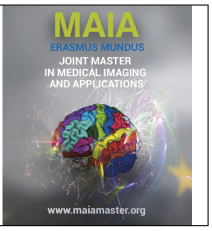
Medical Imaging and Applications

Master Thesis, June 2023



Self-supervised learning for acute ischemic stroke final infarct lesion segmentation in non-contrast CT

Joaquin O. Seia*, Ezequiel de la Rosa[a], Diana M. Sima[a], David Robben[a]

[a] *icometrix, Leuven, Belgium*

Abstract

Ischemic stroke accounts for 87% of all strokes, which are the leading cause of disability and the fifth leading cause of death worldwide. Current stroke management guidelines rely on quantification of ischemic lesion volume to select an appropriate treatment for a patient. Despite the fact that baseline non-contrast CT is not as suitable as it is Perfusion CT or Diffusion Weighted Imaging to obtain this measurement, it is the first imaging modality performed when the patient arrives at the emergency department, it is cheaper, more widely available and faster. Consequently, the development of accurate automated segmentation tools for ischemic lesions on baseline non-contrast CT is a clinically relevant problem that, if satisfactory solved, would represent an improvement in healthcare provision.

Among other difficulties, the low contrast that the ischemic stroke lesion presents on baseline non-contrast CT makes the task of segmenting it very challenging even for expert radiologists. In the case of automated solutions, the difficulty of collecting large, well-curated datasets of baseline and follow-up acute ischemic stroke images further complicates the task. In this work, a self-supervised learning (SSL) pre-training strategy was proposed to exploit large unlabelled non-contrast CT datasets (stroke positive and negative) in the task of acute ischemic stroke infarct segmentation. A robust data pre-processing pipeline was proposed to homogenise the different datasets before using them in a SSL-enhanced version of the well-known self-configuring nnU-Net pipeline. From the experiments conducted, pre-training the nnU-Net encoder in a self-supervised manner with all available non-contrast CT images resulted in an acute ischemic stroke segmentation performance significantly higher than training the same model from scratch and comparable to that obtained by training from scratch using approximately 3.6 times more labelled data.

The code developed for this work is publicly available at: https://github.com/joaco18/stroke-seg-ssl.

Keywords: Acute Ischemic Stroke, Non-Contrast CT, Segmentation, Self-Supervised Learning

1. Introduction

Stroke is a pathology characterised by a focal injury in the central nervous system with a vascular origin. It represents the first cause of disability and the fifth cause of death worldwide (Virani et al. (2021)). A stroke can be classified in two types: ischemic and hemorrhagic. The ischemic type accounts for 87% of all strokes and involves a restriction or reduction of blood flow caused by the occlusion of a blood vessel (Benjamin et al. (2017); Sacco et al. (2013)).

During the management of this emergency, *time is brain*: the longer the brain tissue is deprived of blood supply, the higher the probability of cell death and the worse the prognosis for the patient. Depending on the time elapsed since the onset of the stroke, it can be divided into three categories: *hyperacute* (less than 6 hours from onset), *acute* (less than 24 hours) or *sub-acute* (from 24 hours to 5 days) (Brorson and Cifu (2019)). Each of these stages is accompanied by distinct physiopathological characteristics that define different ways in which the patient should be managed.

The time-dependent fate of the hypoperfused tissue can be spatially described by two clinically relevant zones: *core* and *penumbra*. While the former refers to a highly hypoperfused tissue that is already infarcted (or is inevitably destined to become infarcted regardless of treatment), the latter represents hypoperfused tissue

*Corresponding author

Email address: `joacoseia18@gmail.com` (Joaquin O. Seia)

that is potentially salvageable through reperfusion (Vagal et al. (2019)).

Reperfusion therapy, and endovascular treatment (EVT) in particular, is an effective therapeutic solution for acute ischemic stroke (AIS) patients with a large vessel occlusion. However, it is currently restricted to patients with a small lesion core because the larger this region it is, the higher the risk of an hemorrhage following the reperfusion and the smaller the benefit the patient can get from it. (Byrne et al. (2019); Goyal et al. (2020)).

1.1. Medical images in the context of stroke

As mentioned above, selecting the correct treatment for a patient requires the assessment of the extent of the ischemic lesion, and medical imaging plays a vital role in this process. This is reflected in the current American Heart Association (AHA) guideline for the management of stroke patients, which recommends an emergency brain imaging evaluation before any treatment decision is made (Powers et al. (2019)). This recommendation states that non-contrast computed tomography (NCCT) should be a first-line imaging modality used to rule out intracerebral hemorrhage. Thereafter, patient selection for EVT should follow one of two imaging recommendations depending on the time to last known well. For hyperacute ischemic stroke, the Alberta Stroke Program Early CT Score (ASPECTS) should be computed over the NCCT and an angiographic CT (CTA) should be acquired. In AIS patients, a combination of CTA and perfusion CT (CTP) or magnetic resonance angiography (MRA) and diffusion-weighted magnetic resonance imaging (DWI) is recommended.

The guideline presents three principal non-angiographic imaging modalities: NCCT, CTP and DWI. Although the evaluation of an AIS stroke patient using NCCT alone is not recommended, it is an imaging modality with high potential for stroke lesion assessment. In order to identify what makes NCCT unique, it is necessary to introduce on some basic concepts of these three modalities.

Non-contrast computed tomography

NCCT measures tissue density. Brain tissue undergoing through severe ischemia appears hypodense on NCCT because of increased water content due to ionic edema (Goyal et al. (2020)). ASPECTS is a rating system that uses this biomarker to subjectively assess the extent of early infarction on the NCCT (Mokin et al. (2017)). Although the use of ASPECTS is currently part of the stroke management guidelines, it is characterised by a high inter-observer variability (Farzin et al. (2016)). It has been well described that the ischemic core signal is virtually absent in baseline NCCT images compared to other modalities, making the task of lesion segmentation challenging even for expert neuroradiologists (El-Hariri et al. (2022); Estrada et al. (2022)). Furthermore, this scoring system does not provide a fine quantification of the lesion volume but rather a coarse estimation of extent based on affected vascular regions.

Computed tomography perfusion

In this modality, the focus is placed on blood flow measurement rather than the consequences of ischemia in the brain parenchyma. Through a non-trivial post-processing step, measurements of penumbra and ischemic core volumes can be obtained based on the relative blood flow at each voxel. The recent inclusion of CTP among the AHA recommendations results from the successful use of CTP-derived core and penumbra volumes as part of patient selection criteria for EVT in two large clinical trials, Defuse 3 and DAWN (Albers et al. (2018); Nogueira et al. (2018)).

However, despite being useful, CTP is not free of complications. Differences in results between software solutions and difficulties inherent in the modality itself, such as patient motion or confounding physiological processes, can lead to over- or underestimation of core volume in CTP. To serve as an example, in AIS, CTP is prone to underestimation of baseline ischemic core in cases of *luxury perfusion*, where the core infarction becomes hyperemic because of spontaneous reperfusion or engorgement of the leptomeningeal arteries (Sotoudeh et al. (2019)). In order to rule out this false negatives, a simultaneous review of the baseline NCCT is required, leveraging the complementary information provided by the two modalities (Vagal et al. (2019)).

Diffusion weighted image

The ischemic stroke lesion appears as a high signal on the DWI scan because of the diffusion restriction in the extracellular space caused by the cytotoxic edema (Goyal et al. (2020), Kuang et al. (2021)). Contrary to NCCT, this biomarker is visible within minutes after ischemia onset and is much more conspicuous. Because of its limited availability, the higher cost and longer acquisition time, DWI is usually reserved for a follow-up evaluation and quantification of final infarct (El-Hariri et al. (2022)). These elements make DWI the gold standard for estimating the volume of the ischemic lesion.

However, it is important to note that even though the lesion core volumes computed on the baseline NCCT (or CTP) are highly correlated with those obtained from the DWI image, they may differ. For example, depending on the success of recanalisation or the time elapsed between the baseline image and the treatment, the final infarct extent will differ from the baseline lesion. In addition, very small ischemic lesions associated with small emboli generated during reperfusion may appear on the post-treatment image but not on the baseline image.

1.2. Why NCCT?

As presented, it is clear that advanced imaging techniques such as CTP and DWI can provide a more complete and accurate assessment of the ischemic lesion. However, these modalities are not available in hospitals on a 24/7 basis, and AIS patients are mostly diagnosed by using NCCT images (Kim et al. (2021)).

Despite the fact that baseline NCCT is not the best modality for quantifying the volume of the ischemic core, it is the first line imaging modality performed when the patient arrives to the emergency department, is the cheapest, the most widely available and the fastest technique among the mentioned ones. Consequently, the segmentation of ischemic stroke lesion core on baseline NCCT is a clinically relevant problem.

As stated in Bouslama et al. (2021), a proper quantification of the stroke core on baseline NCCT images could allow centres without advanced imaging techniques or specialised stroke neurologists to ensure access to endovascular therapy for a wider population of patients who could benefit from it. Even when following the current guidelines and using perfusion imaging, NCCT can still provide complementary information that can lead to improved healthcare provision.

In this context, where manual segmentation of stroke lesions on baseline NCCT images is not feasible but would represent an improvement in the management of AIS patients, the development of accurate automated segmentation tools for ischemic lesions on baseline NCCT is a problem that needs to be addressed.

2. State of the art

2.1. Automatic segmentation of AIS lesions

Over the last decade, machine learning, and in particular deep learning (DL), has been successfully applied to many image segmentation taks. The work in Isensee et al. (2020), which presented a robust model that achieved state of the art (SOTA) performance over 53 different medical image segmentation problems, can serve as a clear example of this. The segmentation of acute ischemic lesions has not been the exception, where several methods have tackled the task in MRI images (Clèrigues et al. (2020)) or CTP scans (Amador et al. (2021, 2022); Robben et al. (2020)).

In contrast, on baseline NCCT images, there are only a few well-established approaches for segmenting AIS lesions. Among these solutions, there is a large heterogeneity in their experimental designs, which affects how comparable and transferable they are to other NCCT datasets. Overall, there are three main elements transversal to the literature on this topic:

1. The vast majority of deep learning proposals have successfully applied UNet-like deep convolutional neural networks (DCNN).
2. The use of contextual information in the model design improves the segmentation results. In most cases, inter-hemispheric asymmetries are used as one of the forms of contextual information.
3. The lack of large, well-curated and publicly available datasets containing baseline and follow-up AIS images is not negligible, as most of the publications have worked with private datasets with different patient selection criteria.

In the following for each of this three aspects some salient publications are commented.

2.1.1. U-Net like architecture choice

Among the many works using U-shaped architectures, in Ostmeier et al. (2022) and El-Hariri et al. (2022), the self-configuring model nnU-Net (Isensee et al. (2020)) was shown to be successful in AIS lesion core segmentation on baseline NCCT images. In the first case, the authors showed that nnU-Net achieved non-inferior segmentation results compared to expert neuroradiologists. In the second case, the authors not only showed that nnU-Net was able to achieve high volumetric agreement with ground truth pre-treatment DWI labels, but also pointed out that their model was already part of commercial software, demonstrating the impact this architecture already has in the clinical practice.

2.1.2. Exploiting contextual information

Contextual information has been incorporated into models in many different ways in the literature. Chen et al. (2022) and Kuang et al. (2019) used the difference images generated after a sagittal flipping of the NCCT images. The first one opted for a 2D U-shaped architecture in which the original, flipped and difference images were given as a multi-channel input. The second one opted for a more sophisticated approach using a 3D U-shaped architecture with a multi-path encoder. Four paths were used, covering the original image, the difference image, an infarct location probability map and a distance-to-cerebrospinal-fluid map. In Ni et al. (2022), a 3-step end-to-end trainable 3D asymmetry disentangling network was used to obtain an effective and interpretable AIS segmentation on NCCT. Their method automatically separated pathological asymmetries and intrinsic anatomical asymmetries from the NCCT.

