Multi-input segmentation of damaged brain in acute ischemic stroke patients using slow fusion with skip connection

Luca Tomasetti*[1], Mahdieh Khanmohammadi[1], Kjersti Engan[1], Liv Jorunn Høllesli[1,2], and Kathinka Dæhli Kurz[1,2]

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway
[2]Stavanger Medical Imaging Laboratory (SMIL), Department of Radiology, Stavanger University Hospital, 4019 Stavanger, Norway

Abstract

Time is a fundamental factor during stroke treatments. A fast, automatic approach that segments the ischemic regions helps treatment decisions. In clinical use today, a set of color-coded parametric maps generated from computed tomography perfusion (CTP) images are investigated manually to decide a treatment plan. We propose an automatic method based on a neural network using a set of parametric maps to segment the two ischemic regions (core and penumbra) in patients affected by acute ischemic stroke. Our model is based on a convolution-deconvolution bottleneck structure with multi-input and slow fusion. A loss function based on the focal Tversky index addresses the data imbalance issue. The proposed architecture demonstrates effective performance and results comparable to the ground truth annotated by neuroradiologists. A Dice coefficient of 0.81 for penumbra and 0.52 for core over the large vessel occlusion test set is achieved. The full implementation is available at: `https://git.io/JtFGb`.

1 Introduction

A cerebral stroke is the second most common cause of death among adults worldwide [21]. Cerebral stroke can be divided into two general categories: ischemic and hemorrhagic stroke. Ischemic stroke approximately represents 80% of the totality of the strokes [16]. The ischemic brain tissue is divided into two distinct regions during an ischemic stroke: the ischemic core (infarcted tissue) and the ischemic penumbra, a hypoperfused but viable tissue region. Fast and correct visualization of the salvageable penumbra and the irreversibly damaged core tissues can benefit medical doctors for treatment planning in acute stroke patients (AIS).

Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) are the two of the modalities used to diagnose acute stroke patients [6]. CT is preferred in many centers due to its high sensitivity for detecting hemorrhage, rapid scan times, and widespread availability. Information about clinical severity are calculated, at hospital admission, using the National Institutes of Health Stroke Scale (NIHSS), and color-coded parametric maps (2D PMs) are generated using the 4D CT Perfusion (CTP) imaging usually performed immediately after hospital admission. PMs are estimated to evaluate the changes in the tissue density over the injection of a contrast agent over time. Time-to-peak (TTP), time-to-maximum (T_{Max}), cerebral blood flow (CBF), and cerebral blood volume (CBV), are all examples of PMs, derived from pixel information of a time density curve, generated from a CTP study [11]. Also, the maximum intensity projection (MIP) is found as the maximum Hounsfield unit value over the time sequence of the CTP providing a 3D volume. In addition to diagnosing acute stroke, CT is also necessary for treatment decisions, with CTP being an essential modality with the ability to assess the penumbra and core.

Deep neural network (DNN) models have been proven to be an effective and beneficial tool for classification and segmentation tasks in many medical image analysis applications. Various research

*Corresponding Author: luca.tomasetti@uis.no

https://doi.org/10.7557/18.6223

groups have focused their effort on the study of ischemic strokes, and some methods have been developed for classifying and segmenting the infarct core [2, 5, 9, 13]. These methods rely on a set of PMs as input and ground truth images generated through follow-up images acquired hours after the stroke onset. Nevertheless, the methods mentioned above only segment *core* regions. However, it could be more beneficial to acquire knowledge also about the penumbra regions since it is crucial for the treatment decisions during the first stages of the ischemic stroke [15, 20]. To the best of our knowledge, Tomasetti et al. was the first research group to segment both *core and penumbra* using machine learning and deep learning approaches [19, 20].

It is highly time-consuming to collect and label medical data, and transfer learning is a popular approach to solve problems related to medical images [3, 22]. Over the past years, there have been numerous examples of transfer learning architectures used for various tasks in disparate domains pretrained with the ImageNet dataset [4, 22]. Additionally, early and slow fusion approaches, with or without inflation, have been proven to improve accuracy in video classification [4, 8], and medical diagnosis [12, 14]. An early fusion approach combines input information at the beginning of the process, allowing a network to increase the performances of the system using cross-correlation between data [7]. The slow fusion approach slowly merges input information throughout the model permitting higher layers to access more global information [8]. An inflating technique has been proposed to use pretrained weights from image classification networks in video classification models, expanding the filters from 2D (image-based) to 3D (video-based) [4].

