image-level classification of scans having at least one CMB versus none and report TNR (image-level). Most of them show very high specificity, often close to or equal to 100%, but this is often coupled with an incredible number of either reported or inferred FPs or an unreported and impossible-to-infer number of FPs or precision at the detection level. Only four studies that report TNR also report detection precision: [115], [116], [12], and [117]. Interestingly, those that do so have very good precision (except for [12], which reports 9% precision) but do not report FPavg or FPcmb. When reported or able to be computed, F1 scores range between 0.19 and 0.8, with particularly high values of 0.91 observed in [118] and 0.9836 in [119].

FPs seem to be an issue across studies, with some published papers on CMB segmentation have disproportionate numbers of FPs, with FPavg-FPcmb numbers of 159-unknown [9], 107-5.4 [10], 27-unknown [86], 26-unknown [120], 45-1.5 [121], 45-unknown [38], 56-5.3 [122], 45-1.5 [123], and 69-unknown [114].

Finally, the interpretation of the preceding metrics becomes challenging, given that many are computed at the patch level, whereas our study evaluates at the image level. In the patch-wise evaluation, they preselected CMB patches and a comparable number of non-CMB patches as inputs. Occasionally patches contained multiple CMBs, which complicates any fair comparison with object detection metrics.

### 8.3.4 Performance comparison with existing methods

We attempted a fair comparison with methods that evaluated the DOU dataset, as outlined in Table 8.4. We assumed that the metrics reported in the literature were equivalent to the *glob*-averaged ones computed in our study. Unfortunately, few papers have utilized this dataset for detection-level evaluation and this dataset. Furthermore, we concentrated on the most recent models that employed advanced DL techniques. However, for the selected papers, detailed information about metric computation, the independence of the test sets, and resolutions of ambiguous matches in cases of double detections were often unavailable. Consequently, this comparison is severely limited. Nevertheless, we can at least observe that our method achieves detection

performance comparable to that reported in the modern approaches. Notably, we achieved the very low positive rates comparatively with decent recall. Worth noting, our F1 score outperforms the majority of other methods, with the exception of the results presented in [72], which are of an exceptionally high standard. However, it should be noted that they use a different localization criterion (Point Inside Mask) and no assignment strategy (double matches allowed); which as we saw in the previous section would improve our scores for DOU.

Table 8.4: **Comparison with other Methods** - This table summarizes the performance metrics of different CMB detection methods, focusing solely on precision, recall, F1 score, and false positives per scan; global-averaged.

| **Method** | *glob***Precision** | *glob***Recall** | *glob***F1** | **FPscan** |
|---|---|---|---|---|
| Sundaresan et al. [72] | 0.89 | 0.87 | 0.88 | 0.5 |
| Ferrer et al. [71] | 1 | 0.5 | 0.67 | 0 |
| Dou et al. [11] | 0.49 | 0.8 | 0.61 | 7.7 |
| OUR APPROACH | 0.84 | 0.69 | 0.76 | 0.135 |

# 9 Discussion

In this study, we developed a deep learning model specifically designed to detect CMBs in close to real-life clinical settings, where challenges arise from significant variations in demographics, MRI acquisition parameters, and clinical conditions. To address these challenges, we curated a diverse collection of private and public datasets. We then trained through meticulous data splitting and augmentation strategies a custom 3D U-Net. We divide the training into two phases: a pre-training phase, where synthetic CMBs alongside a large set of negative scans were utilized to enhance the model's discriminatory capabilities between healthy and pathological scans; and the fine-tuning phase, aimed at refining the model's accuracy and generalization capabilities. Additionally, we explored the potential benefits of transfer learning from an in-house model trained on a substantial corpus of clinically relevant data, ultimately finding that training from scratch was more advantageous.

To ensure a rigorous evaluation, we employed two separate and independent datasets for testing the model, thereby addressing known evaluation deficiencies in the literature. The testing involved a public dataset, used for benchmarking, where we achieved highly satisfactory results and a more clinically challenging dataset that demonstrated a marked drop in performance. Further performance analysis indicated that factors such as slice thickness and the overall count of microbleeds significantly impacted model performance. In the subsequent subsection, we will specific facets of this study.

## 9.1 Pros and cons of Transfer Learning

In our comparison of detection, segmentation, and image-classification results between the *Scratch* and *TL* models, it became evident that the *Scratch* model exhibited superior performance on the test sets, leading us to select it for further performance analysis.

It is crucial to note that the training durations for both models differed significantly. The *TL* model required approximately half the training time of the *Scratch* model—16 days versus 34 days respectively—and fewer epochs (197 epochs in 17 days and 108 epochs in 8 days during pretraining; 187 epochs in 17 days and 99 epochs in 8 days for fine-tuning). In all cases, training was halted upon reaching a plateau or observing overfitting, with the best model selected based on the F1 score for pretraining and validation loss for fine-tuning. Initially, one might assume that the difference in training times could significantly impact results. However, this reduction was anticipated and aligns with one of the primary benefits of transfer learning—reduced computation time. In our case, transfer learning reduced the training duration by half, while maintaining comparable performance and reaching a plateau in terms of validation metrics. For this reason, we believe these models can be fairly compared.

An interesting point of discussion is why the pre-trained model did not show superior performance, particularly since the Apollo model was trained on similar data types and pathologies, specifically macrohemorrhages. One potential reason for the underperformance of the TL model relates to the resolution of the training data. The Apollo model was trained with data in a 1x1x1mm$^3$ isotropic space. Consequently, the last filters of the contracting path, which encode high-level features, might expect different proportions relative to the patch size. For instance, if the pre-trained model learned that macrohemorrhages typically occupy half of the patch in a 1mm isotropic space, it might struggle to adapt when encountering a macrohemorrhage in a 0.5mm isotropic space, where the same lesion occupies a greater proportion of the patch due to the higher resolution (doubling in each dimension). This discrepancy could render the pre-learned spatial relationships less effective. However, the low-level features learned in the first filter layers should remain beneficial. The superior performance of the Scratch model might be attributed to its ability to learn high-level features from scratch, without being constrained by any pre-existing biases in the feature space, allowing it to better specialize and potentially avoid overfitting to the specific characteristics of the new task.

Interestingly, the *TL* model did show slightly better performance on the CRB dataset, possibly due to the inclusion of lower-resolution training data in the Apollo dataset, which featured many instances of large slice

thickness. Nonetheless, this advantage was not sufficient.