An approach usually referred to as SOTA in AIS core segmentation in baseline NCCT is the work of Kuang et al. (2021). The authors proposed a multi-task learning approach, called EIS-Net, which was simultaneously trained to segment the stroke lesion and to predict the ASPECTS score from the NCCT. Their model consisted of a 3D U-shaped segmentation CNN architecture with a triple-path encoder. Each path was fed with the NCCT, the sagittally mirrored NCCT and a CT atlas, respectively. The differences between these features were exploited using an ad hoc comparison block. The use of contextual information within a multitask optimisation strategy allowed them to achieve better results than using the plain U-shaped segmentation model.

2.1.3. Data scarcity problem

Overlapping with the previous remark, the work in Giancardo et al. (2023) presents another model that exploits the inter-hemispheric differences. However, the remarkable aspect of this paper is that the authors identified that one of the elements that is delaying the development of automatic AIS segmentation in NCCT/CTA is the difficulty in obtaining large enough samples containing high-quality DWI images with voxel-level ground truth annotations. To circumvent this problem, they used only image-level labels (the stroke core volume size) to train their model and obtained a competitive AIS lesion core segmentation on CTA images.

2.2. Deep learning with labelled data scarcity

In the recent years, self-supervised learning (SSL) strategies have gained popularity for addressing the problem of scarcity of labelled data. As described in the work of Balestriero et al. (2023) SSL stands for a collection of machine learning approaches that can learn from large amounts of unlabelled data. The common practice implies the definition of a pretext task based on unlabelled inputs to produce descriptive and meaningful representations that can be used across different downstream tasks. Self-supervised image representation learning has shown amazing progress in the last five years, achieving a performance in several downstream tasks that is competitive or even superior to supervised learning approaches (Bardes et al. (2022); Caron et al. (2021); Chen and He (2021); Grill et al. (2020); He et al. (2022, 2020)).

One of the most commonly used SSL methods is based on a joint embedding architecture, where two Siamese networks are trained to produce similar embeddings for different views of the same image. In this way, the networks learn to extract semantically meaningful information from the images themselves. The main difficulty in this approach is to avoid *representation collapse*, phenomenon where the networks ignore the inputs and produce identical and constant output vectors.

Recently, among the several existing ways to avoid model collapse, *distillation methods* have been pointed out as achieving better performance than others (Balestriero et al. (2023); Bardes et al. (2022)). In general, distillation methods train a student network to predict the representations of a teacher network. During the training phase, the gradients are only back-propagated through the student network, and the weights of the teacher are a running average of the weights of the student.

One of the most representative distillation SSL methods is the work of Caron et al. (2021). The authors designed an approach termed DINO as an acronym of "*knowledge **di**stillation with **no** labels*". DINO simplified SSL training by optimising the matching of the teacher network's output using a standard cross-entropy loss. Collapse prevention was achieved by including two simple operations in the teacher output, known as *centring* and *sharpening*. DINO could work on both transformer and convolutional architectures achieving SOTA accuracy on ImageNet. More interestingly, the trained encoders could obtain feature representations that explicitly contained a scene layout of the image, which could be used to generate accurate segmentation masks.

These promising models have also found their way into the field of medical imaging. Among the many papers applying SSL to medical images (Jiang et al. (2023); Kalapos and Gyires-Tóth (2023); Manna and Chakraborty (2022)), the work presented in Ye et al. (2022) deserves special attention. In this article, the authors proposed a DINO-based SSL method, called DeSD (**de**ep **s**elf-**d**istillation), which allowed the use of unlabelled data in the context of 3D medical image segmentation. In their model, both a student network and a momentum teacher were built by stacking several sub-encoders. The deep self-distillation supervision implied that the features of every student sub-encoder were optimised to match the teacher's output distribution. This technique resulted in superior pre-training of the segmentation network encoder compared to other existing SSL methods. When tested on seven downstream 3D medical segmentation datasets, their method outperformed training the same segmentation architecture from scratch and achieved state of the art results.

Considering the challenges associated with baseline NCCT AIS lesion segmentation, SSL can be identified as a promising approach to address the problem. The capacity of self-supervised learning techniques to make models extract semantically meaningful information coming from unlabelled datasets, represents a promising approach as a pre-training strategy in the context of AIS lesion segmentation. Given that SSL models have been shown to capture the scene layout of images, these methods may represent an unexplored way to exploit the intrinsic contextual information present in the brain NCCT image that is not necessarily related to the AIS lesion label.

In addition, these methods open the possibility of repurposing large amounts of unused, unlabelled NCCT images (with or without stroke) to improve segmentation performance. In a context of scarcity of good quality labelled AIS datasets, SSL could represent a more efficient use of the manual labelling process, limiting it to a subset of cases used for fine tuning to the downstream segmentation task and validating the results.

Lastly, U-shaped architectures and in particular nnU-Net, represent a standard segmentation baseline across many medical imaging modalities which has also been successful in the context of AIS lesion core segmentation. Therefore, the integration of SSL as a pre-training strategy for nnU-Net encoder, in a similar manner to that done in Ye et al. (2022), may represent a synergistic way to integrate the aforementioned benefits of SSL with the highly successful self-configuring nnU-Net pipeline.

2.3. Contributions

1. A robust pre-processing pipeline for NCCT images which can work across many different datasets of variable quality.
2. A systematic pipeline to enhance nnU-Net autotuning framework with an additional encoder pretrainining in a self-supervised manner.
3. Use of a DeSD-like self supervised pre-training strategy to exploit large unlabelled NCCT (stroke positive and negative) datasets for the task of AIS segmentation.
4. Introduction of an asymmetry based data augmentation technique for achieving better latent representations for the context of stroke lesion segmentation on NCCT.
5. Pre-training nnU-Net's encoder with SSL is found to be an effective way for exploiting large amounts of unlabelled datasets, improving AIS final infarct segmentation performance on baseline NCCT images.

3. Material and methods

3.1. Datasets

In this work four datasets were utilised. A detailed description of each of them is presented below.

Acute Ischemic Stroke Dataset (AISD)

This dataset, published in Li et al. (2021), included cases of AIS with less than 24 hours from symptom onset to NCCT acquisition (n=397). For each case, NCCT, DWI and manual stroke lesion segmentation were provided. NCCT and DWI images were not registered. Patients underwent DWI within 24 hours of CT acquisition. Ground-truth labels were delineated on the NCCT by a physician using the MRI images as a reference. Important clinical information was missing from this dataset, such as the timing of DWI acquisition (pre/post endovascular treatment). As a result, it was not possible to determine whether the provided ground truth corresponded to final infarct or pre-treatment stroke core lesions. After visual inspection, five cases were discarded from the dataset due to large motion artefacts or non-overlapping ground truth with baseline imaging.

A Paired CT-MRI dataset for Ischemic Stroke Segmentation (APIS)

This dataset corresponded to the publicly released training subset of the ISBI 2023 APIS Challenge (Gómez et al. (2023)). The dataset (n=60) included patients over 18 years of age, collected from two Colombian clinics (FOSCAL and FOSUNAB), eight of whom were healthy controls. Each case included an NCCT image, the apparent diffusion coefficient (ADC) map derived from the DWI and a manual delineation of the stroke lesion core. No treatment was applied between NCCT and ADC (stroke lesion core as ground truth). Two neuroradiologists with more than five years of experience delineated the affected tissue over the DWI/ADC images. Eight cases were discarded due to high image corruption (missing slices, evident lesion mask misplacement), leaving a total count of forty-four stroke-positive cases.

In APIS dataset, NCCT and ADC maps were provided registered and skull-stripped. However, after a visual inspection, the registration process was found suboptimal. This had two negative consequences: non-brain structures (e.g. bone) were visible in the skull-stripped image and the labels were misregistered.

icometrix Acute Ischemic Stroke Dataset (icoAIS).

This was an in-house private set of cases provided by the Klinikum rechts der Isar (Munich, Germany). The collection included acute and early subacute stroke patients (n=159) and healthy patients (n=8), all over 18 years of age. Three images were available for each case: NCCT, DWI and ADC map. The MRI images were provided already skull-stripped and registered to a common space. MRI images were acquired after successful revascularisation therapy. The collection included a wide range of infarct patterns in all vascular territories, even including posterior circulation infarcts, which are common in clinical practice but not commonly studied in the literature.

Ground truth labels for icoAIS were obtained in two different ways. In half of the cases (n=79), the voxel-level labels involved a high-quality hybrid human-algorithm annotation process described in Hernandez Petzsche et al. (2022). Since the process involved neuroradiologists with more than ten years of experience reviewing the MRI images, this subset was referred to as *gold standard* labels. For the remaining stroke-positive cases (n=80), *silver standard* labels were obtained by running SEALS -the publicly available[1] ISLES22 winning solution- over the DWI images. A stroke expert reviewed the annotations and found them to be adequate and highly correlated with the available gold standard annotations.

Collaborative European Neuro-Trauma Effectiveness Research in Traumatic Brain Injury Dataset (CENTER-TBI)

This collection of NCCT images was a multi-centre, multi-scanner dataset presented in Maas et al. (2015). From the complete dataset, only a selection of NCCT scans identified by expert review as not having abnormal TBI-related findings was kept (n=637). This re-

[1] https://github.com/Tabrisrei/ISLES22_SEALS

sulted in a very diverse dataset of healthy (non-stroke) patients.

In all cases, due to the retrospective nature of this work and the rigorous anonymisation of the data, it was not necessary to obtain informed consent from the patients. In the particular case of APIS, the data was used in agreement with the APIS challenge signed informed consent.

3.1.1. *Dataset partitioning*

The complete dataset was split into three subsets: training, validation and test (70-20-10% respectively). The partitioning was done individually for each dataset at the patient level. For stroke-positive datasets, the partitioning was done stratified by lesion location and size. This ensured that all subsets had equal representation of right, left and bilateral lesions and lesion sizes. In the specific case of icoAIS the dataset was also stratified by the labelling standard (gold vs. silver).

3.2. *Preprocessing*

As a first pre-processing step, all the images were turned into NIfTI format. Cases that where acquired in "tilted" fashion were "un-tilted" during the DICOM to NIfTI conversion using the NITRC conversion tool (Li et al. (2016)).

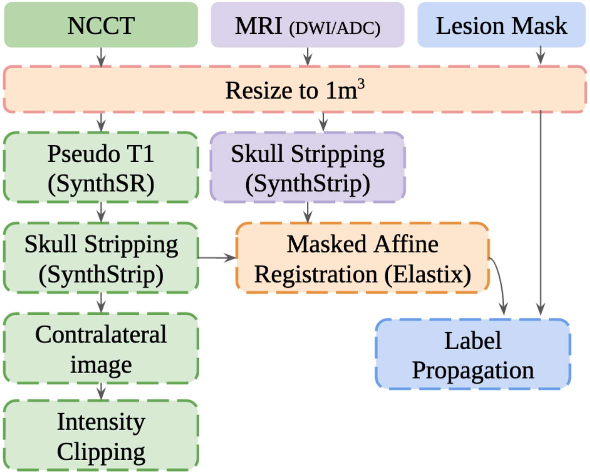


Figure 1: Preprocessing pipeline

It is noteworthy that there were significant differences between the datasets used. Diversity is desirable in the context of SSL, but to reduce the risk of the algorithms exploiting meaningless shortcuts or biases (i.e. distinguishing data origins by their skull-stripping quality), the datasets were homogenised as much as possible. To do this, a robust preprocessing pipeline (summarised in Figure 1) was applied to all datasets. It comprised six steps:

a. Resampling. All volumes were resampled to $1mm^3$ resolution using a linear interpolation for NCCT, ADC and DWI images and a nearest neighbour interpolation for the ground truth mask.

b. Skull Stripping. NCCT and DWI/ADC images are not characterised by having high contrast differences between the different brain soft tissues. As a consequence, popular skull stripping methods (Ashburner and Friston (2005); Isensee et al. (2019); Lutkenhoff et al. (2014)) mainly developed to work on high resolution T1 MRI images gave poor results when applied to the desired modalities. In addition, for APIS, all methods failed due to the pre-existing sub-optimal preprocessing. Robust results were obtained by combining two models from the publicly available FreeSurfer toolbox: SynthSR and SynthStrip (Hoopes et al. (2022); Iglesias et al. (2023)).