A proper understanding of *both* ischemic regions is a major requisite for initial treatment decisions; however, it has not been fully explored in the previous researches; thus, in this work, we propose a DNN architecture to simultaneously segment both core and penumbra regions in AIS patients. We implemented a structure that was inspired by a multiscale model proposed by Wetteland et al. [22]; however, while Wetteland's model used the same image but with different magnifications as input, we propose a different approach. Our model uses a multi-input CNN with slow fusion, based on transfer learning from VGG-16 models pre-trained on ImageNet. We want to investigate if the usage of the PMs in combination with MIP volume and NIHSS as input can produce meaningful results in the segmentation of ischemic regions, as this is already calculated and in use in clinical settings. The paper contributes with:

- A fully-automatic DNN method (Fig. 1) to segment both *core and penumbra*, using color-coded PMs acquired shortly after hospital admission when an AIS is expected.
- A slow fusion multi-input approach is tested combining the PMs with other images and/or patient information.
- The model is trained and tested using a dataset of patients affected by different levels of vessel occlusion for generalizing the input.
- Manual annotations based on experts' assessment of the CTP with PMs and MIP are used as ground truth.

2 Dataset

The study was approved by the Regional ethic committee project 2012/1499. The dataset included CTP scans from 152 patients collected at Stavanger University Hospital (SUS) between January 2014 and August 2020. NIHSS score was available for all patients. Based on the level of vessel occlusion using CT angiography, large vessel occlusion (LVO) was defined as occlusion of a large, proximal artery. Non-LVO was defined as patients with perfusion deficits with occlusion of a smaller, more distal artery or with perfusion deficits without visible artery occlusion. Patients were divided into three groups based on the vessel occlusion severity: 77 patients with LVO, 60 Non-LVO patients, and 15 without ischemic stroke (WIS).

Two neuroradiologists with 16 and 3.5 years of clinical experience delineated ground truth core and penumbra regions in an in-house developed software tool. These delineations were based on visual information in the CBF, CBV, TTP, T_{Max} and MIP images acquired directly after the CTP at hospital admission. PMs and MIP have a 512×512 pixel resolution, displaying only the brain region. Each patient study contains a number of brain slices between 13 and 27 for each scan. Furthermore, the MRI examination performed within one to three days after the CT examination was studied and used in assistance to generate the ground

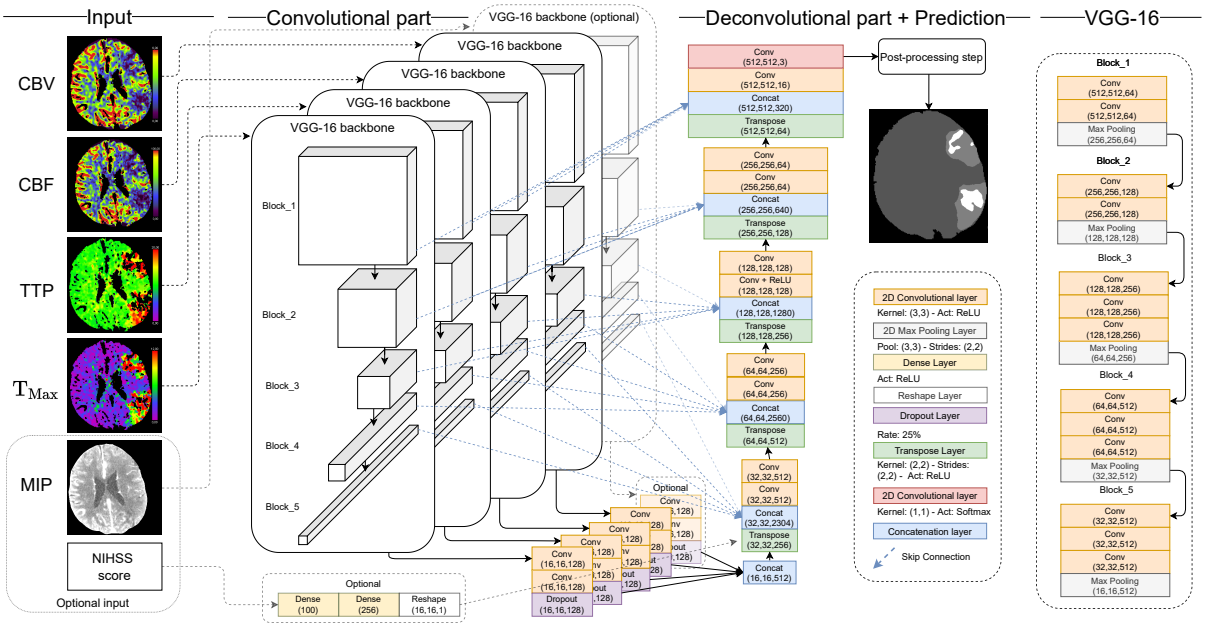

Figure 1: Proposed architecture with multi-input and slow fusion. The network has a convolutional part composed of four to five VGG-16 models. Skip connections combined with feature concatenation from VGG-16 models ensure slow data fusion. On the left, an overview of the feature extraction part of the VGG-16 architecture.

truth images since follow-up MRI exhibits the final infarct core. The dataset was randomly split into training (58%), validation (20%), and testing sets (22%). The training set resulted in 89 patients (42 LVO, 38 Non-LVO, 9 WIS); the validation set resulted in 30 patients (16 LVO, 11 Non-LVO, 3 WIS), and the remaining 33 patients for testing (19 LVO, 11 Non-LVO, 3 WIS).