## 9.2 The problem of Annotations

During our examination of the annotated data, we identified several significant issues that put in doubt the integrity and reliability of the annotations used:

- **Microbleed Size in MOMENI:** The majority of microbleeds in the MOMENI dataset fall below the clinically accepted minimum radius of 1mm, suggesting that these annotations may not represent clinically significant CMBs and perhaps protocol was not correctly followed
- **Inconsistency in MOMENI:** We observed instances within the MOMENI dataset where specific CMBs were annotated in some scans but were missing in others for the same patient, despite being clearly visible. This inconsistency highlights potential false negatives in the ground truth annotations.
- **Outliers in CRB Dataset:** The CRB dataset exhibits outliers in all measured characteristics of CMBs and scan-level features, indicating possible errors in the annotation or pre-processing.
- **Lack of Double Reads:** Unlike best practices suggested in the literature, all our datasets currently rely on single-reader annotations. Optimal annotation processes often include double readings followed by arbitration or consensus to resolve discrepancies, enhancing the reliability of the annotations.

These findings suggest broader systemic issues in the annotation of CMBs. A review of the literature reinforced our concerns, particularly regarding the methods used to assess annotator agreement:

- **Flawed Inter-annotator Agreement Metrics:** Metrics such as the Intraclass Correlation Coefficient (ICC) are limited to counting microbleeds and do not effectively capture agreement on the individual detection of CMBs. This limitation can obscure true performance, as FPs may be mistakenly counted as TPs, and FNs as true negatives TNs, giving a misleading impression of accuracy. Moreover, metrics like kappa, which simply dichotomize outcomes based on the total number of detected CMBs irrespective of their specific locations, fail to provide a nuanced assessment of annotator consistency.
- **Model Performance vs. Rater Agreement:** Considering the deficiencies in rater agreement metrics, it is pragmatic to assess our model's performance against simplified dichotomization scenarios or basic counts of microbleeds. This approach is particularly justified given the absence of detailed annotator agreement for exact CMB detection in published studies. Without such measures, accurately gauging our model's proximity to human performance remains challenging, as the expected agreement is unrealistic.

These observations underscore the need for rigorous review and potential revision of annotation protocols and inter-annotator agreement to improve the reliability of GT data used for automatic CMB detection models.

## 9.3 Model Performance

The most obvious observation after our performance analysis is that **performance varies significantly between the CRB and DOU datasets**. While the performance on the DOU dataset was robust and satisfactory across all settings, the CRB dataset exhibited less favorable outcomes. This discrepancy underscores the complexities inherent in the CRB test set, which presents unique challenges not typically encountered in other datasets used for CMB detection published in the literature. Specifically, the CRB dataset includes:

- **Poor Slice Thickness:** This leads to tubular-looking CMBs that are not predicted as CMBs because they no longer appear spherical.

- **High Incidence of CMBs:** Many scans contain a large number of CMBs per scan, which often results in more frequent model failures.
- **High Variability:** There is greater variability of scan-, patient-, and demographic-level features than is typical.

In any case, we identified several key points regarding the performance of our model:

- **Tendency to Under-predict:** Our model tends to under-predict when uncertain about a CMB rather than sacrificing precision, creating more FNs than FPs, a trend that is opposite from those observed in the literature.
- **Problem of Redundant Predictions:** The model occasionally makes double predictions on the same CMB, which negatively skews metrics by creating artificial false positives.
- **Very Good Precision:** When compared to other methods, and even without resolving the issue of redundant predictions, the precision on the DOU test set is excellent. The rate of false positives is significantly lower than that of most other published methods.
- **Inability to Improve by Post-processing on the masks:** The problem of having more false negatives than false positives is challenging to address. While maintaining high precision is desirable, we might consider the benefits of tweaking the loss function to reduce false negatives at the cost of increased false positives, combined with a candidate verification stage.
- **Failure with Large Slice Thickness:** Large slice thickness and a high number of CMBs per scan have a profoundly negative effect on detection accuracy, creating tubular-looking CMBs that resemble veins, one of the most common mimics. We might consider assigning less weight to this dimension somehow.
- **Equal Performance across CMB locations**: model showed equal performance regardless of location in the brain
- **Struggle to detect elongated CMBs**: the model demonstrated a clear ability to disregard elongated CMBs, resulting in the generation of false negatives in low-resolution scenarios.

## 9.4 Limitations

In the course of our study, we identified several limitations that affected the overall performance and feasibility of our CMB detection model. These limitations are critical for understanding the boundaries within which our results should be interpreted. Specifically:

- **Location Analysis Precision:** The precision of location analysis might be compromised as SynthSeg was not very precise in demarcating anatomical boundaries for SWI and T2S sequences, likely due to insufficient contrast in these sequences.
- **Training Duration:** The training time was excessively long, with total training times ranging between 2 and three weeks. This duration makes it unfeasible to perform cross-validation, which would be desirable to enhance the robustness of the findings. Accelerating convergence through performance optimization is a critical need, especially for 3D models.
- **Inference Time:** The high inference time, requiring 1 minute and 30 seconds to predict on one patient using a 16GB GPU with a batch size of 14, to which SynthSeg prediction and preprocessing steps must be added, could be impractical for clinical applications and needs improvement.
- **Preprocessing Decisions:** Some decisions in preprocessing may have been suboptimal. For instance, growing 3D spheres should have been performed in isotropic space post-resampling, rather than before, to allow more controlled adjustments even though sphere creations accounted for voxel dimensions in the original space

- **Evaluation Resolution:** Perhaps the evaluation should have been conducted at the original image resolution by undoing the resampling. This approach ensures that no additional effects occur due to downsampling and might justify growing spheres in the original space.
- **Hyperparameter Tuning:** The hyperparameter tuning phase was brief, which might have hindered faster convergence of the model.
- **Lack of Cross-Validation:** The absence of cross-validation due to computational demand in our study limits the generalizability and robustness of the results.
- **Comparative Analysis:** It is virtually impossible to compare our approach with other methods in the literature due to non-standardized evaluation and lack of benchmarks.
- **Clinical Variability:** An inherent challenge in MIA is the inability to test models against all possible clinical variabilities. Our approach tried to approximate the ideal by evaluating not just on an internal test split of a specific cohort but across broader scenarios.
- **Global Context in Model:** The model lacks sufficient global context, which might be crucial for making accurate decisions about individual detections of CMB-like structures.