In both cases, the respective authors used a clever synthetic data generation technique to obtain robust models across multiple resolutions and contrasts. The authors show that SynthSR is able to generate a high-resolution T1 MRI out of any brain MRI image and has a reasonable performance when using CT scans. SynthStrip is a brain segmentation model that is very robust across different brain MRI modalities, but it did not work very well when applied directly to the NCCT image. Instead, generating a pseudo-T1 first and applying SynthStrip over it gave the best results. SynthStrip was applied directly to ADC and DWI MRI images with good results.

c. Registration. For APIS and icoAIS, a brain-masked affine registration of the MRI image to the NCCT space was performed using the Elastix toolbox (Shamonin (2013)). This involved a pyramidal registration with mutual information as the objective function. After this, the stroke lesion mask was propagated to the NCCT space using the same transformation but with nearest neighbour interpolation.

d. Contralateral image. To obtain the mirrored brain with respect to the inter-hemispheric plane, the NCCT was first registered to an NCCT MNI space template (Rorden et al. (2012)) using brain-masked affine registration. A left-right flip was then performed and the resulting image was masked-affine registered to the original NCCT image. In this way, the desired image ended in the original patient space and the effect of gross normal asymmetries in brain shape was reduced.

e. Intensity clipping. Following the literature and the recommendations done by an expert in stroke imaging, the NCCT and the contralateral NCCT images were intensity clipped to the range [-100, 400], leaving unchanged the range in which both brain soft tissue and stroke lesions have their intensities.

3.3. *Method overview*

As mentioned in the introduction, this work uses a variant of the two-step SSL paradigm presented in Ye et al. (2022). Unlike the cited work, the nnU-Net self-configuring model is used as the base architecture and is enhanced by adding SSL pre-training to its encoder part.

Introduced in the work of Isensee et al. (2020), nnU-Net is a self-configuring method for deep-learning biomedical image segmentation. In terms of implementation, it consists of a robust pipeline with two stages: first, it finds the best configuration of the UNet model for any new dataset, and then, based on the conclusions of this step, the tailored training can be performed. In the first step, the data pre-processing, network architecture, training details and post-processing stages are decided. Two core elements are involved in this process: a *dataset fingerprint* and a *pipeline fingerprint*. The first one is a standardised dataset representation comprising key properties such as the image size, voxel spacing and class ratios. The second one registers a collection of choices made during the automatic optimal method design (i.e. batch size, patch size, network topology, etc.) and serves as a recipe followed during training to instantiate the model and the training machinery.

Given the success of this self-tuned pipeline, we used it as a basis for a four steps training strategy:

1. nnU-Net configuration according to the dataset used for SSL.
2. Self-Supervised pre-training of nnU-Net encoder.
3. nnU-Net configuration according to the dataset used for supervised learning.
4. Transfer learning and supervised training of nnU-Net segmentation network.

In the first step, the objective was to use the optimal architecture configuration and data pre-processing of nnU-Net, but for self-supervised training. To this end, an nnU-Net dataset containing the full set of NCCT images (stroke and healthy cases) was generated and the nnU-Net self-configuration pipeline was run over it. By default, during the dataset fingerprint extraction, nnU-Net computes the intensity statistics -that are later used for normalising the images- from foreground region defined by the segmentation mask. To be able to process healthy cases and to obtain a more general normalisation strategy, nnU-Net pipeline was fed with the brain masks as if they were the segmentation targets, so that all the whole brain region was treated as foreground. Three results were kept from this step: the pre-processed images, the dataset fingerprint and the pipeline fingerprint.

In the second step, the pipeline fingerprint was used to instantiate the complete UNet model, discarding the decoder part and adding the additional modules required for deep self-distillation training (see details in Section 3.4). This distillation mechanism was implemented to dynamically adapt to the number of encoding steps defined by the nnU-Net pipeline. Finally, the encoder was trained in SSL fashion and its weights were re-adjusted (removing the extra SSL modules) to the original nnU-Net encoder structure.

In the third step, the self-configuring nnU-Net pipeline was run a second time. This time, only the subset of desired labelled images were included and the stroke lesion masks were given as segmentation targets. The resulting dataset fingerprint was modified by replacing the intensity statistics with those from the full NCCT dataset, so that the pre-processed inputs were in the same intensity space used to pre-train the encoder. The preprocessing was then run and the pipeline fingerprint was generated.

Finally, the nnU-Net supervised training pipeline was run using the pre-trained weights of the encoder as a starting checkpoint.

3.4. Self-supervised learning details

The SSL method implemented in this work shared the same overall structure with DeSD (Ye et al. (2022)). Its general outline can be appreciated in Figure 2. The overall strategy was based on knowledge distillation, training a student network g_{θ_s} to match the output of a teacher network g_{θ_t} (where θ_s and θ_t were their parameters respectively).

Both networks roughly shared the same architecture, but the student one was decoupled into N sub-encoders: $g^i_{\theta_s}$ for i from 1 to N. The network g_{θ_s} was generated by adding a projection head at the end of each of the 6 encoding stages determined by the nnU-Net configuration pipeline. Naming the projection head h, each stage of nnU-Net encoder f^i $(i = 1, ..., N)$ and the complete encoder f, the overall network could be formally written as: $g^i_s = h^i \circ f^i_s$. The same holds for the teacher network, but using only the complete encoder: $g_t = h_t \circ f_t$.

In all cases, the projection head (h) consisted of a multi-layer perceptron (MLP) of three layers plus a final weight-normalised fully connected layer of dimension K. The two MLP hidden layers were of dimension 2048, with batch normalisation and Gaussian Error Linear Units (GELU) activation function. The MLP output had a dimension of 256, with no batch normalisation or non-linearity applied. In summary, the complete projection head mapped the output of each stage of the student network and the final teacher network into a K dimensional representation, with $K = 65536$ in our case.

As done in Caron et al. (2021), after the K dimensional teacher representations were obtained, a centring was applied to avoid model collapse. The centring depended on the first order batch statistics and can be interpreted as adding a bias term c to the teacher: $g_t(x) \leftarrow g_t(x) + c$. The centre c was updated with an exponential moving average rule:

$$c \leftarrow mc + (1 - m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i), \tag{1}$$

where $m = 0.9$ and B is the batch size.

Given an input image x, the resulting representations from the N=6 student sub-encoders and the teacher one

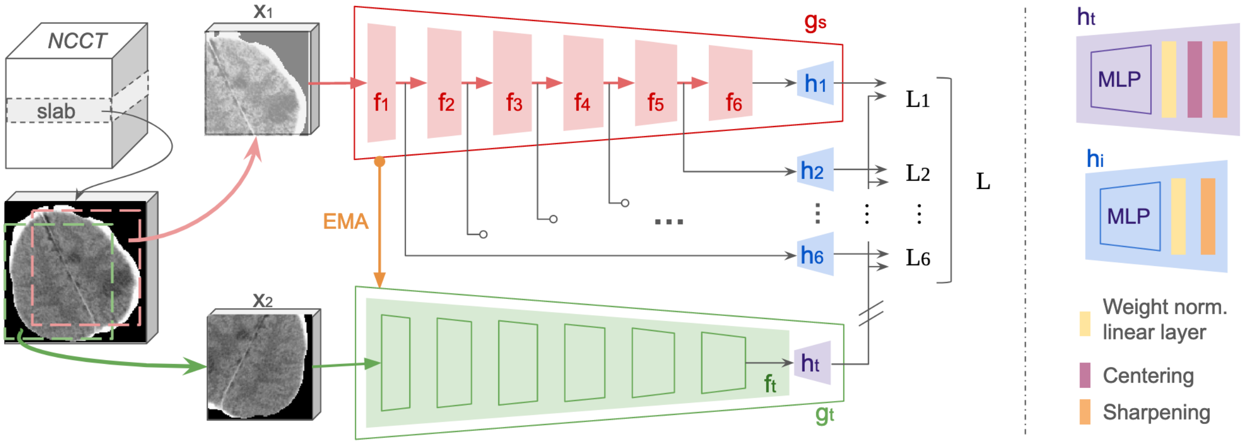


Figure 2: Self-supervised training method, see Section 3.4 for details. Subindex *s* and *t* stands for student and teacher respectively, *L* for loss, *EMA* for exponential moving average and *MLP* for multi-layer perceptron.

were transformed into probability distributions over the K dimensions (denoted P_s^i for $i = 1, ..., N$ and P_t respectively). Each probability P was obtained by normalising the output of the corresponding network g with a softmax function:

$$P(x)^{(j)} = \frac{exp(g(x)^{(j)}/T}{\sum_{k=1}^{K} exp(g(x)^{(k)}/T}, \qquad (2)$$

where $j = 1, ..., K$ and $T > 0$ was a temperature parameter that, as indicated in Caron et al. (2021), had a sharpening effect that reduced the possibility of representation collapse.

As usually done in SSL, from a given image, a set $V = \{x_1, x_2\}$ was generated with two differently distorted views (crops) of it. Then, x_1 and x_2 were passed through both the student and the teacher networks, in order to obtain their respective embeddings.

During training, only the student weights were updated through gradient back-propagation. The training objective was to match each student embedding with the teacher's one by minimising the cross-entropy loss:

$$\min_{\theta_s} \frac{1}{N} \sum_{i=1}^{N} \sum_{x \in V} \sum_{\substack{x' \in V \\ x' \neq V}} H(P_t(x), P_s(x')), \qquad (3)$$

where $H(a, b) = -a \log b$.

The weights of the teacher network were updated using an exponential moving average of the student ones (momentum encoder). The updating rule was defined by: $\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$. Where λ followed a cosine schedule from 0.9996 to 1 during training.

Finally, once the model was trained, the features used in downstream task are the ones from the backbone f_t, dropping the projection head.

3.4.1. Distorted views strategy

The SSL literature suggests that inputs to Siamese networks should follow two recommendations: The view size should contain more than 50% of the original image, and that the larger the batch size used during training, the better. However, as the proposed models had 3D inputs, which increased their GPU memory consumption, a trade-off was made between these two requirements. An anisotropic patch size of 112x112x16 (favouring inter-hemispheric contextual content) and a batch size of 64 (the largest the memory would hold) were chosen.

Given that the median image size in our datasets was 169x138x139, to avoid sampling two completely different views from the volume, an online patch sampling strategy of two stages was used. First, from the region defined by the brain's bounding box, a slab of 24 slices was sampled along the z axis and then the two patches x_1 and x_2 were sampled from within the slab. This ensured a high degree of overlap between the paired views. In addition, to prioritise slabs with higher brain content, if the sampled slab contained less than 40% of brain, it was replaced by a new sample with a probability equal to the background content percentage (the less brain content, the more chances to take another sample).

Several data augmentations were applied to both views: flipping, scaling, Gaussian noise, Gaussian blur, gamma intensity transformation, change of image brightness and contrast. These were the same ones used in Ye et al. (2022) but applied on a volume fashion and not slice-wise.

3.4.2. Training and evaluation details

Evaluating the progress of SSL methods is not a simple task. Common ways to evaluate the quality of the obtained representations are linear probing with a classification task, k-nearest neighbours clustering or directly training over the downstream task at each epoch. In our case, training the decoder stage of the segmentation network at each epoch was computationally prohibitive. For this reason, the training dynamics were evaluated by checking the training and validation loss curves and using *RankMe*.

RankMe is a recently proposed metric to evaluate SSL training performance in a fully unsupervised manner (Garrido et al. (2023)). It represents an indirect way to evaluate the information content of a set of representations by assessing an approximation to their rank. The higher the RankMe, the more linearly independent the representation components are and more the variance of the data is distributed among them, showing an empirical positive correlation with their performance in different downstream tasks. Following the cited publication, RankMe was computed over the latent representation of 256 elements produced by the MLP of the teacher's projection head.

Preliminary experiments were carried out to become familiar with the learning dynamics of the models and to find a suitable set of hyperparameters and training choices.