3 Method

The baseline model takes as input four PMs (CBF, CBV, TTP, T_{Max}) of each 2D brain slice, MIP images and, for some experiments, we also use the patient's NIHSS score as input. The slow fusion is done both by concatenating the feature vectors from the different inputs and introducing skip connections from the different levels of the convolutional part.

The architecture presents a structure that resembles the U-Net model [17]. Fig. 1 displays a representation of the architecture. The convolutional part of the network is based on four to five distinct VGG-16 networks. This part is used to extract low-resolution features. A detailed overview of a single VGG-16 architecture is displayed on the right of Fig. 1. At the end of each VGG-16 network, two convolutional layers followed by a dropout layer are inserted to regularize the model. The VGG-16 models were pre-trained with the ImageNet dataset. The deconvolutional section is a series of transpose layers followed by convolutional layers. The transpose layers receive in input a concatenation of the previous layer and the skip connection of the last convolutional layer for each block in the VGG-16 architectures. This step is performed through a concatenation layer, where various inputs are concatenated together along the channel axis, providing a slow fusion of the low-level features and the features of the VGG-16 models. All the convolutional layers use a ReLU activation function, except the last layer, which uses a softmax activation function that generates a probability vector for the three classes involved (i.e., core, penumbra, and healthy brain).

The model's output is a single 2D brain slice image with the same height and width dimensions as

the input PMs. The 3D volume is generated by concatenating all the 2D brain slice images sequentially. To evaluate the accuracy of the predicted outputs, we use three distinct metrics: the Dice Coefficient, the Hausdorff distance, and the difference in volume among the predictions (V_p) and the ground truth images (V_g): $\Delta V = |V_g - V_p|$.

3.1 Class imbalance & Loss function

The dataset has imbalanced classes: 93.1% of the pixels belong to the healthy brain class, 6.2% penumbra, and the remaining 0.7% core. This issue is even more pronounced in the Non-LVO group, where 0.2% is core, 2.2% penumbra, and the remaining 97.6% belongs to healthy brain tissue. To overcome this issue, we build our model to focus on two aspects: the *loss function* and the *Non-LVO group*.

A generalized focal loss, based on the Tversky index (TI) [1, 18] was adopted. The TI is a generalization of the Dice similarity coefficient. The selected loss function was developed to address the data imbalance problem in medical image segmentation, improving the trade-off between precision and recall when training on small structures. γ, α, and β are hyper-parameters of the Focal Tversky loss (FTL) [1]. Furthermore, during training, we emphasized the misclassification of penumbra and core class, in patients in the Non-LVO group, with a higher penalty in the loss because of the evident imbalance among classes in this sub-group. A post-processing step is performed before generating the predicted outcome: a binary mask of the entire brain slice is created based on the MIP image to force the segmentation inside a valid area (the brain tissue). Subsequently, from the softmax activation function, the highest probability value for each pixel was selected.

4 Experiments & Results

In the reminder of the paper we define the models as: $\mathrm{SF}_w(input)$, where $w \in$ {**F**rozen, **U**nfrozen, **G**radual fine-tuning} and the input are combination of PMs, MIP images (M), and NIHSS score (N). We use the FTL as the loss function for our network. We perform three different experiments: *Exp-1)* Hyper-parameter search for the poposed method (Fig. 1). *Exp-2)* combination of inputs and freeze/unfreeze variations of VGG-16; *Exp-3)* comparison of different input-fusion methods (Fig. 2). For all experiments, the same setting was used: Adam [10] was used as the optimizer function. The batch size was set to 2, and each model was trained for 1000 epochs. The validation FTL was monitored, and an early stopping was invoked if there was no improvement after 25 consecutive epochs.

In *Exp-1* a hyper-parameter search is done running a large number of hyper-parameter combinations (over the same model) for finding the optimal values for the given task. We ran experiments with distinct values for FTL hyper-parameters $\gamma \in [1, 3], \alpha \in [0, 1]$, and $\beta \in [0, 1]$ as shown in Table 1, where each value represents the average of the patients in the validation set for the different severity levels.

Table 1: *Exp-1*: hyper-parameters search for the FTL loss, model SF_F(PMs). Each value represents the average of the patients in the validation set for the different severity levels and their standard deviation (SD).