## 9.5 Future Work

Based on our findings, picture several areas for future research that could potentially enhance the viability of the of CMB automatic detection task. These suggestions aim to address the limitations encountered and to explore new methodologies that could bring significant improvements:

- **Investigating Mimics:** Explore the presence of mimics that may vary between datasets, such as more in CRB and fewer in DOU, which could affect model performance. Thinking of some heuristics to filter them out might really help.
- **Vessel Segmentation:** Implement vessel segmentation to flag dubious cases near vessels as potential false positives.
- **Efficient Architectures:** Research more efficient 2D architectures, like YOLO, which allow for bounding box heuristics based on the expected CMB sizes, which are fixed.
- **Annotation Consistency:** Investigate the effects of using different rating scales on annotation consistency, particularly how scales like MARS and BOMBS influence the detection of varying CMB sizes.
- **Statistical Analysis:** Conduct more detailed statistical analysis to understand performance based on the characteristics of CMBs or scans, with bigger test sets that allow for significant conclusions.
- **Model Training Phases:** Evaluate the impact of omitting the pre-training phase on model effectiveness.
- **Different Fine-tuning strategy**: Considering different weights of carrying out fine-tuning in this task. For instance, we could freeze the filters on the encoder and just train the decoder, which has been shown beneficial in some studies [58].
- **Loss Function:** Consider using a loss function that incorporates a detection-based metric to potentially improve detection specifically.
- **Instance Segmentation:** Explore instance segmentation models to manage the overlap of possible closely situated CMBs.
- **Scan TE Effects:** Further investigate how the echo time (TE) of the scan affects observed CMB size and consider informing the model about this factor to enhance detection accuracy.

- **Post-processing on Logits:** do post-processing strategies on the model logits to avoid under-segmentation of cases with tubular shapes
- **Adapt to Clinical Need:** More closely mimics radiological assessments by focusing on getting counts per location area; or classifying them into binary groups; reflecting clinical priorities more accurately.
- **Prospective Studies:** Plan and execute prospective studies to measure the practical benefits of model assistance, such as time saved for raters and the reliability of model predictions versus human annotations in blinded experiments.
- **Health Status Distinction:** Enhance model evaluation by including a test set to evaluate classification into healthy versus unhealthy scans. This would test the model's ability to distinguish between these two states
- **Global Context Integration:** Integrate global context using techniques akin to positional encoding in natural language processing. For instance, simply using coordinates of the patch center within the whole image as an extra input.
- **Whole-image Attention:** Develop whole-image attention mechanisms that encode the entire image at different levels of resolution, providing the model with comprehensive contextual information during patch predictions.

## 9.6 Conclusion

In conclusion, our study highlights several critical findings and underscores the complexities involved in CMB detection. Unlike prevailing trends in the literature, where precision is the primary concern due to the large number of false positives, our model encounters greater challenges in achieving a high recall due to the abundance of false negatives. Interestingly, transfer learning did not enhance performance as anticipated, though it did reduce training time. Our model excels with datasets characterized by high resolution and SWI, yet struggles with more complex datasets like CRB—which we specifically developed to test challenging scenarios—primarily due to insufficient slice thickness and a high occurrence of CMBs. We thus recommend employing our model in scenarios where MRI anisotropy is minimal, as accurate 3D feature recognition is essential for effective classification.

Our results are difficult to compare directly with published methods due to the absence of a benchmark dataset, reliance on private data, non-standardized metric reporting and computation, and use of tests sets of limited clinical variability within the field. This highlights the urgent need for standardized evaluations and the development of benchmark datasets. Additionally, the questionable quality of annotations and the lack of detailed inter-annotator agreement using detection-level metrics rather than global counts reveal fundamental issues in how current CMB detection rating protocols are assessed and highlight the need for reevaluating this expected agreement and possibly enhance rating scales.

Overall, we have successfully developed a sequence-agnostic model that performs well in high-resolution environments but encounters challenges at lower resolutions. Our analysis revealed significant data issues that complicate CMB detection and require further investigation. Additionally, we identified crucial methodological shortcomings in the expert rating and assessment of CMBs, as well as in the evaluation methodologies used for testing the automated approaches. These findings underscore the need for enhanced rating protocols and standardization across evaluations to advance the field effectively.

# Bibliography


[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] 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.

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

[4] S. Yan, X. Jin, X. Zhang, S. Zhang, D. S. Liebeskind, and M. Lou, "Extensive cerebral microbleeds predict parenchymal haemorrhage and poor outcome after intravenous thrombolysis." [Online]. Available: http://jnnp.bmj.com/

[5] S. Akoudad, F. J. Wolters, A. Viswanathan, R. F. D. Bruijn, A. V. D. Lugt, A. Hofman, P. J. Koudstaal, M. A. Ikram, and M. W. Vernooij, "Association of cerebral microbleeds with cognitive decline and dementia," *JAMA Neurology*, vol. 73, pp. 934–943, 8 2016.

[6] 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

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

[8] 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.

[9] A. Fazlollahi, F. Meriaudeau, V. Villemagne, C. Rowe, P. Desmond, P. Yates, O. Salvado, P. Bourgeat, V. L. Villemagne, C. C. Rowe, P. M. Desmond, and P. A. Yates, "Automatic detection of small spherical lesions using multiscale approach in 3d medical images," 2013. [Online]. Available: https://hal.science/hal-00860520

[10] S. R. Barnes, E. M. Haacke, M. Ayaz, A. S. Boikov, W. Kirsch, and D. Kido, "Semiautomated detection of cerebral microbleeds in magnetic resonance images," *Magnetic Resonance Imaging*, vol. 29, pp. 844–852, 7 2011.

[11] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P. A. Heng, "Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1182–1195, 5 2016.

[12] S. Momeni, A. Fazlollahi, P. Yates, C. Rowe, Y. Gao, A. W. C. Liew, and O. Salvado, "Synthetic microbleeds generation for classifier training without ground truth," *Computer Methods and Programs in Biomedicine*, vol. 207, 8 2021.

[13] C. H. Sudre *et al.*, "Where is valdo? vascular lesions detection and segmentation challenge at miccai 2021," *MICCAI Conference Proceedings*, 2022. [Online]. Available: http://arxiv.org/abs/2208.07167

[14] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, 2015.

[15] J. del Pozo Lerida, "Automatic segmentation of cerebral microbleeds," https://github.com/jorgedelpozolerida/Segmentation_CMB/blob/main/JorgedelPozoLerida_ResearchProject_AutomatedSegmentationofCMB.pdf, 2023.

[16] Imaios, "Introduction to MRI sequences," https://www.imaios.com/en/e-Courses/e-MRI/MRI-Sequences/MRI-sequences, 2009, accessed: 2022-04-29.

[17] S. Haller, E. M. Haacke, M. M. Thurnher, and F. Barkhof, "Susceptibility-weighted imaging: Technical essentials and clinical neurologic applications," pp. 3–26, 4 2021.

[18] K. Nandigam, "Swan mri revealing multiple microhemorrhages secondary to septic emboli from mucormycosis," pp. 199–200, 7 2013.

[19] C. Cordonnier, R. A.-S. Salman, and J. Wardlaw, "Spontaneous brain microbleeds: Systematic review, subgroup analyses and standards for study design and reporting," pp. 1988–2003, 2007.