An AdamW optimiser was used with a batch size of 64. The learning rate was linearly ramped up during the first 10 training epochs to its base value of 0.1. After this warm-up the learning rate was decayed with a cosine schedule. Weight decay also followed a cosine schedule from 0.04 to 10^{-5}. The temperature T_s was set to 0.1 while a linear warm-up from 0.04 to 0.07 during the first 10 epochs was used for T_t. In order to reduce memory consumption 16-bit floating precision was used for the model weights. Independently on the dataset used, in all the experiments SSL was trained for 100 epochs after which train loss, validation losses and RankMe curves reached a plateau. Each epoch consisted of 9600 iterations. Since there were no clear signs of overfitting or model performance decay (or improvement) at the final plateau, the model from the last epoch was kept as the pre-trained checkpoint.

3.4.3. Latent representation projection and model interpretability

After training the encoder in the SSL fashion, two techniques were used to interpret the obtained representations: attribution maps and t-SNE dimensional reduction to observe clustered patterns.

Attribution maps were computed using Integrated Gradients method (Sundararajan et al. (2017)). Since our network did not output class wise predictions, a sum of all the elements in the final representation was added at the end of the network before applying the attribution method. The resulting maps highlighted the voxels whose change most affected the entire latent representation.

To analyse the projections in a more systematic way, a t-distributed Stochastic Neighbour Embedding (t-SNE) projection of the teacher representations onto a two-dimensional space was obtained (van der Maaten and Hinton (2008)). Given the resulting clustered nature of the resulting 2D space, a set of representative points of the visible groupings were selected and their six nearest neighbours were obtained. For these 6 cases, axial slices of the NCCT volumes and their attribution maps were plotted with the aim of detecting common salient features used by the network to generate the representations (see Figure 8 for an example).

3.5. Supervised Segmentation details

Among the different experiments performed in the supervised segmentation training, the main objective was to compare the impact of the different training strategies on the performance of the final infarct segmentation on baseline NCCT images. In this sense, we decided to focus on the icoAIS dataset for supervised segmentation. This was the best described dataset, the one with the best image quality and the only one that was certain to contain only final infarct ground truth.

As a first step, several experiments were performed to train the nnU-Net model from scratch with the icoAIS dataset to gain insight into the training dynamics. From these experiments it was found that the 3D nnU-Net architecture outperformed the 2D version and therefore the former was retained throughout the work.

After running the nnU-Net self-configuring pipeline over the dataset used for supervised learning, the resulting model architecture is illustrated in Figure 3. The model details follow the overall rules defined in Isensee et al. (2020). However, the main design specifications are presented below.

The model consisted of six encoding and six decoding stages 3D U-shaped architecture, using only plain convolutions, instance normalisation and Leaky ReLU non-linear activation function. Each resolution stage of both the encoder and the decoder consisted of two computational blocks of convolution-normalisation-activation function. Down-sampling was done with strided convolutions and up-sampling with transposed ones. The initial number of feature maps was set to 32 and doubled (halved) with each down-sampling (up-sampling) operation, the number of feature maps across the network was capped at 320 to limit the chance of overfitting.

During training, each epoch implied 250 iterations, where the minibatch size was 2. Stochastic gradient descent with Nesterov momentum ($\mu = 0.99$) and an initial learning rate of 0.01 was used to optimise the network weights. The learning rate was decayed through the training according to the 'poly' learning rate policy, $(1 - epoch/epoch_{max})^{0.9}$. The loss function was the average of binary cross-entropy and soft dice losses.

The network was trained with deep supervision, where additional losses are added in the decoder at all but the two lowest resolutions, each using a down sampled version of the ground truth mask. The training objective was the sum of the losses at all resolutions, $L = w_1 x L_1 + ... + w_4 x L_4$, where the weights $w_i = \frac{1}{2^i} w_1$ were later normalised to sum to 1.

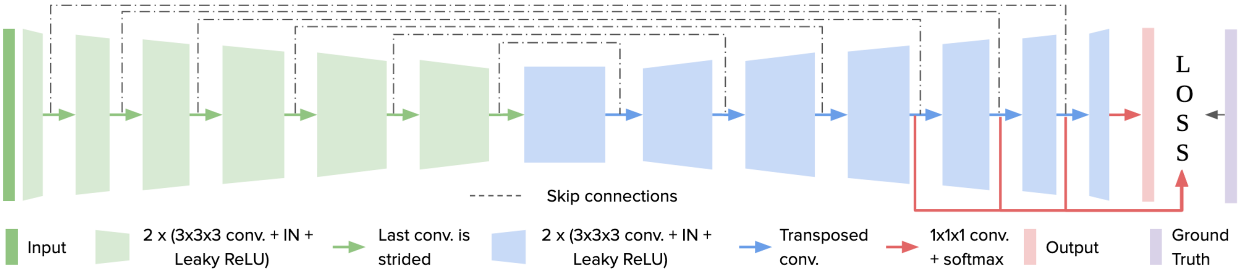


Figure 3: Alternative representation of the U-Net model, highlighting the encoding-decoding nature of the architecture

Since the resampling was performed as part of our preprocessing pipeline, the nnU-Net preprocessing consisted only of normalising the images using the foreground voxels statistics. The 0.5 and 99.5 percentiles were used for clipping, and the post-calculated mean and standard deviation were used for z-score normalisation. Samples for the mini-batches were selected from random training cases. Class unbalance was handled with over sampling by forcing that at least half of the samples in the minibatch had to contain stroke on it. The patch size used was 160x160x96. A variety of data augmentations were applied on the fly during training: rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, low resolution simulation, gamma correction and mirroring.

Unless otherwise stated, all networks were trained for 100 epochs. After these epochs, the model trained from scratch on the icoAIS dataset reached a plateau on the validation exponential moving average (EMA) pseudo-dice, and the validation and train loss curves began to show overfitting patterns (the former stopped decreasing, while the latter continued to decrease). In all cases, the models were trained on the training set, model selection was done with the validation set, and the test set was left aside in both procedures.

The nnU-Net pipeline retains the best model from all training epochs as the one that maximises the pseudo-dice EMA over the validation set. This dice approximation is computed by considering each batch of samples from the validation cases as a case itself. This strategy was shown in the original paper to be a good compromise between computational cost and performance impact. However, in the problem addressed in this work, it resulted in a very noisy validation pseudo-dice curve, which - even when smoothed by the EMA - affected the selection of the best model. In the preliminary experiments, it was observed that when the model was trained for 300 epochs, the best selected model did not achieve the highest performance over the full images in the validation set. This suggested that, in our specific problem, nnU-Net tended to select slightly overfitted models as best, which was “manually” prevented by limiting training to 100 epochs.

Another preliminary finding was that when nnU-Net supervised training was repeated on the same cases and experimental design, but with different random initialisations, there was significant variability in performance on the full validation cases. In an attempt to minimise this confounding noise, three different runs with different random initialisations were made for each supervised experiment. For each of them, the best model was selected according to the nnU-Net criteria and the majority voting ensemble was computed from the predictions of these three models.

Finally, during inference time, images were predicted using a sliding window approach with a window size of 96x160x160 and a stride of 48. Gaussian importance weighting was applied, increasing the weight of central voxels in the softmax aggregation. Test time data augmentation was applied by mirroring all axes.

3.5.1. Training nnU-Net with pretrained weights

When performing transfer learning from a pre-trained SSL model to a downstream task, two main strategies can be used to prevent the encoder from “forgetting” what it has learned: freezing the weights of the pre-trained encoder (for a fraction of epochs or for all epochs) or using smaller learning rates for the encoder. In order not to modify the nnU-Net pipeline too much, we decided to go for the freezing strategy. Four strategies were briefly explored: leaving all parameters unfrozen, unfreezing the encoder after 33 epochs or after 66 epochs, and leaving the encoder frozen for the entire training procedure. The best resulting strategy was to leave all parameters unfrozen and was therefore used in all experiments presented in this work.

3.6. Metrics

The performance of the models in the segmentation task was evaluated using a set of different metrics. Among the metrics commonly reported in the medical imaging community, the Dice Score Coefficient (DSC) and the 95% Hausdorff Distance (HD95) were used. The former evaluates the voxel overlap between the segmentation and the ground truth, ignoring the true negatives in the background, however Dice is not well suited to detecting outliers in the contour prediction. HD compares the greatest distance between predicted and ground truth contours. For both the DSC and HD95, calculated at subject level, we report the mean, median and interquartile range

From a clinical perspective, since lesion volume plays a key role in patient selection criteria, it is important to measure how close the predicted volume was to the ground truth. In this sense, three measures were included: absolute volume difference, volumetric Spearman correlation and interclass correlation coefficient (two-way mixed effects, single rater consistency definition (ICC(3,1)) (Koo and Li (2016)). All metrics were calculated at the subject level.

Since the ground truths are derived from post-treatment DWI images, minor embolic lesions of no clinical significance and characterised by volumes $< 3mL$ may be included in the ground truth whereas they were not present in the baseline image. Following Giancardo et al. (2023) and the suggestions from a stroke imaging expert, we decided to report the metrics for two different scenarios. One where all lesions are retained in the ground truth and a second one, where only lesions larger than 3 mL are retained.

For this two scenarios both the results on the independent validation and test sets are reported in the results section.

3.6.1. Statistical Analysis

Dice scores across experiments were statistically compared using the Wilcoxon signed-rank test after rejecting normality with the Shapiro-Wilk test and QQ plot analysis.

3.7. Experiments

3.7.1. Baselines

The effect of using an SSL pre-training strategy was compared against two baselines:

- Training the nnU-Net model from scratch using only the icoAIS dataset, run hereafter referred to as **FS-STKi**, following the convention (training mechanism)-(supervised dataset), where *FS* stands for From Scratch, *STK* for stroke and *i* for the particular icoAIS dataset.

- Training the nnU-Net model from scratch using all the available labelled data coming from the three stroke positive datasets: AISD, APIS, icoAIS. Similar as before, this run is referred to as **FS-STKp**, where the *p* stands for positive.

It is important to note that for the second baseline, the training was conducted for 300 epochs. Because of the increase in the training data, the training convergence took longer. At around 300 epochs, the same rationale mentioned for selecting 100 epochs was fulfilled.

3.7.2. Exploring different datasets for SSL pre-training

In order to evaluate the benefits of using the pre-trained encoder, several experiments were carried out. The first one involved studying the influence of using different datasets during SSL pre-training. Three scenarios were investigated:

- Training on all the available NCCT images (abbreviated as *"ALL"*). This involved using the AISD, APIS, icoAIS and CENTER-TBI datasets, with the hypothesis that including the greater diversity of images in the dataset would allow the model to learn better representations.

- Training with all stroke-positive NCCT images (*STKp*). This involved using the AISD, APIS, icoAIS datasets, hypothesising that including of only stroke-positive images might allow the model to somehow capture better suited representations for the problem under assessment.

- Training only on healthy/non-stroke patients (*STKn*). This implied using only the CENTER-TBI dataset, with the idea that pre-training on healthy patients and fine-tuning on stroke-positive cases might allow the model somehow exploit the difference (presence of lesions) between these image sets.

For all these different datasets, the SSL pre-training was done as described in Subsection 3.4.2. Once the encoder was pre-trained, the supervised training was done *using only the icoAIS dataset*. In the following, these supervised pre-trained experiments are identified respectively as **ALL-STKi**, **STKp-STKi**, **STKn-STKi**, following a convention close to the one defined above (SSL dataset)-(supervised dataset).

3.7.3. Symmetry focused data augmentation technique

In self-supervised learning the data augmentation techniques used to generate the two different views from the same image play an important role in the representations learned by the model. In an attempt to integrate the inter-hemispheric asymmetries more explicitly into the SSL pre-training pipeline, a specific data augmentation technique was developed. In detail, when a patch was sampled from the original NCCT volume, the same patch location was sampled from the contralateral image and both views were later subjected to the regular data augmentation techniques.