Model	Parameters			Dice Coefficient (Avg.) ± SD			
				LVO		Non-LVO	
	γ	α	β	Penumbra	Core	Penumbra	Core
SF_F(PMs)	1	1	1	0.68±0.2	0.29±0.3	**0.30±0.3**	0.13±0.2
		0.3	0.7	0.36±0.2	0.37±0.3	0.06±0.1	0.21±0.3
		0.5	0.5	0.66±0.2	0.35±0.3	**0.30±0.4**	0.19±0.2
		0.7	0.3	0.68±0.2	0.31±0.3	0.28±0.3	0.17±0.2
	4/3	0.3	0.7	0.68±0.2	0.35±0.3	0.29±0.4	0.17±0.2
		0.5	0.5	0.69±0.2	0.36±0.3	**0.30±0.3**	0.17±0.3
		0.7	0.3	**0.71±0.1**	0.37±0.3	0.27±0.3	**0.22±0.3**
	1.5	0.3	0.7	0.67±0.2	0.30±0.3	0.27±0.3	0.14±0.2
		0.5	0.5	0.67±0.2	0.30±0.3	0.29±0.3	0.14±0.2
		0.7	0.3	0.70±0.3	0.37±0.2	0.29±0.3	0.20±0.3
	2	0.3	0.7	0.67±0.2	0.00±0.0	**0.30±0.4**	0.00±0.0
		0.5	0.5	0.70±0.2	0.00±0.0	0.28±0.3	0.00±0.0
		0.7	0.3	0.70±0.2	**0.39±0.3**	0.29±0.3	0.20±0.3
	3	0.3	0.7	0.00±0.0	0.00±0.0	0.00±0.0	0.00±0.0
		0.5	0.5	0.70±0.2	0.00±0.0	**0.30±0.4**	0.00±0.0
		0.7	0.3	0.68±0.2	0.37±0.3	0.29±0.3	0.19±0.3

For *Exp-2* we have combined our slow fusion multi-input baseline model with different combinations of inputs to understand if diversity in the input can improve training the model. The various combinations are shown in Table 2 with the corresponding results. From the baseline input (PMs), we added MIP images as input or the NIHSS score or combined both MIP images and NIHSS score. Models for all these experiments were based on a multi-input with slow fusion. VGG-16 models were trained with frozen weights, unfrozen weights, and a gradual fine-tuning approach for each experiment. The latter setting was developed in three steps:

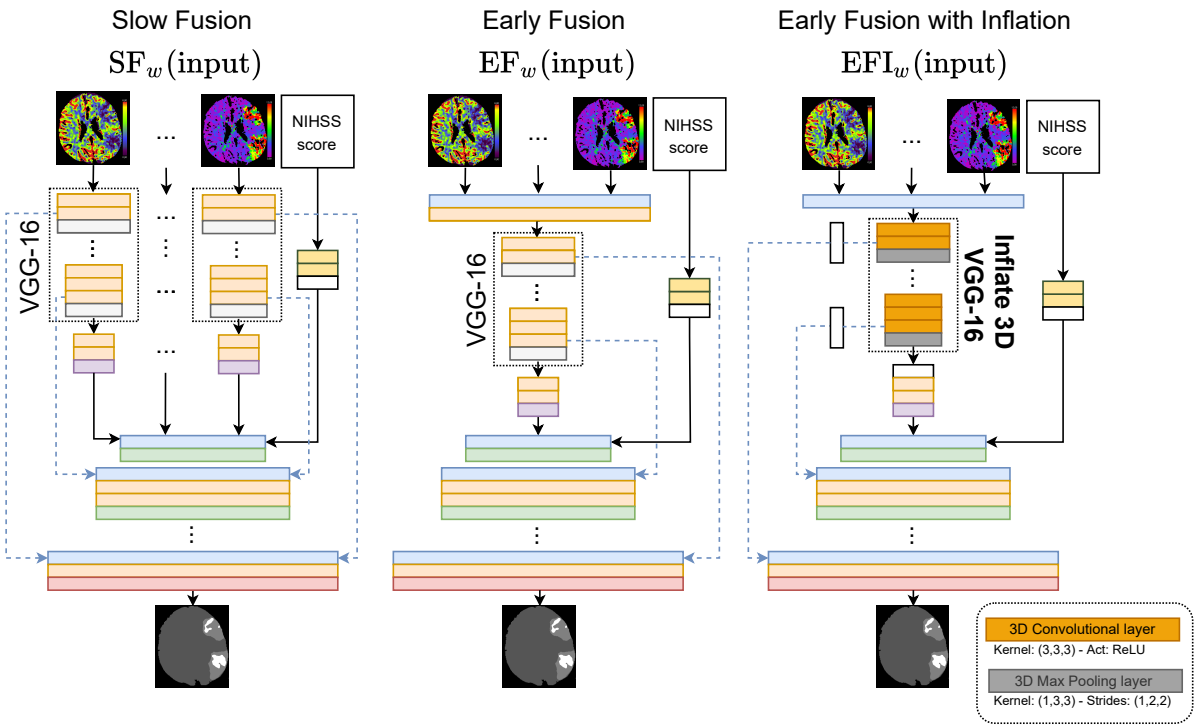

Figure 2: Overview of the three models used for comparison in *Exp-3*: the early fusion model ($EF_w(input)$); the early fusion with inflation ($EFI_w(input)$); $SF_w(input)$ is described in Fig. 1.

first, the model was trained with all the VGG-16 weights frozen; secondly, after monitoring the validation loss having no improvements for 25 consecutive epochs, the bottom half of the weights were unfrozen, and the training continued; at last, when no improvement in the validation loss was detected again, all weights were unfrozen, the training continued, and the validation loss was monitored. We have selected SF_G(PMs,N) as the proposed model.