[20] L. Puy, M. Pasi, M. Rodrigues, S. J. V. Veluw, G. Tsivgoulis, A. Shoamanesh, and C. Cordonnier, "Cerebral microbleeds: From depiction to interpretation," pp. 598–607, 6 2021.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] H. Bokura, R. Saika, T. Yamaguchi, A. Nagai, H. Oguro, S. Kobayashi, and S. Yamaguchi, "Microbleeds are associated with subsequent hemorrhagic and ischemic stroke in healthy elderly individuals," *Stroke*, vol. 42, pp. 1867–1871, 7 2011.

[28] 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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] S. M. Gregoire, U. J. Chaudhary, M. M. Brown, T. A. Yousry, F. C. Kallis, H. R. Jäger, and F. D. J. Werring, "The microbleed anatomical rating scale (mars) reliability of a tool to map brain microbleeds," 2009. [Online]. Available: www.neurology.org

[35] 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.

[36] J. D. Goos, W. M. V. D. Flier, D. L. Knol, P. J. Pouwels, P. Scheltens, F. Barkhof, and M. P. Wattjes, "Clinical relevance of improved microbleed detection by susceptibility- weighted magnetic resonance imaging," *Stroke*, vol. 42, pp. 1894–1900, 7 2011.

[37] J. C. Purrucker, M. Wolf, K. Haas, T. Siedler, T. Rizos, S. Khan, P. U. Heuschmann, and R. Veltkamp, "Microbleeds in ischemic vs hemorrhagic strokes on novel oral anticoagulants," *Acta Neurologica Scandinavica*, vol. 138, pp. 163–169, 8 2018.

[38] H. J. Kuijf, S. J. van Veluw, M. A. Viergever, K. L. Vincken, and G. J. Biessels, "How to assess the reliability of cerebral microbleed rating?" 2013.

[39] 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.

[40] 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.

[41] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," 12 2022.

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

[43] 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.

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

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

[46] 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.

[47] 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.

[48] O. Bernard *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2514–2525, 11 2018.

[49] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," 11 2018. [Online]. Available: http://arxiv.org/abs/1811.02629

[50] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 11 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[51] 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

[52] G. Valiant, "A theory of the learnable."

[53] J. Quinonero-Candela, *Dataset shift in machine learning*. MIT Press, 2009.

[54] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, pp. 521–530, 2012.

[55] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. G. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken, A. Fedorov, P. Abolmaesumi, B. Platel, and W. M. Wells, "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation," 2 2017. [Online]. Available: http://arxiv.org/abs/1702.07841http://dx.doi.org/10.1007/978-3-319-66179-7_59

[56] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, and E. Burnaev, "Domain shift in computer vision models for mri data analysis: An overview," 10 2020. [Online]. Available: http://arxiv.org/abs/2010.07222

[57] S. J. Pan and Q. Yang, "A survey on transfer learning," pp. 1345–1359, 2010.

[58] D. Karimi, S. K. Warfield, and A. Gholipour, "Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations," *Artificial Intelligence in Medicine*, vol. 116, 6 2021.

[59] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 5 2019.

[60] V. Cheplygina, "Cats or cat scans: Transfer learning from natural or medical image source data sets?" pp. 21–27, 3 2019.

[61] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[62] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," pp. 545–563, 8 2021.

[63] 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.

[64] 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.

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

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

[67] 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.

[68] 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.

[69] 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.

[70] 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

[71] 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

[72] 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.

[73] 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.

[74] 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.

[75] J.-H. Kim, Y. Noh, H. Lee, S. Lee, W.-R. Kim, K. M. Kang, E. Y. Kim, M. A. Al-Masni, and D.-H. Kim, "Toward automated detection of microbleeds with anatomical scale localization: A complete clinical diagnosis support using deep learning."

[76] P. Xia, E. S. Hui, B. J. Chua, F. Huang, Z. Wang, H. Zhang, H. Yu, K. K. Lau, H. K. Mak, and P. Cao, "Deep-learning-based mri microbleeds detection for cerebral small vessel disease on quantitative susceptibility mapping," *Journal of Magnetic Resonance Imaging*, 2023.

[77] T. Xia, R. Zhang, Z. Chen, G. Xie, X. Wu, Z. Lv, and L. Guo, "Progressive learning based knowledge distillation for low resolution cerebral microbleed segmentation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1856–1860.

[78] K. A. Ellis *et al.*, "The australian imaging, biomarkers and lifestyle (aibl) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease," *International Psychogeriatrics*, vol. 21, no. 4, pp. 672–687, 2009.

[79] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 1310–1320, 6 2010.

[80] T. Tillin, A. D. Hughes, J. Mayet, P. Whincup, N. Sattar, N. G. Forouhi, P. M. McKeigue, and N. Chaturvedi, "The relationship between metabolic risk factors and incident cardiovascular disease in europeans, south asians, and african caribbeans: Sabre (southall and brent revisited) - a prospective population-based study," *Journal of the American College of Cardiology*, vol. 61, pp. 1777–1786, 4 2013.

[81] M. A. Ikram, A. van der Lugt, W. J. Niessen, P. J. Koudstaal, G. P. Krestin, A. Hofman, D. Bos, and M. W. Vernooij, "The rotterdam scan study: design update 2016 and main findings," *European Journal of Epidemiology*, vol. 30, pp. 1299–1315, 12 2015.

[82] J. L. Molinuevo, N. Gramunt, J. D. Gispert, K. Fauria, M. Esteller, C. Minguillon, G. Sánchez-Benavides, G. Huesa, S. Morán, R. Dal-Ré, and J. Camí, "The alfa project: A research platform to identify early pathophysiological features of alzheimer's disease," *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, vol. 2, pp. 82–92, 6 2016.

[83] M. W. Vernooij, A. V. D. Lugt, M. A. Ikram, P. A. Wielopolski, W. J. Niessen, A. Hofman, G. P. Krestin, and M. M. B. Breteler, "Prevalence and risk factors of cerebral microbleeds the rotterdam scan study," 2008.

[84] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. Breteler, H. Chabriat, C. DeCarli, F. E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. van Oostenbrugge, L. Pantoni, O. Speck, B. C. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P. B. Gorelick, and M. Dichgans, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," pp. 822–838, 8 2013.

[85] 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.

[86] A. Fazlollahi, F. Meriaudeau, L. Giancardo, V. L. Villemagne, C. C. Rowe, P. Yates, O. Salvado, and P. Bourgeat, "Computer-aided detection of cerebral microbleeds in susceptibility-weighted imaging," *Computerized Medical Imaging and Graphics*, vol. 46, pp. 269–276, 12 2015.