The idea behind this experiment was that with this augmentation, the resulting model representations would be more agnostic to brain asymmetries. Derived from this, training the encoders with healthy patients only, the model would disregard normal asymmetries, which could have an impact when trained in a supervised manner with the presence of stroke-induced asymmetries.

To explore this idea, we pre-trained the encoder with the three dataset configurations presented in the previous section, but added the symmetry augmentation with a probability of 0.7 each time an image was sampled (regardless of its stroke content).

Once the encoders were pretrained, the complete supervised nnU-Net was trained using only the icoAIS dataset. These experiments are identified as **ALL-STKi-SA**, **STKp-STKi-SA**, **STKn-STKi-SA**, following a close convention to the one defined above (SSL pretraining dataset)-(supervised dataset)-(symmetry augmentation).

3.7.4. Additional experiments

An early phase of this work involved the participation in the ISBI 2023 APIS Challenge. Following the reported benefits of including inter-hemispheric differences in ischemic stroke segmentation models, a 3D nnU-Net was trained over the complete stroke-positive training set (*STKp*) with the same specifications as above, but using a different input. The NCCT and the difference from its sagittally mirrored version were used as a multi-channel input to the model. This model achieved first place in the competition by a wide margin over the other participants (for further details see Appendix A).

With the aim of combining this initial achievement with the main line of research of this work, a further set of experiments was carried out using the multi-channel 3D input to train both the encoder with SSL and the complete nnU-Net in a supervised fashion (with or without pre-training). For the sake of clarity, the details and results of these side experiments are included in Appendix B.

3.8. Computational resources

All the models were implemented using Python 3.10.9 and PyTorch 2.0. All the experiments were run on a 64-bit GNU/Linux (Ubuntu 20.04) server with an 8-core AMD EPYC 7R32 CPU (2.8 GHz) with 32 GB of RAM and a single NVIDIA A10G GPU card with 24 GB GDDR6 of memory using CUDA 11.6.

4. Results

4.1. Results over icoAIS validation set

In Tables 1 and 2 the quantitative segmentation results obtained on the icoAIS validation set are presented. The first table shows the results considering all lesion sizes and the second one only considering lesions bigger than 3 mL.

Firstly, from the two result summaries, focusing on DSC as usually done in the literature, we can see that pre-training the U-shape model's encoder with SSL gave an improvement in performance compared to training from scratch. Almost all the methods pre-trained with SSL had a higher mean and median Dice than *FS-STKi*. Secondly, when considering the influence of the dataset used in SSL pre-training, using all the available NCCTs (*ALL*) was on par to using only the stroke positive datasets (*STKp*). However, both achieved superior Dice scores than using only stroke-negative (healthy) data (*STKn*). Finally, when evaluating the use of the ad hoc symmetry augmentation technique, it can be seen that it slightly improved the performance when pretraining with all the NCCTs or only the stroke-positive datasets.

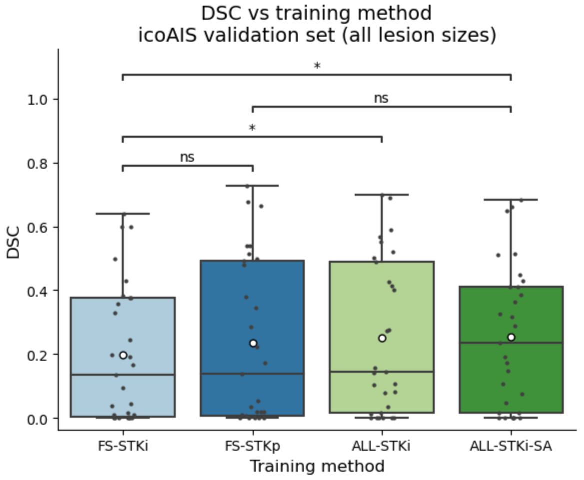


Figure 4: Dice Score Coefficient for the best performing training methods computed over the *icoAIS validation set* considering all lesion sizes. The statistical significance were determined with a paired Wilcoxon rank test, where *ns* indicates $0.05 < p <= 1$ and * indicates $0.01 < p <= 0.05$

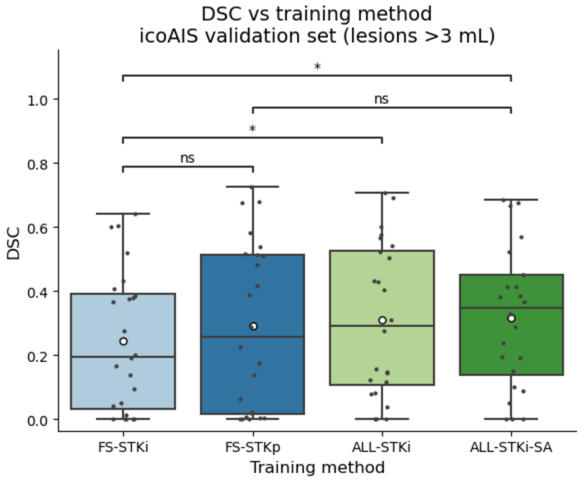


Figure 5: Dice Score Coefficient for the best performing training methods computed over the *AIS validation set* considering lesions $> 3mL$. The statistical significance were determined with a paired Wilcoxon rank test, where *ns* indicates $0.05 < p <= 1$ and * indicates $0.01 < p <= 0.05$

Again focusing on Dice, from all the SSL pre-training variants, the best performance was achieved when using all available NCCTs with the symmetric data augmen-

Table 1: Performance measures on *icoAIS validation set* considering all lesion sizes (cases n=29).

	DSC ↑		HD95 [mm] ↓		AVD [mL] ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	0.1981	0.135 (0.373)	45.19	47.27 (14.48)	**19.74**	9.51 (17.03)	0.52	0.70
FS-STKp	0.2361	0.138 (0.487)	**39.39**	**35.52** (23.84)	21.83	10.23 (27.23)	0.65	**0.74**
ALL-STKi	0.2510	0.144 (0.473)	45.72	50.32 (32.85)	21.56	**8.61** (21.01)	0.65	0.63
ALL-STKi-SA	**0.2554**	**0.237** (0.395)	44.33	47.69 (33.45)	21.84	9.23 (19.76)	**0.70**	0.63
STKp-STKi	0.2503	0.166 (0.431)	49.94	50.47 (29.35)	20.29	9.74 (17.47)	0.60	0.67
STKp-STKi-SA	0.2525	0.186 (0.424)	43.76	48.19 (30.80)	20.27	9.55 (17.70)	0.61	0.65
STKn-STKi	0.2195	0.133 (0.396)	52.43	48.28 (40.01)	24.23	11.22 (20.41)	0.58	0.61
STKn-STKi-SA	0.2096	0.083 (0.336)	49.86	47.93 (38.67)	26.54	12.73 (25.60)	0.64	0.54

Table 2: Performance measures on *icoAIS validation set* considering lesions > 3*mL* (cases n=24).

	DSC ↑		HD95 [mm] ↓		AVD [mL] ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	0.2451	0.1958 (0.358)	44.37	43.60(26.20)	**21.18**	9.91 (19.86)	0.50	0.7
FS-STKp	0.2720	0.2981 (0.405)	44.87	41.75(38.07)	27.99	14.57 (34.48)	0.69	**0.73**
ALL-STKi	0.3101	0.2926 (0.420)	43.12	39.16(32.38)	23.40	**7.00** (24.26)	0.65	0.62
ALL-STKi-SA	**0.3171**	**0.3477** (0.312)	**38.52**	**34.70**(37.78)	24.01	11.50 (20.73)	**0.71**	0.63
STKp-STKi	0.3103	0.3307 (0.358)	51.64	45.91(41.46)	21.07	8.91 (17.02)	0.69	0.66
STKp-STKi-SA	0.3122	0.3589 (0.376)	44.74	45.16(35.71)	21.32	7.61 (20.05)	0.66	0.65
STKn-STKi	0.2725	0.2705 (0.412)	47.76	41.04(44.06)	27.10	13.07 (18.84)	0.62	0.61
STKn-STKi-SA	0.2608	0.1986 (0.369)	45.92	38.59(34.48)	29.95	15.91 (35.32)	0.65	0.52

tation (*ALL-STKi-SA*). See Figure 4 for a comparison of Dice for the best performing methods. Considering all lesion sizes, a mean Dice of 0.2554 ± 0.225 and a median Dice of 0.237 ± 0.395 were obtained, which were significantly higher than those obtained for training from scratch only with icoAIS data (mean DSC of 0.1981 ± 0.214 and median of 0.135 ± 0.373).

More interestingly, the results obtained with SSL pretraining on all the data (*ALL-STKi-SA*) were superior to those obtained when the supervised model was trained from scratch with all labelled datasets (*FS-STKp*) (mean DSC 0.2361 ± 0.255 and median 0.138 ± 0.487).

When considering the results obtained by including only lesions larger than 3 mL, two things must be noted. The results stated in the previous paragraph still hold in this case -as can be seen in Table 2- with *ALL-STKi-SA* SSL pre-trained model outperformed its counterpart trained from scratch (see Figure 5 for a boxplot comparison of the best methods).

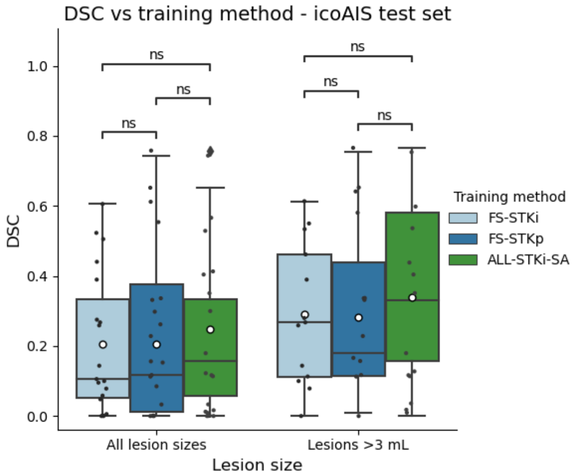


Figure 6: Dice Score Coefficient for the best performing training methods computed over the *icoAIS test set* ablated by lesion size. The statistical significance were determined with a paired Wilcoxon rank test, where *ns* indicates 0.05 < *p* <= 1

For the other quantitative measures presented in Tables 1 and 2, the results were not as clear as in the case of DSC. In the case of the 95% Hausdorff distance, regardless of the lesion size considered, *ALL-STKi-SA* slightly outperformed training from scratch with the same supervised learning dataset. However, this was not the case when using all labelled cases, where the SSL model was superior only when excluding the small volumes. When considering the absolute volume difference, the results are even more dispersed, indicating that the best performing method across all runs was *FS-STKi* if we consider the mean value, or *ALL-STKi* if we consider

Table 3: Performance measures on **icoAIS test set** for both lesion sizes criteria (cases n=19 for *All* and n=14 for $> 3mL$).