To understand if using multi-input and slow fusion is suitable for this task, in *Exp-3* we compared it with two models: an early fusion (EF_G(PMs,N)), and early fusion with inflation (EFI_G(PMs,N)), adopting the same multi-input idea but with different fusion approacheas. The inflation approach converts 2D into 3D layers, adding a temporal dimension. The inflation followed the idea of the I3D network by Carreira et al. [4], where they introduced video classification models with an inflated ImageNet pre-trained image classification architecture. The setting and hyper-parameters of these two new architectures are the same as the selected model to maintain a fair comparison. Fig. 2 presents an overview of the model architectures. For EF_G(PMs,N), the four input parametric maps are concatenated together over the channel axis to generate a single input volume passing through a 2D convolutional layer to reduce the channel dimension to feed it to a single VGG-16 backbone. Differently, EFI_G(PMs,N)'s inputs are concatenated over the time dimension; the filters and pooling kernels of the VGG-16 architecture are inflated, remodeling squared filters into cubic filters. The bottom of Table 2 shows the results of these two models in comparison with the other experiments performed.

SF_G(PMs,N) is seen as having good overall performance and is chosen as the proposed model. The model is tested with a previously unseen test set, and the performance is compared with manual annotations from two different experts. Table 3 presents the test results and the inter-observer variability in comparison with two expert neuroradiologists. Fig. 3 shows examples of predictions

from six patients from the test set using the proposed model.

5 Discussion

We developed a multi-input CNN with early and slow fusion using transfer learning in this work. The proposed network aims to simultaneously segment dead (core) and salvageable tissues (penumbra) in AIS patients. The proposed method is learned on patients with or without ischemic stroke and for different vessel occlusion severities. This generalization helps the model correctly segment most ischemic regions, regardless of the patient group. After a series of experiments, multi-input including NIHSS, by a slow fusion approach, SF_G(PMs,N), is chosen as the proposed method based on the total of the results.

A hyper-parameter search was performed for the first experiment. Table 1 shows the average Dice coefficient for selecting the optimal hyper-parameters for the FTL function. Our observations showed that $\gamma = 4/3$, $\alpha = 0.7$, and $\beta = 1 - \alpha$ give satisfactory results for all the classes in the two

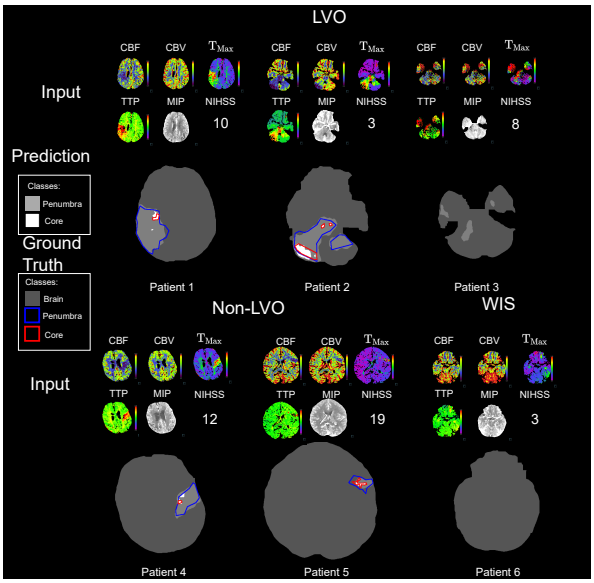

Figure 3: Prediction results for six test patients using the selected model, the set of all possible inputs, and the relative ground truth images.

Table 2: Statistical results over the validation set for the models divided for *Exp-2* and *Exp-3*. The last two rows contain results for *Exp-3*. Dice coefficient, Hausdorff distance, and the volume difference are the metrics considered to select the best model. Each value exhibits the average of the patients in the validation set and the standard deviation (SD) for the distinct groups. For the metrics: ⇑ indicates that higher values are better, while with ⇓ lower values are preferable. Highlighted values represent the best result for a specific class and metric. The selected model is highlighted inside a red rectangle.