[87] B. Billot, C. Magdamo, Y. Cheng, S. E. A. ID, S. I. Das, and J. E. Iglesias, "Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets," 2023. [Online]. Available: https://doi.org/10.1073/pnas.2216399120

[88] P. N. Jensen, T. Rashid, J. B. Ware, Y. Cui, C. M. Sitlani, T. R. Austin, W. T. Longstreth, A. G. Bertoni, E. Mamourian, R. N. Bryan, I. M. Nasrallah, M. Habes, and S. R. Heckbert, "Association of brain microbleeds with risk factors, cognition, and mri markers in mesa," *Alzheimer's and Dementia*, vol. 19, pp. 4139–4149, 9 2023.

[89] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "Synthstrip: skull-stripping for any brain image," *NeuroImage*, vol. 260, 10 2022.

[90] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 6 2016. [Online]. Available: http://arxiv.org/abs/1606.06650

[91] I. Valverde, G. Gomez-Ciriza, T. Hussain, C. Suarez-Mejias, M. N. Velasco-Forte, N. Byrne, A. Ordoñez, A. Gonzalez-Calle, D. Anderson, M. G. Hazekamp, A. A. Roest, J. Rivas-Gonzalez, S. Uribe, I. El-Rassi, J. Simpson, O. Miller, E. Ruiz, I. Zabala, A. Mendez, B. Manso, P. Gallego, F. Prada, M. Cantinotti, L. Ait-Ali, C. Merino, A. Parry, N. Poirier, G. Greil, R. Razavi, T. Gomez-Cia, and A. R. Hosseinpour, "Three-dimensional printed models for surgical planning of complex congenital heart defects: An international multicentre study," *European Journal of Cardio-thoracic Surgery*, vol. 52, pp. 1139–1148, 12 2017.

[92] C. Wachinger, M. Reuter, and T. Klein, "Deepnat: Deep convolutional neural network for segmenting neuroanatomy," *NeuroImage*, vol. 170, pp. 434–445, 4 2018.

[93] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," 7 2017. [Online]. Available: http://arxiv.org/abs/1707.03237http://dx.doi.org/10.1007/978-3-319-67558-9_28

[94] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," 6 2017. [Online]. Available: http://arxiv.org/abs/1706.05721

[95] 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

[96] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 12 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[97] N. H. Shah, A. Milstein, and S. C. Bagley, "Making machine learning models clinically useful," pp. 1351–1352, 10 2019.

[98] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," 10 2019.

[99] L. Maier-Hein *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nature Methods*, vol. 21, no. 2, pp. 195–212, 2024.

[100] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing,, 2008.

[101] O. U. Aydin, A. A. Taha, A. Hilbert, A. A. Khalil, I. Galinovic, J. B. Fiebach, D. Frey, and V. I. Madai, "On the usage of average hausdorff distance for segmentation performance assessment: hidden error when used for ranking," *European Radiology Experimental*, vol. 5, 12 2021.

[102] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2 2021.

[103] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability (Switzerland)*, vol. 13, pp. 1–29, 2 2021.

[104] A. Popovic, M. de la Fuente, M. Engelhardt, and K. Radermacher, "Statistical validation metric for accuracy assessment in medical image segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 2, pp. 169–181, 2007.

[105] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, 8 2015.

[106] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, pp. 427–437, 7 2009.

[107] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged f 1 and macro-averaged f 1 scores," *Applied Intelligence*, vol. 52, pp. 4961–4972, 3 2022.

[108] D. J. Rumala, "How you split matters: Data leakage and subject characteristics studies in longitudinal brain mri analysis," 9 2023. [Online]. Available: http://arxiv.org/abs/2309.00350

[109] B. F. Stanley and S. W. Franklin, "Effective feature extraction for cerebral microbleed detection using edge emphasized weber maximum directional co-occurance matrix," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 13 683–13 696, 10 2023.

[110] ——, "Automated cerebral microbleed detection using selective 3d gradient co-occurance matrix and convolutional neural network," *Biomedical Signal Processing and Control*, vol. 75, 5 2022.

[111] V. Sundaresan, C. Arthofer, G. Zamboni, R. A. Dineen, P. M. Rothwell, S. N. Sotiropoulos, D. P. Auer, D. J. Tozer, H. S. Markus, K. L. Miller, I. Dragonu, N. Sprigg, F. Alfaro-Almagro, M. Jenkinson, and L. Griffanti, "Automated detection of candidate subjects with cerebral microbleeds using machine learning," *Frontiers in Neuroinformatics*, vol. 15, 1 2022.

[112] T. Li, Y. Zou, P. Bai, S. Li, H. Wang, X. Chen, Z. Meng, Z. Kang, and G. Zhou, "Detecting cerebral microbleeds via deep learning with features enhancement by reusing ground truth," *Computer Methods and Programs in Biomedicine*, vol. 204, 6 2021.

[113] M. L. Seghier, M. A. Kolanko, A. P. Leff, H. R. Jäger, S. M. Gregoire, and D. J. Werring, "Microbleed detection using automated segmentation (midas): A new method applicable to standard clinical mr images," *PLoS ONE*, vol. 6, 2011.

[114] H. Liu, T. Rashid, and M. Habes, "Cerebral microbleed detection via fourier descriptor with dual domain distribution modeling." Institute of Electrical and Electronics Engineers Inc., 4 2020.

[115] S. Wang, C. Tang, J. Sun, and Y. Zhang, "Cerebral micro-bleeding detection based on densely connected neural network," *Frontiers in Neuroscience*, vol. 13, 2019.

[116] P. Doke, D. Shrivastava, C. Pan, Q. Zhou, and Y. D. Zhang, "Using cnn with bayesian optimization to identify cerebral micro-bleeds," vol. 31. Springer, 7 2020.

[117] S. Lu, S. Liu, S. H. Wang, and Y. D. Zhang, "Cerebral microbleed detection via convolutional neural network and extreme learning machine," 9 2021.

[118] M. A. Ferlin, M. Grochowski, A. Kwasigroch, A. Mikołajczyk, E. Szurowska, M. Grzywińska, and A. Sabisz, "A comprehensive analysis of deep neural-based cerebral microbleeds detection system," *Electronics (Switzerland)*, vol. 10, 9 2021.

[119] Z. Lu, Y. Yan, and S. H. Wang, "Cmb-net: a deep convolutional neural network for diagnosis of cerebral microbleeds," *Multimedia Tools and Applications*, vol. 81, pp. 19 195–19 214, 6 2022.

[120] T. L. van den Heuvel, A. W. van der Eerden, R. Manniesing, M. Ghafoorian, T. Tan, T. M. Andriessen, T. V. Vyvere, L. van den Hauwe, B. M. ter Haar Romeny, B. M. Goraj, and B. Platel, "Automated detection of cerebral microbleeds in patients with traumatic brain injury," *NeuroImage: Clinical*, vol. 12, pp. 241–251, 2016.

[121] W. Bian, C. P. Hess, S. M. Chang, S. J. Nelson, and J. M. Lupo, "Computer-aided detection of radiation-induced cerebral microbleeds on susceptibility-weighted mr images," *NeuroImage: Clinical*, vol. 2, pp. 282–290, 2013.