	DSC ↑		HD95 [mm] ↓		AVD [mL] ↓		Lesion size
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)	
FS-STKi	0.2053	0.106 (0.279)	46.99	48.75 (19.71)	26.81	8.92 (30.37)	All
FS-STKp	0.2059	0.117 (0.366)	**45.23**	**42.40** (33.66)	23.77	11.28 (28.79)	All
ALL-STKi-SA	**0.2468**	**0.156** (0.275)	45.81	42.56 (27.89)	**21.51**	**6.70** (21.01)	All
FS-STKi	0.2915	0.268 (0.349)	42.91	41.27 (28.61)	30.43	5.47 (50.83)	$> 3mL$
FS-STKp	0.2833	0.180 (0.325)	**36.69**	**37.47** (38.61)	**22.24**	9.53 (31.37)	$> 3mL$
ALL-STKi-SA	**0.3406**	**0.332** (0.424)	42.66	42.56 (41.39)	23.27	**5.32** (29.38)	$> 3mL$

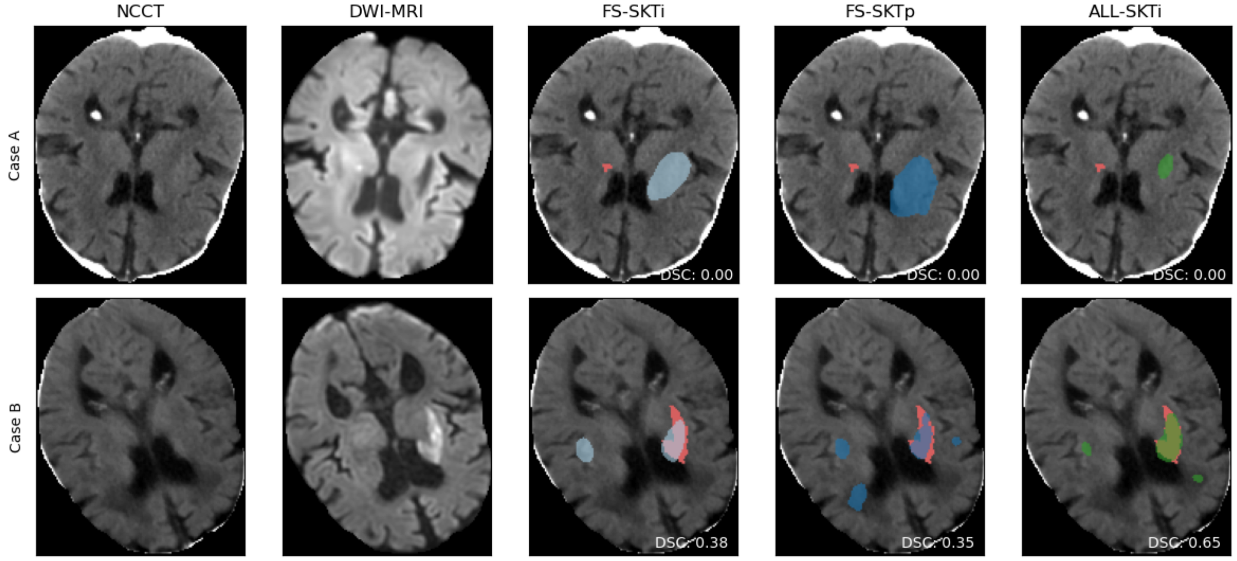


Figure 7: Example results from the **icoAIS test set**. In the upper row a case (A) in which all the methods failed, in the bottom row a case (B) in which all the methods had reasonable performance. In all the cases the GT is shown in red with the model prediction overlaid in colours different from red

the median. Finally, the model selected as best by the Dice criteria showed the highest Spearman volume correlation, and all the models showed a moderate ICC, both when including and excluding lesions smaller than 3mL.

4.2. Results over icoAIS test set

In Table 3 the quantitative segmentation results obtained on the icoAIS test set are presented. The upper part shows the results considering all lesion sizes (n=19) and the lower part only considering lesions bigger than 3 mL (n=14). In both cases, only the best performing SSL pre-trained method according to the validation set was compared against the two baselines.

When considering Dice score as the main metric (results summarised in Figure 6), although the differences were not statistically significant, the same trend as seen in the validation sets could be observed. Pre-training the nnU-Net encoder with all the available NCCTs in the SSL fashion, and then fine-tuning with only the cases from the icoAIS dataset, resulted in a better performance (median DSC for all lesion sizes: 0.156, and for lesions $> 3mL$: 0.332) than training from scratch (median DSC for all lesion sizes: 0.106, and for lesions $> 3mL$: 0.268), and even slightly better than training from scratch with all the labelled data. Although the SSL pre-trained model did not achieve the best HD95, it slightly outperformed its counterpart trained from scratch, and it achieved the best average volume difference of the three models considered. Spearman correlation and ICC values were omitted due to the low confidence in their results given the small size of the test set.

Figure 7 shows qualitative results for two cases from the test partition of the icoAIS dataset. The top row persents a challenging case that none of the methods could segment correctly, and the bottom row shows a case where all methods performed well. In both cases, it is important to note that the SSL pre-trained model (last column) was more specific than its counterparts trained from scratch.

4.3. Interpreting the SSL pre-trained encoder

Figure 8, shows the results of applying some techniques to interpret the latent representations obtained with the best SSL pre-trained encoder. The top sub-figure depicts the low-dimensional projections of the NCCT volumes from the test set. As can be seen, a clustered structure emerged from the data points that was not related to the origin of the data or the presence of stroke lesions (colour of the dots).

Visual inspection was done on examples of each cluster, without identifying clear patterns, biases or shortcuts used by the model to aggregate the cases. In this sense, the middle and the bottom sub-figures show the middle slice of the 6 nearest neighbouring volumes from some representative points in the scatter plot (black markers). In the middle one, it can be seen that in general there are some common elements along the rows (neighbours). For example, in the first and second rows there is a similarity in the orientation of the volumes, and in the third row the cases seem to have large or abnormal ventricle sizes. The last sub-picture shows the attribution maps obtained from these cases. In the cases from the first row, the model was strongly influenced by some voxels in the frontal region, in the second row, some attention was given to the anterior voxels outside the brain (possibly related to the overall orientation of the case), and in the third row, the ventricular regions are highlighted. Overall, it is important to note that even when attention was paid to voxels outside the brain, the representations were able to capture information that was mostly related to the brain region.

5. Discussion

In this work, the use of a DeSD-like self-supervised pre-training strategy was proposed to exploit large unlabelled NCCT (stroke positive and negative) datasets in the task of AIS final infarct segmentation.

As a first remark, it is worth highlighting the software engineering contributions made in order to enhance the nnU-Net model with SSL pre-training of its encoder. In this work, an SSL training infrastructure was designed in such a way that it was automatically adapted to the guidelines resulting from the robust nnU-Net self-configuration pipeline. In this sense, the proposed method, represents a contribution beyond the problem or datasets addressed in this work, allowing the use of SSL pre-training in any other 3D medical image segmentation problem.

In the experiments conducted for AIS, the segmentation performance obtained by pre-training the nnU-Net encoder in a self-supervised learning fashion and then doing supervised segmentation training was significantly better than training nnU-Net from scratch on the same supervised learning dataset. These results were consistent with the findings of Ye et al. (2022), despite the different nature of the problem and the different network architecture used.

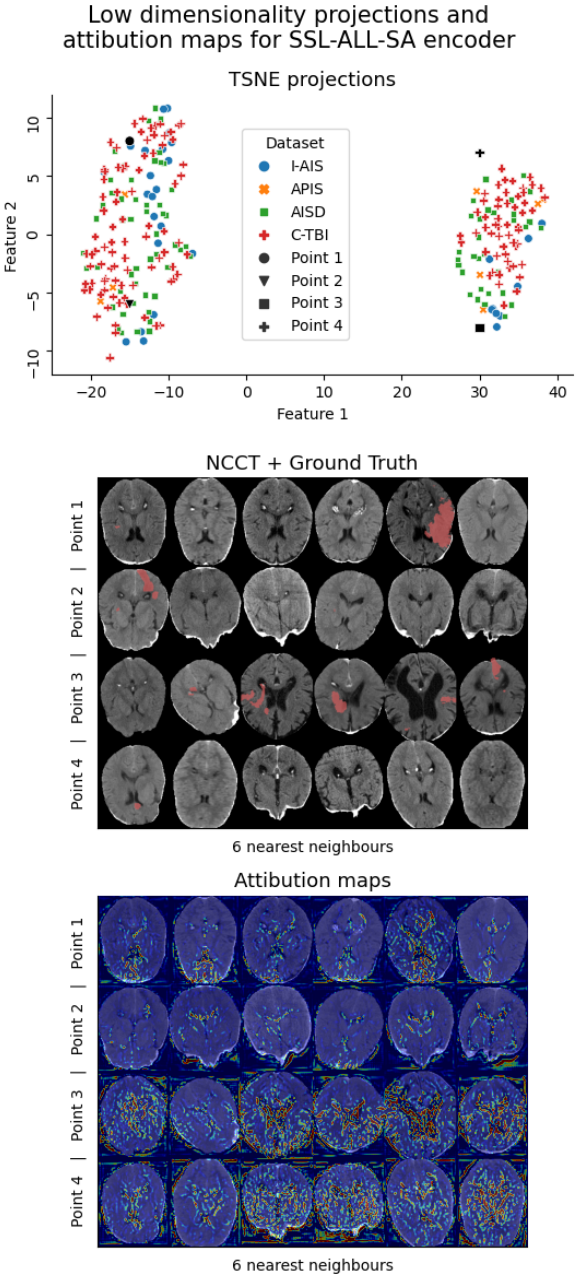


Figure 8: T-SNE low dimensional projections of the image representations and NCCT with the respective Attribution map for representative points.

In terms of the dataset used to pre-train the encoder, it was interesting that only a small difference to was found between using all the stroke positive and negative cases and using only the AIS positive datasets, while both had a large performance gain with compared to using only the stroke negative cases for pre-training. There might be several explanations for this. One of them is that the inclusion of stroke positive cases may be a key element in achieving a good pre-training for the supervised task at hand. However, the attention maps obtained for the encoder and the clusters shown by the low representations did not suggest that the pre-trained encoder itself captured any evident information about the stroke lesions. A more resonable explanation lies in the inclusion of cases in the training set for SSL that were also present in the supervised training set, allowing the model to take into account the particular image characteristics of this dataset. In any case, it was clear that the model did not benefit from the contrast of pre-training on healthy cases and then using only pathological ones for the supervised task.

In relation to the exploitation of contextual information, it was shown that including the proposed symmetry augmentation technique during pre-training led to a small improvement in the segmentation performance. This was not surprising, as it was consistent with many publications in the field showing that models achieve better AIS segmentation performance when forced to exploit inter-hemispheric symmetry/asymmetry. However, despite the initial belief that the symmetry augmentation would be more beneficial for the downstream task when pre-training on non-stroke cases (becoming normal asymmetry agnostic), the results obtained were exactly the opposite. As there is no immediate rational explanation for this phenomenon, further experiments should be conducted to better understand what is the effect of this augmentation on the obtained representations.

Also in relation to the contextual information, in the introduction it was initially hypothesised that the use of SSL could lead to meaningful representations of the NCCT images that could exploit the contextual information already present in the image itself. However, although the encoder's attribution maps showed that the pre-trained encoder was mostly influenced by brain voxels and the shape/location of certain structures such as the ventricles, it is difficult to determine the impact of this information on the final segmentation. The inclusion of additional pretext tasks during the SSL training, as done in Giancardo et al. (2023), is a strategy that could lead to representations better suited to the downstream problem and should be explored in future work.

One of the other important results of this work was that the best SSL pre-training strategy was able to achieve performances at least on par with the ones obtained by training the supervised segmentation model trained from scratch with almost 3.6 times more labelled cases. As commented previously, there was a high variability in the quality of the datasets, in their initial preprocessing, in the volume of the lesions in them, and only icoAIS had certain infarct ground truth masks. In this context, SSL pre-training provided a robust way of extracting meaningful information from these cases, becoming independent of their variable labelling standards.

5.1. Comparison with other approaches

From the results section, it is clear that models developed for other medical image segmentation problems achieve better Dice scores. However, AIS segmentation on NCCT is a particularly challenging task, due to the cross-domain nature of the labels, the lack of visibility of the lesions in NCCT, and many other reasons previously exposed.

Especially in the context of AIS, it is difficult to even compare with other approaches presented in the literature. In this work, the icoAIS dataset was chosen as the main dataset; this choice had a major drawback, which is the impossibility -in the time of the project- to compare our results with those of other methods, since neither their approaches nor our dataset are publicly available. However, this dataset was chosen because it was better than the publicly available ones in terms of size, type and quality of the labels. Therefore, in this work, the performance evaluation of the proposed methods was done by comparing the relative improvements of the proposed methods with a widely accepted baseline such as nnU-Net.

Having said this, and taking into account that the strict comparison with other reported methods may be misleading due to several differences (number of cases, label origin, minimum lesion size, lesion location, etc.), our results are in the same range as the those reported by other methods on datasets similar to ours. For example, in Kuang et al. (2021) a 3D UNet applied to AIS segmentation is reported to achieve a mean Dice of 0.308 (sd: 0.283), in Giancardo et al. (2023) their model achieved a mean Dice of 0.26 and a plain nnU-Net one of 0.14 and in El-Hariri et al. (2022) a 3D nnU-Net trained for AIS had a mean Dice of 0.377 or 0.346 (sd: 0.276 and 0.275 respectively) depending on the reader used as ground truth.