Model	Input			Dice Coeff. (Avg.) ± SD ⇑		Hausdorff Dist. (Avg.) ± SD ⇓		ΔV (Avg.) ± SD (ml) ⇓		
				LVO	Non-LVO	LVO	Non-LVO	LVO	Non-LVO	WIS
	PMs	MIP	NIHSS	Penumbra				Core		
Exp-2 - Layer weights: **Frozen**										
SF_F(PMs)	X			**0.37±0.3** / 0.71±0.1	0.22±0.3 / 0.27±0.3	2.7±1.8 / 5.9±1.0	0.9±0.8 / 3.2±1.5	9.4±20 / 27.0±28	0.8±1.3 / 10.0±15	0.5±0.5 / 9.8±8.0
SF_F(PMs,M)	X	X		0.36±0.3 / 0.69±0.2	0.20±0.3 / 0.29±0.3	3.0±1.7 / 5.9±0.9	1.0±0.8 / 3.1±1.4	10.0±21 / 25.9±27	0.8±1.0 / 5.5±6.0	0.5±0.3 / 8.1±7.0
SF_F(PMs,N)	X		X	0.36±0.3 / 0.70±0.2	0.16±0.2 / 0.29±0.3	**2.3±1.9** / 5.6±1.2	0.7±0.6 / 2.3±1.3	5.0±9.0 / 27.0±37	0.6±1.2 / 4.1±4.8	0.1±0.1 / 4.0±3.4
SF_F(PMs,M,N)	X	X	X	0.34±0.3 / 0.68±0.2	0.18±0.3 / 0.3±0.3	2.5±1.8 / 5.7±1.4	0.8±0.7 / 2.4±1.2	6.3±12 / 29.9±37	0.5±1.0 / **2.8±3.0**	**0.0±0.0** / 3.9±4.0
Exp-2 - Layer weights: **Unfrozen**										
SF_U(PMs)	X			0.34±0.3 / 0.70±0.2	0.24±0.3 / 0.29±0.4	2.7±1.7 / 5.4±1.3	0.7±0.7 / 2.1±1.5	6.5±14 / 29.8±36	0.5±0.7 / 4.2±6.9	**0.0±0.0** / **0.1±0.1**
SF_U(PMs,M)	X	X		0.36±0.3 / 0.70±0.2	**0.24±0.3** / 0.34±0.3	2.5±1.8 / 5.6±1.4	0.6±0.6 / 2.3±1.6	6.5±8.1 / 28.1±40	0.4±0.8 / 4.9±9.8	**0.0±0.0** / 0.9±1.5
SF_U(PMs,N)	X		X	0.36±0.3 / **0.72±0.2**	0.23±0.3 / 0.29±0.3	2.6±1.8 / 5.6±1.1	0.9±0.7 / 2.7±1.4	5.2±10 / 28.8±28	0.5±0.7 / 7.8±12	0.1±0.1 / 3.4±2.7
SF_U(PMs,M,N)	X	X	X	0.36±0.3 / 0.71±0.2	0.22±0.3 / 0.32±0.3	2.8±1.7 / 5.7±1.0	1.0±0.8 / 2.4±1.6	7.0±14 / 29.0±25	0.8±0.9 / 9.0±18	0.1±0.1 / 5.5±4.5
Exp-2 - Layer weights: **Gradual Fine-tuning**										
SF_G(PMs)	X			0.35±0.3 / 0.71±0.2	0.19±0.3 / 0.30±0.3	2.4±1.8 / 5.8±1.0	0.8±0.7 / 3.0±1.4	4.7±7.6 / 29.3±25	0.6±0.8 / 11.8±22	**0.0±0.0** / 6.2±5.5
SF_G(PMs,M)	X	X		0.35±0.3 / 0.69±0.2	0.22±0.4 / **0.34±0.6**	2.4±1.8 / 5.4±1.4	**0.6±0.5** / **1.9±1.2**	5.2±9.4 / 27.8±38	**0.3±0.3** / 3.3±3.7	**0.0±0.0** / 2.4±3.8
SF_G(PMs,N)	X		X	**0.37±0.3** / **0.72±0.2**	0.21±0.3 / 0.31±0.3	2.4±1.7 / **5.3±1.4**	0.7±0.6 / 2.3±1.4	**4.4±7.0** / 26.0±35	0.5±0.7 / 6.5±14	**0.0±0.0** / 1.0±1.7
SF_G(PMs,M,N)	X	X	X	0.34±0.3 / 0.68±0.3	0.18±0.3 / 0.31±0.3	2.5±1.8 / 5.5±1.3	0.7±0.6 / 2.3±1.3	5.1±9.6 / 26.9±35	0.5±0.8 / 5.5±9.2	**0.0±0.0** / 0.9±1.1
Exp-3 - Layer weights: **Gradual Fine-tuning**										
EF_G(PMs,N)	X		X	0.26±0.3 / 0.68±0.2	0.19±0.3 / 0.32±0.3	2.4±1.9 / 5.6±1.4	0.7±0.6 / 2.3±1.5	8.0±21 / 31.8±47	0.7±0.9 / 5.3±10	**0.0±0.0** / 1.8±0.3
EFI_G(PMs,N)	X		X	0.29±0.3 / 0.67±0.2	0.16±0.2 / 0.29±0.3	2.5±1.9 / 6.0±0.8	0.7±0.6 / 3.3±1.5	5.3±10 / **25.5±28**	0.6±1.2 / 6.3±7.1	**0.0±0.0** / 6.7±4.7

Table 3: Evaluation metrics for the selected method (SF_G(PMs,N)) over the 33 test patients (19 LVO, 11 Non-LVO, 3 WIS) manually annotated by two experts neuroradiologists (NR_1 and NR_2).