[122] T. Ateeq, M. N. Majeed, S. M. Anwar, M. Maqsood, Z. ur Rehman, J. W. Lee, K. Muhammad, S. Wang, S. W. Baik, and I. Mehmood, "Ensemble-classifiers-assisted detection of cerebral microbleeds in brain mri," *Computers and Electrical Engineering*, vol. 69, pp. 768–781, 7 2018.

[123] M. A. Morrison, S. Payabvash, Y. Chen, S. Avadiappan, M. Shah, X. Zou, C. P. Hess, and J. M. Lupo, "A user-guided tool for semi-automated cerebral microbleed detection and volume segmentation: Evaluating vascular injury and data labelling for machine learning," *NeuroImage: Clinical*, vol. 20, pp. 498–505, 1 2018.

# A Training Details

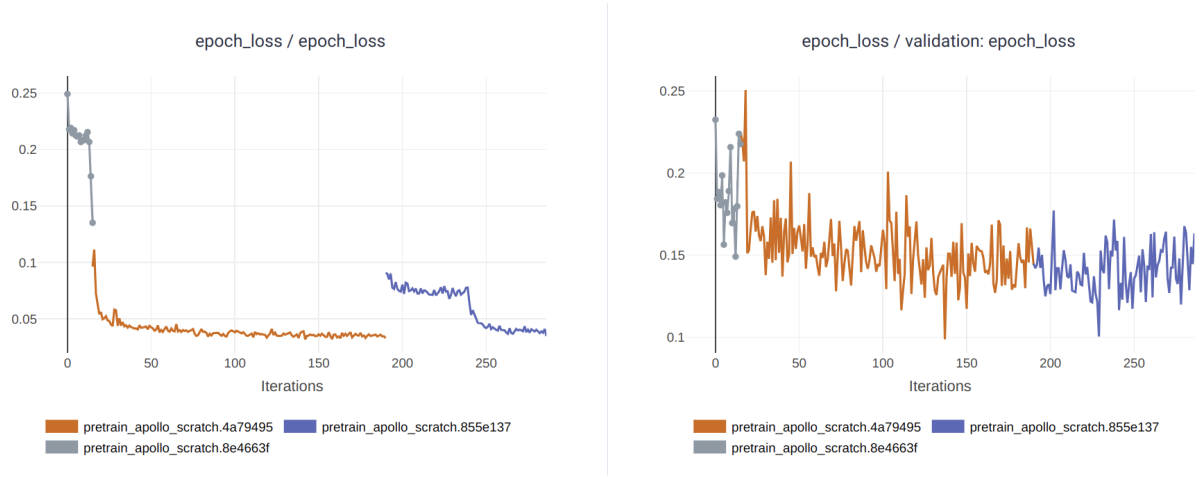



Figure A.1: Training and validation loss during the pre-training phase for the Scratch model. Training was conducted in consecutive parts, each shown with one different color


Figure A.2: Training and validation loss during the fine-tuning phase for the Scratch model.

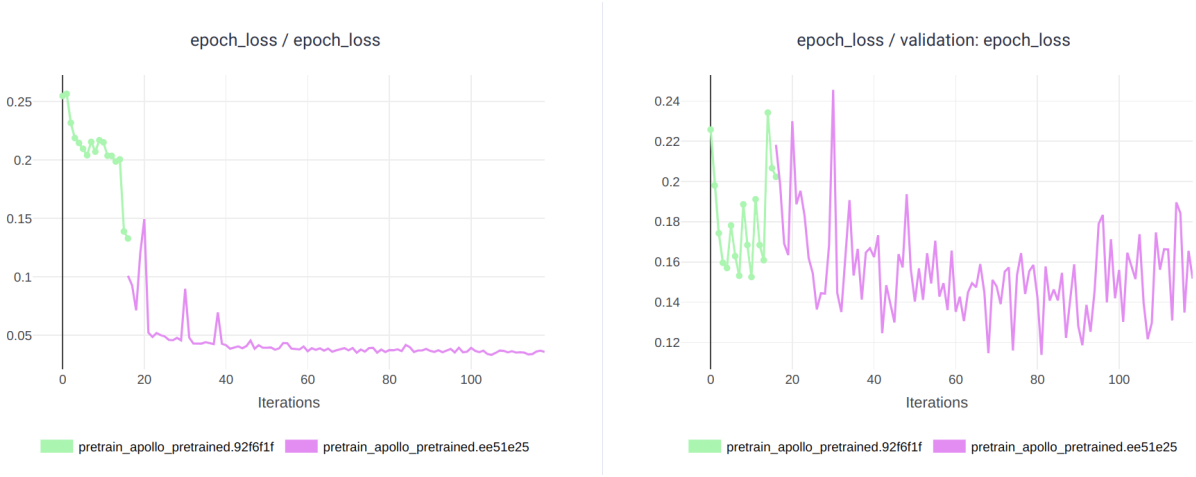



Figure A.3: Training and validation loss during the pre-training phase for the TL model. Training was conducted in consecutive parts, each shown with one different color

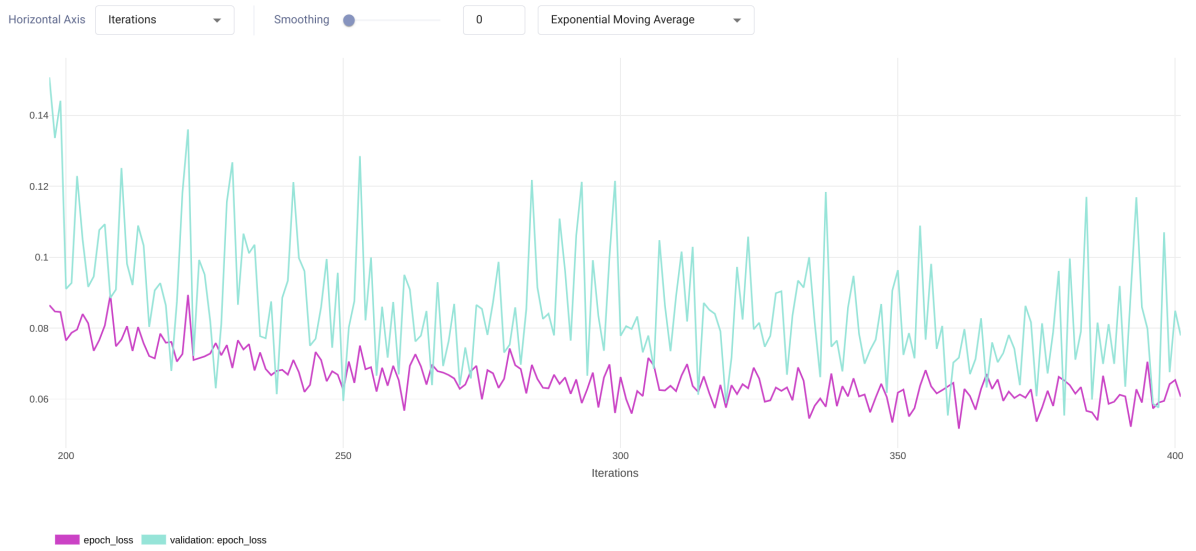



Figure A.4: Training and validation loss during the pre-training phase for the TL model.