5.2. Limitations

Choosing which metric to report and focus on is not an easy task in AIS segmentation. Although the performance based on 95% Hausdorff distance and absolute volume difference was reported and commented in the results section, the focus in this work was placed on the Dice Score coefficient.

The 95% Hausdorff distance is not considered a clinically relevant metric in the field of AIS segmentation and was only included for consistency with other image analysis works. Due to the lack of contrast of AIS in NCCT images, it is hard to achieve accurate contour matching, making it very difficult to get a clear picture of the overall model performance using HD95. AVD, on the other hand, is a very relevant measure from a clinical perspective, nevertheless the values obtained for it should be put into context. With Dice values not surpassing 0.35, it is difficult to tell if a model is better than another simply because it achieved a smaller volume difference. A smaller AVD indicates that the volumes are similar, but in our problem they are most likely mislocalised, so this should be interpreted roughly linked to how specific the models are.

Finally, DSC is not without its complications. The Dice score is biased by the size of the lesion volume, i.e. low spatial overlap for a big lesion might generate a high Dice and vice versa. As a consequence, the increase in DSC reported in the results for all models when the lesions smaller than 3mL were removed from the ground truth, could have two explanations. On the one hand, it could indicate that all models struggled to segment very small lesions, but on the other hand, it could be a consequence of the limitations of the metric itself. Due to the wide range of lesion volumes, depending on the case, small lesions may have a large influence on the metric, which is inconsistent with the lower clinical significance assigned to them.

Although Dice was chosen as the metric to focus on, all of the above concerns should be taken into account when interpreting the results obtained. For future work, a ranking system, such as the one implemented in (Maier et al. (2017)), could be used to collect and ponder the information provided by the different metrics.

Regarding the experimental design, some pitfalls in the dataset partitioning strategy need to be pointed out. Firstly, the distribution of cases between the training, validation and test sets could have been done better. A minimum number of cases should have been guaranteed to be left in the test set to allow for more powerful statistical analysis of the results.

Secondly, as previously commented, in this work a stratified partitioning was done according to data origin, lesion location and lesion size. However, it might have been advantageous not to restrict the location to the sides of the brain, but to specify the nervous system structures affected, as lesions in the cerebellum and brain stem may be more difficult to segment due to bone-related imaging artefacts.

Lastly, more robust results could have been obtained by validating the models using a k-fold cross-validation procedure. However, in a context of limited computational resources, it was preferred to run each supervised nnU-Net experiment multiple times and enssembling the resulting models to reduce the impact of nnU-Net random initialisation on segmentation performance.

6. Conclusions

In this work, a DeSD-like self-supervised pre-training strategy was proposed to exploit large unlabelled NCCT (stroke positive and negative) datasets in the task of AIS final infarct segmentation. A robust data pre-processing pipeline was proposed to homogenise the different datasets before using them in an SSL-enhanced version of the well-known self-configuring nnU-Net model.

From the conducted experiments, pre-training the nnU-Net's encoder in a self-supervised manner with all the available NCCT images (stroke-positive and stroke-negative) resulted in an AIS segmentation performance significantly higher than training the same model from scratch and comparable to that obtained by using approximately 3.6 times more labelled data.

In acute ischemic stroke, is very difficult to have access access to high quality datasets with both baseline and follow up images to accurately asses the extension of the final infarct (or ischemic lesion core). In this work, we presented a successful method to exploit large amounts of unlabelled baseline NCCT images, which are much easier to obtain from hospitals and are currently neglected, proving they can be used to improve final infarct lesion segmentation.

Acknowledgements

I would like to express my deepest gratitude to my supervisors for their trust, feedback and for the freedom they gave me throughout this project. I am also grateful to my family and friends who, despite the physical distances, have always been my emotional support. Lastly, I would like to mention Enzo Ferrante, a crucial link in a long chain of events that made this thesis possible.

Finally, this endeavour would not have been possible without the financial support of the Erasmus Mundus programme of the European Union and the computing resources provided by *icometrix*, Leuven, Belgium.

References

Albers, G.W., Marks, M.P., Kemp, S., Christensen, S., Tsai, J.P., Ortega-Gutierrez, S., McTaggart, R.A., Torbey, M.T., Kim-Tenser, M., Leslie-Mazwi, T., et al., 2018. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. New England Journal of Medicine 378, 708–718. doi:10.1056/nejmoa1713973.

Amador, K., Wilms, M., Winder, A., Fiehler, J., Forkert, N., 2021. Stroke lesion outcome prediction based on 4d CT perfusion data using temporal convolutional networks, in: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning, PMLR. pp. 22–33.

Amador, K., Winder, A., Fiehler, J., Wilms, M., Forkert, N.D., 2022. Hybrid spatio-temporal transformer network for predicting ischemic stroke lesion outcomes from 4d CT perfusion imaging. Lecture Notes in Computer Science , 644–654doi:10.1007/978-3-031-16437-8_62.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851. doi:10.1016/j.neuroimage.2005.02.018.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., Lecun, Y., Goldblum, M., 2023. A cookbook of self-supervised learning .

Bardes, A., Ponce, J., LeCun, Y., 2022. VICReg: Variance-Invariance-Covariance Regularization for self-supervised learning, in: International Conference on Learning Representations. URL: .

Benjamin, E.J., Blaha, M.J., Chiuve, S.E., Cushman, M., Das, S.R., Deo, R., de Ferranti, S.D., Floyd, J., Fornage, M., Gillespie, C., et al., 2017. Heart disease and stroke statistics—2017 update: A report from the American Heart Association. Circulation 135. doi:10.1161/cir.0000000000000485.

Bouslama, M., Ravindran, K., Harston, G., Rodrigues, G.M., Pisani, L., Haussen, D.C., Frankel, M.R., Nogueira, R.G., 2021. Noncontrast computed tomography e-stroke infarct volume is similar to rapid computed tomography perfusion in estimating postreperfusion infarct volumes. Stroke 52, 634–641. doi:10.1161/strokeaha.120.031651.

Brorson, J.R., Cifu, A.S., 2019. Management of Patients With Acute Ischemic Stroke. JAMA 322, 777–778. doi:10.1001/jama.2019.10436.

Byrne, D., Walsh, J.P., MacMahon, P.J., 2019. An acute stroke CT imaging algorithm incorporating automated perfusion analysis. Emergency Radiology 26, 319–329. doi:10.1007/s10140-019-01675-2.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised Vision Transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) doi:10.1109/iccv48922.2021.00951.

Chen, W., Wu, J., Wei, R., Wu, S., Xia, C., Wang, D., Liu, D., Zheng, L., Zou, T., Li, R., et al., 2022. Improving the diagnosis of acute ischemic stroke on non-contrast CT using Deep Learning: A multicenter study. Insights into Imaging 13. doi:10.1186/s13244-022-01331-3.

Chen, X., He, K., 2021. Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15750–15758.

Clèrigues, A., Valverde, S., Bernal, J., Freixenet, J., Oliver, A., Lladó, X., 2020. Acute and sub-acute stroke lesion segmentation from multimodal MRI. Computer Methods and Programs in Biomedicine 194, 105521. doi:10.1016/j.cmpb.2020.105521.

El-Hariri, H., Souto Maior Neto, L.A., Cimflova, P., Bala, F., Golan, R., Sojoudi, A., Duszynski, C., Elebute, I., Mousavi, S.H., Qiu, W., et al., 2022. Evaluating nnU-net for early ischemic change segmentation on non-contrast computed tomography in patients with acute ischemic stroke. Computers in Biology and Medicine 141, 105033. doi:10.1016/j.compbiomed.2021.105033.

Estrada, U.M., Meeks, G., Salazar-Marioni, S., Scalzo, F., Farooqui, M., Vivanco-Suarez, J., Gutierrez, S.O., Sheth, S.A., Giancardo, L., 2022. Quantification of infarct core signal using CT imaging in acute ischemic stroke. NeuroImage: Clinical 34, 102998. doi:10.1016/j.nicl.2022.102998.

Farzin, B., Fahed, R., Guilbert, F., Poppe, A.Y., Daneault, N., Durocher, A.P., Lanthier, S., Boudjani, H., Khoury, N.N., Roy, D., et al., 2016. Early CT changes in patients admitted for thrombectomy. Neurology 87, 249–256. doi:10.1212/wnl.0000000000002860.

Garrido, Q., Balestriero, R., Najman, L., Lecun, Y., 2023. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank doi:10.48550/arXiv.2210.02885.

Giancardo, L., Niktabe, A., Ocasio, L., Abdelkhaleq, R., Salazar-Marioni, S., Sheth, S.A., 2023. Segmentation of acute stroke infarct core using image-level labels on CT-Angiography. NeuroImage: Clinical 37, 103362. doi:10.1016/j.nicl.2023.103362.

Goyal, M., Ospel, J.M., Menon, B., Almekhlafi, M., Jayaraman, M., Fiehler, J., Psychogios, M., Chapot, R., van der Lugt, A., Liu, J., et al., 2020. Challenging the ischemic core concept in acute ischemic stroke imaging. Stroke 51, 3147–3155. doi:10.1161/strokeaha.120.030620.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., 2020. Bootstrap your own latent a new approach to self-supervised learning, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Gómez, S., Florez, S., Mantilla, D., Camacho, P., Tarazona, N., Martínez, F., 2023. An attentional U-Net with an auxiliary class learning to support acute ischemic stroke segmentation on CT. Medical Imaging 2023: Image Processing doi:10.1117/12.2654269.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al., 2022. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific Data 9. doi:10.1038/s41597-022-01875-5.

Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M., 2022. Synthstrip: Skull-stripping for any brain image. NeuroImage 260, 119474. doi:10.1016/j.neuroimage.2022.119474.

Iglesias, J.E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S.E., Das, S., Edlow, B.L., Alexander, D.C., Golland, P., Fischl, B., 2023. SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. Science Advances 9. doi:10.1126/sciadv.add3607.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2020. nnU-net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18, 203–211. doi:10.1038/s41592-020-01008-z.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H., Heiland, S., Wick, W., et al., 2019. Automated brain extraction of multisequence MRI using artificial neural networks. Human Brain Mapping 40, 4952–4964. doi:10.1002/hbm.24750.

Jiang, Y., Sun, M., Guo, H., Yan, K., Lu, L., Xu, M., 2023. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3D medical image segmentation. arXiv preprint arXiv:2302.05615 .

Kalapos, A., Gyires-Tóth, B., 2023. Self-supervised pretraining for 2D medical image segmentation. Lecture Notes in Computer Science , 472–484doi:10.1007/978-3-031-25082-8_31.

Kim, Y., Lee, S., Abdelkhaleq, R., Lopez-Rivera, V., Navi, B., Kamel, H., Savitz, S.I., Czap, A.L., Grotta, J.C., McCullough, L.D., et al., 2021. Utilization and availability of advanced imaging in patients with acute ischemic stroke. Circulation: Cardiovascular Quality and Outcomes 14. doi:10.1161/circoutcomes.120.006989.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15, 155–163. doi:10.1016/j.jcm.2016.02.012.

Kuang, H., Menon, B.K., Qiu, W., 2019. Automated infarct segmentation from follow-up non-contrast CT scans in patients with acute ischemic stroke using dense multi-path contextual generative adversarial network. Lecture Notes in Computer Science , 856–863doi:10.1007/978-3-030-32248-9_95.

Kuang, H., Menon, B.K., Sohn, S.I., Qiu, W., 2021. EIS-net: Segmenting early infarct and scoring aspects simultaneously on non-contrast CT of patients with acute ischemic stroke. Medical Image Analysis 70, 101984. doi:10.1016/j.media.2021.101984.