Model	Dice Coefficient (Avg.) ± SD ⇑			Hausdorff Distance (Avg.) ± SD ⇓			ΔV (Avg.) ± SD (ml) ⇓			
	LVO	Non-LVO	All	LVO	Non-LVO	All	LVO	Non-LVO	WIS	All
	Penumbra						Core			
SF_G(PMs,N) vs (NR_1 & NR_2)	0.52±0.2 / 0.81±0.1	0.12±0.2 / 0.48±0.3	0.34±0.3 / 0.63±0.3	3.2±1.5 / 5.0±1.2	0.7±1.0 / 2.2±0.9	2.1±2.8 / 3.7±2.0	6.1±5.1 / 15.5±12	0.7±1.4 / 4.4±4.4	0.0±0.0 / 0.2±0.4	3.7±4.8 / 10.5±11
NR_1 vs. NR_2	0.44±0.2 / 0.78±0.1	0.15±0.2 / 0.65±0.2	0.30±0.3 / 0.67±0.2	3.2±1.4 / 5.1±±1.0	0.5±0.9 / 1.9±1.4	2.0±1.8 / 3.6±2.2	5.6±4.3 / 33.3±28	0.7±1.9 / 5.5±9.2	0.0±0.0 / 0.0±0.0	3.5±4.2 / 21.0±26
SF_G(PMs,N) vs. NR_1	0.57±0.2 / 0.84±0.1	0.12±0.2 / 0.48±0.3	0.37±0.3 / 0.64±0.3	3.0±1.4 / 4.9±1.1	0.7±1.0 / 2.2±1.0	1.9±1.7 / 3.6±1.9	6.2±5.7 / 11.2±11	0.7±1.4 / 4.4±4.4	0.0±0.0 / 0.3±0.4	3.8±5.2 / 7.9±9.4
SF_G(PMs,N) vs. NR_2	0.44±0.2 / 0.78±0.1	0.1±0.2 / 0.43±0.3	0.29±0.3 / 0.59±0.3	3.3±1.4 / 5.4±1.1	0.5±0.8 / 2.7±1.4	2.1±1.8 / 4.0±2.1	6.8±5.9 / 29.2±30	0.3±0.6 / 8.2±9.6	0.0±0.0 / 0.3±0.4	4.0±5.5 / 19.6±26

groups. Thus we apply these parameters in the following experiments. From the validation results of Table 2 it can be seen that unfreezing the weights of the VGG-16 feature extractors improves the models but gives an overestimation of the volume for both the classes and also that the gradual fine-tuning approach gives a slight improvement when compared to unfreezing all weights from the start. The choice of freezing the parameters reduces training time since a smaller set of parameters needs to be learned; however, the statistical results are less than satisfactory; this could be since PMs are not included in the ImageNet dataset, then the weights are not optimized for these images. Therefore, at the cost of a longer training time, fully unfreezing the weights or using a gradual fine-tuning technique will allow the model to familiarize and learn this particular dataset more accurately.

From the validation results of the different gradual fine-tuned models, it is clear that multi-input fusion, including the NIHSS or MIP images, is better than only PMs. However, it is not entirely clear if including both NIHSS and MIP images improves the models compared to only including NIHSS in addition to PMs. Based on the results presented in Table 2, we select SF_G(PMs,N) as the proposed model. One can argue that SF_U(PMs,N) yields similar results, but SF_G(PMs,N) is chosen because of its lower ΔV for the LVO group, high Dice coefficient over the entire dataset, and satisfactory results for all the other metrics. Furthermore, *Exp-3* favored the proposed slow fusion approach over the two early fusion approaches, EF_G(PMs,N) and EFI_G(PMs,N).

Results from the 33 randomly selected patients constituting the test set using our selected proposed model (Table 3) show an average Dice coefficient of 0.34 for core and 0.63 for penumbra over the entire test set and, on the LVO set, 0.81 and 0.52, respectively. These results present higher or analogous values compared to the inter-observer variability in most of the metrics, regardless of the stroke's severity. A separate comparison of the predicted outputs with both the neuroradiologists' annotations demonstrates that the proposed architecture can achieve high statistic values, regardless of the neuroradiologist. This achievement can be considered valuable throughout the first stages of an AIS.