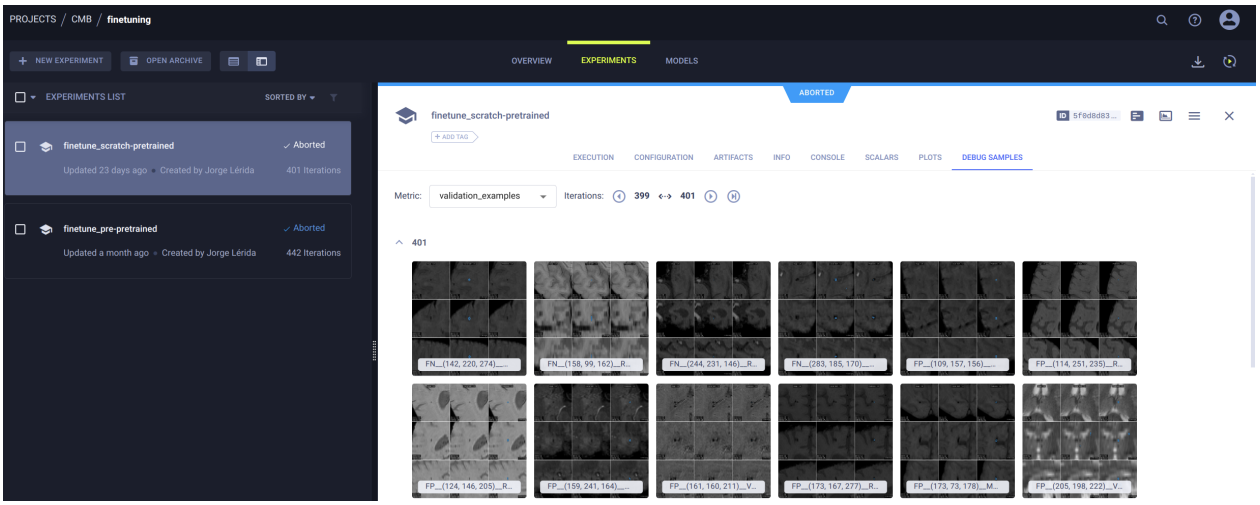



Figure A.5: Example of visual tracking of model's prediction as seen on ClearML.

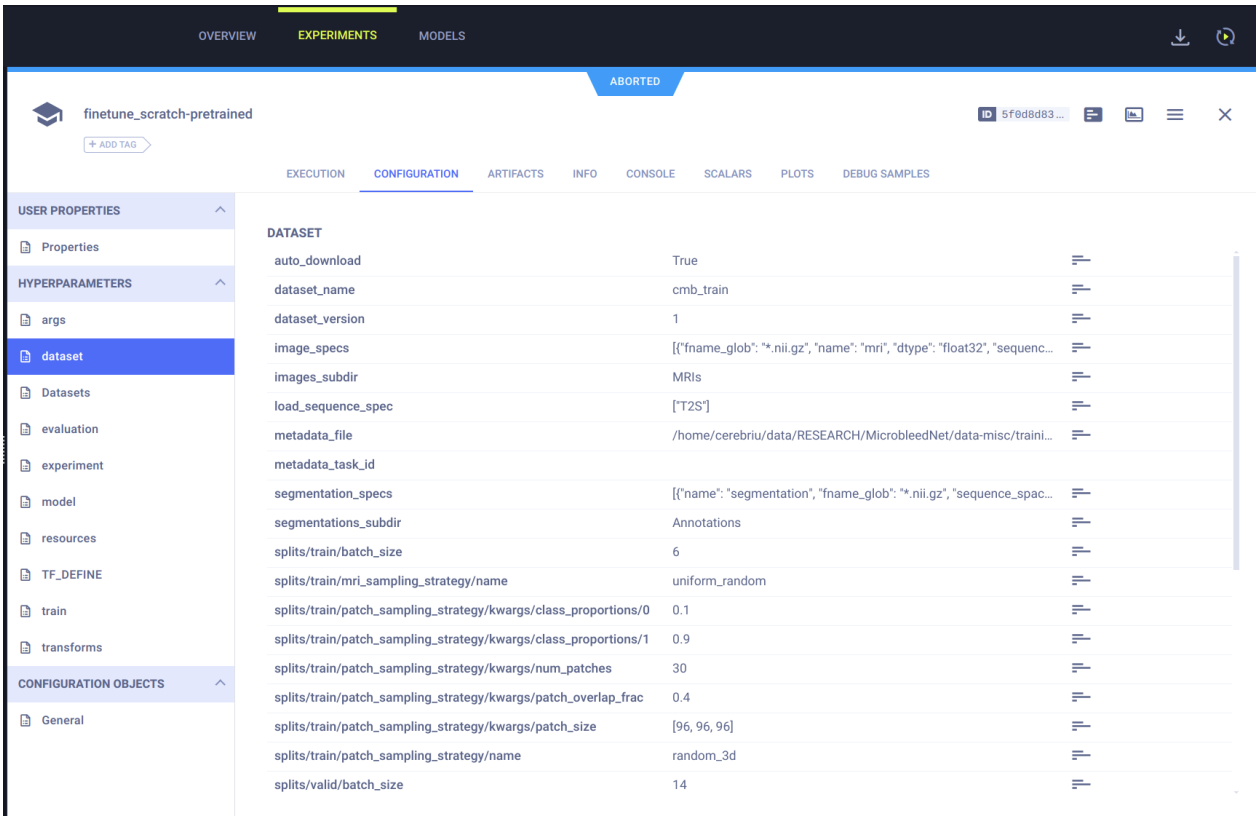



Figure A.6: Example of dataset configuration for fine-tuning task as seen on ClearML