Li, S., Zheng, J., Li, D., 2021. Precise segmentation of non-enhanced computed tomography in patients with ischemic stroke based on multi-scale U-net Deep Network model. Computer Methods and Programs in Biomedicine 208, 106278. doi:10.1016/j.cmpb.2021.106278.

Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C., 2016. The first step for neuroimaging data analysis: DICOM to NIFTI conversion. Journal of Neuroscience Methods 264, 47–56. doi:10.1016/j.jneumeth.2016.03.001.

Lutkenhoff, E.S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J.D., Owen, A.M., Monti, M.M., 2014. Optimized brain extraction for pathological brains (optibet). PLoS ONE 9. doi:10.1371/journal.pone.0115551.

Maas, A.I., Menon, D.K., Steyerberg, E.W., Citerio, G., Lecky, F., Manley, G.T., Hill, S., Legrand, V., Sorgner, A., 2015. Collaborative european neurotrauma effectiveness research in traumatic brain injury (center-TBI). Neurosurgery 76, 67–80. doi:10.1227/neu.0000000000000575.

van der Maaten, L., Hinton, G., 2008. Viualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. ISLES 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269. doi:10.1016/j.media.2016.07.009.

Manna, S., Chakraborty, S., 2022. BYOLMed3D: Self-supervised representation learning of medical videos using gradient accumulation assisted 3D BYOL framework doi:10.48550/arXiv.2208.00444.

Mokin, M., Primiani, C.T., Siddiqui, A.H., Turk, A.S., 2017. ASPECTS (Alberta stroke program early CT score) measurement using Hounsfield unit values when selecting patients for stroke thrombectomy. Stroke 48, 1574–1579. doi:10.1161/strokeaha.117.016745.

Ni, H., Xue, Y., Wong, K., Volpi, J., Wong, S.T., Wang, J.Z., Huang, X., 2022. Asymmetry disentanglement network for interpretable acute ischemic stroke infarct segmentation in non-contrast CT scans. Lecture Notes in Computer Science , 416–426doi:10.1007/978-3-031-16452-1_40.

Nogueira, R.G., Jadhav, A.P., Haussen, D.C., Bonafe, A., Budzik, R.F., Bhuva, P., Yavagal, D.R., Ribo, M., Cognard, C., Hanel, R.A., et al., 2018. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. New England Journal of Medicine 378, 11–21. doi:10.1056/nejmoa1706442.

Ostmeier, S., Heit, J., Axelrod, B., Li, L.J., Zaharchuk, G., Verhaaren, B., Mahammedi, A., Christensen, S., Lansberg, M., 2022. Non-inferiority of deep learning model to segment acute stroke on non-contrast CT compared to neuroradiologists doi:10.48550/arXiv.2211.15341.

Powers, W.J., Rabinstein, A.A., Ackerson, T., Adeoye, O.M., Bambakidis, N.C., Becker, K., Biller, J., Brown, M., Demaerschalk, B.M., Hoh, B., et al., 2019. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke. Stroke 50. doi:10.1161/str.0000000000000211.

Robben, D., Boers, A.M., Marquering, H.A., Langezaal, L.L., Roos, Y.B., van Oostenbrugge, R.J., van Zwam, W.H., Dippel, D.W., Majoie, C.B., van der Lugt, A., et al., 2020. Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. Medical Image Analysis 59, 101589. doi:10.1016/j.media.2019.101589.

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., Karnath, H.O., 2012. Age-specific CT and MRI templates for spatial normalization. NeuroImage 61, 957–965. doi:10.1016/j.neuroimage.2012.03.020.

Sacco, R.L., Kasner, S.E., Broderick, J.P., Caplan, L.R., Connors, J.B., Culebras, A., Elkind, M.S., George, M.G., Hamdan, A.D., Higashida, R.T., et al., 2013. An updated definition of stroke for the 21st Century. Stroke 44, 2064–2089. doi:10.1161/str.0b013e318296aeca.

Shamonin, D., 2013. Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer's disease. Frontiers in Neuroinformatics 7. doi:10.3389/fninf.2013.00050.

Sotoudeh, H., Bag, A.K., Brooks, M.D., 2019. "code-stroke" CT perfusion; challenges and pitfalls. Academic Radiology 26, 1565–1579. doi:10.1016/j.acra.2018.12.013.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for Deep Networks, in: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR. pp. 3319–3328.

Vagal, A., Wintermark, M., Nael, K., Bivard, A., Parsons, M., Grossman, A.W., Khatri, P., 2019. Automated CT perfusion imaging for acute ischemic stroke. Neurology 93, 888–898. doi:10.1212/wnl.0000000000008481.

Virani, S.S., Alonso, A., Aparicio, H.J., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Cheng, S., Delling, F.N., et al., 2021. Heart disease and stroke statistics—2021 update. Circulation 143. doi:10.1161/cir.0000000000000950.

Ye, Y., Zhang, J., Chen, Z., Xia, Y., 2022. DeSD: Self-supervised learning with deep self-distillation for 3D medical image segmentation. Lecture Notes in Computer Science , 545–555doi:10.1007/978-3-031-16440-8$_5$2.

Appendix A. ISBI 2023 APIS Challenge solution

As mentioned in Section 3.7.4, our submission to the ISBI 2023 APIS Challenge won the competition. The proposed model achieved a Dice score of 0.20 ± 0.25 and a Hausdorff distance of 66.02 ± 24.22 on the hidden test set, which were significantly better than those of the second best solution (0.11 ± 0.30 and 59.64 ± 22.88, respectively). Since both solutions used a 3D nnU-Net model, the difference in performance was mainly explained by the use of the robust preprocessing strategy presented in Section 3.2, which allowed us to use a larger set of training cases, and by the inclusion of the difference image as an additional input channel.

To obtain this model, the inter-hemispheric difference image was first generated by subtracting the NCCT from its contralateral version, resulting in an image in which both normal and abnormal inter-hemispheric differences appeared highlighted. Then, a training procedure was performed in the same way as that described in Section 3.7.1 for the baseline *FS-STKp*.

Appendix B. Addition of interhemispherical difference image as another input channel

In line with Appendix A, for this experiment the interhemispheric difference image was generated and then used as an additional input channel in both the SSL pre-training and the supervised experiments.

The two baselines presented in Section 3.7.1 were run for the multi-channel input. Additionally, the nnU-Net encoder was SSL pre-trained with the three dataset configurations presented in Section 3.7.2. Once the encoder was pre-trained, the full supervised nnU-Net was trained using only the icoAIS dataset. All training procedures followed exactly the same specifications as described in section 3.

Tables A.4 and A.5 show the quantitative results obtained for the segmentation task on the icoAIS validation set. Contrary to the results presented in Section 4, it can be seen here that pre-training the nnU-Net encoder with SSL and fine-tuning the full architecture with supervised training led to worse performance than training nnU-Net from scratch on the same dataset.

Comparing the results between the supervised models trained from scratch with and without the inclusion of the difference image, it can be seen that, in line with the

Table B.4: Performance measures for multichannel input with difference image + NCCT, on (icoAIS val. set) considering all lesions sizes (n=29).

	DSC ↑		HD95 ↓		AVD ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	**0.2445**	0.130 (0.470)	44.39	41.15 (32.84)	**22.00**	15.77 (23.23)	**0.7**	**0.69**
FS-STKp	0.2380	**0.185** (0.420)	**38.52**	**34.05** (19.33)	24.48	**14.04** (23.74)	0.69	0.60
ALL-STKi	0.1897	0.122 (0.267)	57.20	55.66 (35.55)	51.22	43.82 (33.68)	0.46	0.26
STKp-STKi	0.1989	0.139 (0.303)	53.51	54.48 (21.97)	47.01	34.42 (33.55)	0.34	0.30
STKn-STKi	0.1953	0.119 (0.332)	55.91	54.08 (26.03)	58.09	48.42 (50.82)	0.42	0.25

Table B.5: Performance measures for multichannel input with difference image + NCCT, on *icoAIS val. set* considering lesions > 3*mL* (n=24).

	DSC ↑		HD95 ↓		AVD ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	0.2890	**0.314** (0.435)	41.78	**36.47** (34.01)	**25.04**	**12.47** (25.68)	0.57	**0.64**
FS-STKp	**0.2920**	0.293 (0.495)	**41.63**	36.60 (38.42)	25.21	17.76 (27.50)	**0.74**	0.58
ALL-STKi	0.2253	0.144 (0.238)	64.73	66.40 (37.36)	53.25	33.67 (35.06)	0.28	0.17
STKp-STKi	0.2361	0.162 (0.285)	59.81	62.44 (27.27)	48.25	33.36 (39.39)	0.23	0.21
STKn-STKi	0.2317	0.149 (0.295)	62.27	62.78 (32.34)	60.70	42.83 (58.37)	0.26	0.16

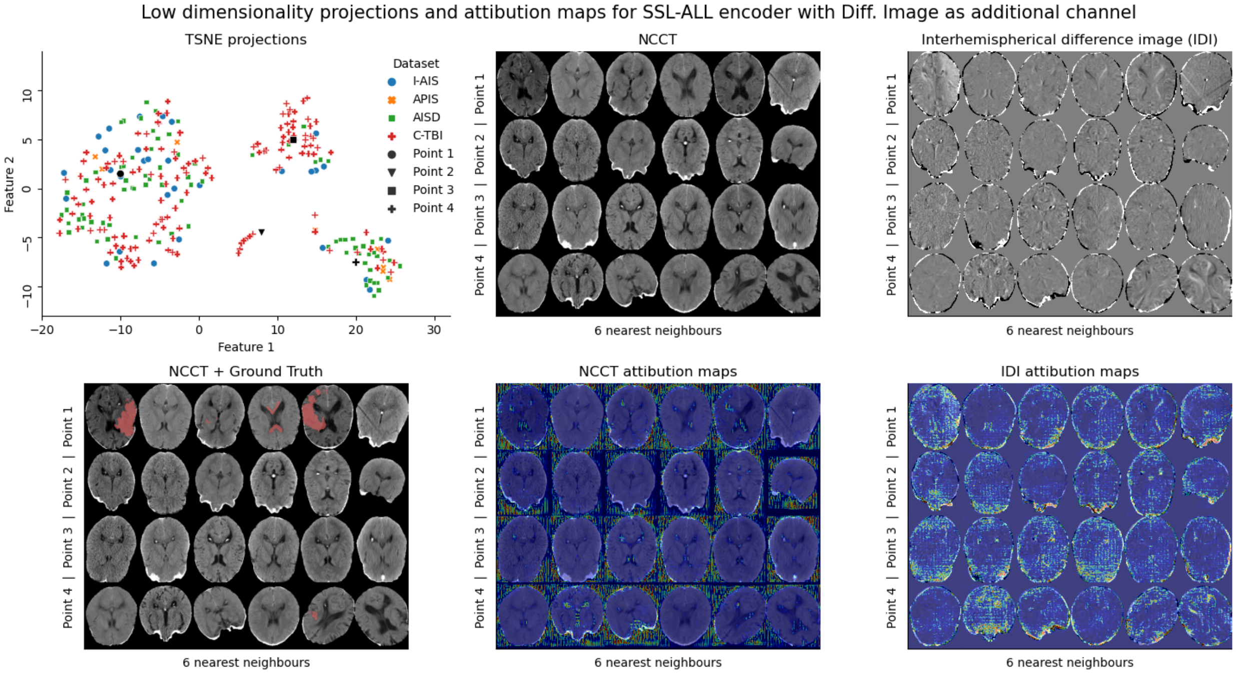


Figure B.9: T-SNE low dimensional projections of the image representations and NCCT and interhemispherical difference images examples (with their the respective attribution maps) for representative points.

literature, the inclusion of this additional input channel was beneficial.

As can be seen in Figure B.9, the attention maps of the encoders pre-trained with SSL show that the output of the model was strongly influenced by the brain region in the difference image channel and, unexpectedly, by the non-brain region of the NCCT channel. This finding, in addition to the benefits seen in training from scratch, suggests that further experiments should be conducted to explore different ways of exploiting the potential of the two images during self-supervised pre-training (i.e. as data augmentation techniques, separate paths for each image, etc.).