MODEL CONFIGURATION

```
name = "UNetApollo3D"
input_shape = [96, 96, 96]
n_channels = 1
filters_per_depth = [32, 64, 128, 256, 320]
filters_out_bottom_conv = 320
kernel_size = [3, 3, 3]
n_classes = 2
conv_kwargs {
  padding = "same"
}
n_conv_per_depth = 3
n_conv_per_depth_upsampling = 2
n_conv_bottom = 2
residual_conv_index = 1
hidden_activation = "LeakyReLU"
hidden_activation_kwargs {
  alpha = 0.01
}
output_activation = "softmax"
output_activation_kwargs {}
sampling_factor = 2
down_sampling_type = "strided"
up_sampling_type = "transposed"
up_sampling_conv_kwargs {}
up_sampling_filters_factor = 1
merge_layer = "Concatenate"
merge_layer_kwargs = null
merge_order = "skip_first"
normalization = "InstanceNormalization"
normalization_kwargs {
  axis = null
  epsilon = 1e-05
}
kernel_regularizer = "L2"
kernel_regularizer_kwargs {
  l2 = 1e-05
}
deep_supervision = false
```

Figure A.7: Model's architecture specific parameters

Training split stratification:

| Column | Value | Count | Proportion (%) |
|---|---|---|---|
| Dataset | MOMENI | 314 | 67.82% |
| Dataset | RODEJA | 87 | 18.79% |
| Dataset | VALDO | 62 | 13.39% |
| healthy_all | True | 294 | 63.5% |
| healthy_all | False | 169 | 36.5% |
| seq_type | SWI | 401 | 86.61% |
| seq_type | T2S | 62 | 13.39% |
| res_level | low | 350 | 75.59% |
| res_level | high | 113 | 24.41% |
| field_strength | 3 | 314 | 67.82% |
| field_strength | 1.5/3 | 149 | 32.18% |
| CMB_level | Unspecified | 294 | 63.5% |
| CMB_level | low | 135 | 29.16% |
| CMB_level | high | 34 | 7.34% |
| TE | 20.0 | 314 | 67.82% |
| TE | Unspecified | 87 | 18.79% |
| TE | 25.0 | 62 | 13.39% |
| stratify_label | MOMENI-True-SWI-low-3-Unspecified-20.0 | 240 | 51.84% |
| stratify_label | MOMENI-False-SWI-low-3-low-20.0 | 65 | 14.04% |
| stratify_label | RODEJA-False-SWI-high-1.5/3-low-Unspecified | 33 | 7.13% |
| stratify_label | RODEJA-True-SWI-high-1.5/3-Unspecified-Unspecified | 27 | 5.83% |
| stratify_label | VALDO-False-T2S-low-1.5/3-low-25.0 | 23 | 4.97% |
| stratify_label | VALDO-True-T2S-high-1.5/3-Unspecified-25.0 | 19 | 4.1% |
| stratify_label | RODEJA-False-SWI-high-1.5/3-high-Unspecified | 14 | 3.02% |
| stratify_label | VALDO-False-T2S-high-1.5/3-low-25.0 | 11 | 2.38% |
| stratify_label | VALDO-False-T2S-high-1.5/3-high-25.0 | 9 | 1.94% |
| stratify_label | MOMENI-False-SWI-low-3-high-20.0 | 9 | 1.94% |
| stratify_label | RODEJA-True-SWI-low-1.5/3-Unspecified-Unspecified | 8 | 1.73% |
| stratify_label | RODEJA-False-SWI-low-1.5/3-low-Unspecified | 3 | 0.65% |
| stratify_label | RODEJA-False-SWI-low-1.5/3-high-Unspecified | 2 | 0.43% |

Validation split stratification:

| Column | Value | Count | Proportion (%) |
|---|---|---|---|
| Dataset | MOMENI | 56 | 68.29% |
| Dataset | RODEJA | 16 | 19.51% |
| Dataset | VALDO | 10 | 12.2% |
| healthy_all | True | 50 | 60.98% |
| healthy_all | False | 32 | 39.02% |
| seq_type | SWI | 72 | 87.8% |
| seq_type | T2S | 10 | 12.2% |
| res_level | low | 62 | 75.61% |
| res_level | high | 20 | 24.39% |
| field_strength | 3 | 56 | 68.29% |
| field_strength | 1.5/3 | 26 | 31.71% |
| CMB_level | Unspecified | 50 | 60.98% |
| CMB_level | low | 26 | 31.71% |
| CMB_level | high | 6 | 7.32% |
| TE | 20.0 | 56 | 68.29% |
| TE | Unspecified | 16 | 19.51% |
| TE | 25.0 | 10 | 12.2% |
| stratify_label | MOMENI-True-SWI-low-3-Unspecified-20.0 | 40 | 48.78% |
| stratify_label | MOMENI-False-SWI-low-3-low-20.0 | 14 | 17.07% |
| stratify_label | RODEJA-False-SWI-high-1.5/3-low-Unspecified | 6 | 7.32% |
| stratify_label | RODEJA-True-SWI-high-1.5/3-Unspecified-Unspecified | 5 | 6.1% |
| stratify_label | VALDO-False-T2S-low-1.5/3-low-25.0 | 4 | 4.88% |
| stratify_label | RODEJA-False-SWI-high-1.5/3-high-Unspecified | 3 | 3.66% |
| stratify_label | VALDO-True-T2S-high-1.5/3-Unspecified-25.0 | 3 | 3.66% |
| stratify_label | RODEJA-True-SWI-low-1.5/3-Unspecified-Unspecified | 2 | 2.44% |
| stratify_label | VALDO-False-T2S-high-1.5/3-low-25.0 | 2 | 2.44% |
| stratify_label | MOMENI-False-SWI-low-3-high-20.0 | 2 | 2.44% |
| stratify_label | VALDO-False-T2S-high-1.5/3-high-25.0 | 1 | 1.22% |

Initial split without sMOMENI and CRBneg. Train: 463, Valid: 82

Figure A.8: Stratified splits LOG for fine-tuning

# B More on Datasets

Table B.1: Counts for each combination of main high-level groups used during stratification

| Dataset | Seq_type | Res_level | Healthy | Healthy_all | Field_strength | CMB_level | Count |
|---|---|---|---|---|---|---|---|
| CRB | SWI | low | no | False | 1.5/3 | high | 5 |
| | | | | | | low | 2 |
| | T2S | high | no | False | 1.5/3 | high | 4 |
| | | | | | | low | 6 |
| | | low | no | False | 1.5/3 | high | 1 |
| DOU | SWI | high | no | False | 3 | high | 6 |
| | | | | | | low | 14 |
| MOMENI | SWI | low | no | False | 3 | high | 11 |
| | | | | | | low | 46 |
| | | | yes | False | 3 | low | 33 |
| RODEJA | SWI | high | no | False | 1.5/3 | high | 17 |
| | | | | | | low | 39 |
| | | low | no | False | 1.5/3 | high | 2 |
| | | | | | | low | 3 |
| VALDO | T2S | high | no | False | 1.5/3 | high | 10 |
| | | | | | | low | 13 |
| | | low | no | False | 1.5/3 | low | 27 |
| sMOMENI | SWI | low | no | False | 3 | high | 3700 |