Predictions from six brain slices by six patients in the test set are shown in Fig. 3. Results display comparable regions as the ground truth images, both with LVO and Non-LVO groups, showing high Dice coefficient and promising results with the model. However, the third example in Fig. 3 shows false-positive regions in the brain: this is possibly due to artifacts present in the generated color-coded PMs. We have noticed similar false-positive with all the other architectures and hyper-parameters as well. These false-positive might be avoided using 4D raw CTP datasets instead of the pre-generated PMs.

Several researchers have proposed methods with promising results to segment the ischemic core [2, 5, 9]. They all use PMs derived from CTP studies for their architectures. Nevertheless, their only focus was to segment the ischemic core without considering the penumbra; thus, excluding a critical aspect for medical treatment decisions. Furthermore, there was no differentiation in the severity level of AIS for the patients involved in their research.

6 Conclusion

The ability to achieve comparable results with two expert neuroradiologists for all the segmented classes (core and penumbra) is valuable to neuroradiologists during the first stages of an ischemic stroke. The proposed model might be a supportive tool that can help doctors to make treatment decisions.

References

[1] N. Abraham and N. M. Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687. IEEE, 2019. doi: 10.1109/ISBI.2019.8759329.

[2] S. M. Abulnaga and J. Rubin. Ischemic stroke lesion segmentation in ct perfusion scans using pyramid pooling and focal loss. In *International MICCAI Brainlesion Workshop*, pages 352–363. Springer, 2018. doi: 10.1007/978-3-030-11723-8_36.

[3] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, and Y. Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7): 1590, 2021. doi: 10.3390/cancers13071590.

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR.2017.502.

[5] A. Clèrigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, and X. Lladó. Acute ischemic stroke lesion core segmentation in ct perfusion images using fully convolutional neural networks. *Computers in Biology and Medicine*, 115:103487, 2019. doi: 10.1016/j.compbiomed.2019.103487.

[6] E. S. O. E. E. Committee, E. W. Committee, et al. Guidelines for management of ischaemic stroke and transient ischaemic attack 2008. *Cerebrovascular diseases*, 25(5):457–507, 2008. doi: 10.1159/000131083.

[7] K. Gadzicki, R. Khamsehashari, and C. Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6. IEEE, 2020. doi: 10.23919/FUSION45008.2020.9190246.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. doi: 10.1109/CVPR.2014.223.

[9] A. S. Kasasbeh, S. Christensen, M. W. Parsons, B. Campbell, G. W. Albers, and M. G. Lansberg. Artificial neural network computer tomography perfusion prediction of ischemic core. *Stroke*, 50(6):1578–1581, 2019. doi: 10.1161/STROKEAHA.118.022649.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. doi: 10.48550/arXiv.1412.6980.

[11] K. Kurz, G. Ringstad, A. Odland, R. Advani, E. Farbu, and M. Kurz. Radiological imaging in acute ischaemic stroke. *European journal of neurology*, 23:8–17, 2016. doi: 10.1111/ene.12849.

[12] R. LaLonde, I. Tanner, K. Nikiforaki, G. Z. Papadakis, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci. Inn: inflated neural networks for ipmn diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–109. Springer, 2019. doi: 10.1007/978-3-030-32254-0_12.

[13] C. Lucas, A. Kemmling, A. M. Mamlouk, and M. P. Heinrich. Multi-scale neural network for automatic segmentation of ischemic strokes on acute perfusion images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1118–1121. IEEE, 2018. doi: 10.1109/ISBI.2018.8363767.

[14] Ø. Meinich-Bache, S. L. Austnes, K. Engan, I. Austvoll, T. Eftestøl, H. Myklebust, S. Kusulla, H. Kidanto, and H. Ersdal. Activity recognition from newborn resuscitation videos. *IEEE journal of biomedical and health informatics*, 24(11):3258–3267, 2020. doi: 10.1109/JBHI.2020.2978252.

[15] B. Murphy, A. Fox, D. Lee, D. Sahlas, S. Black, M. Hogan, S. Coutts, A. Demchuk, M. Goyal, R. Aviv, et al. Identification of penumbra and infarct in acute ischemic

stroke using computed tomography perfusion–derived blood flow and blood volume measurements. *Stroke*, 37(7):1771–1777, 2006. doi: 10.1161/01.STR.0000227243.96808.53.

[16] S. Ojaghihaghighi, S. S. Vahdati, A. Mikaeilpour, and A. Ramouz. Comparison of neurological clinical manifestation in patients with hemorrhagic and ischemic stroke. *World journal of emergency medicine*, 8(1):34, 2017. doi: 10.5847/wjem.j.1920-8642.2017.01.006.

[17] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

[18] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017. doi: 10.1007/978-3-319-67389-9_44.

[19] L. Tomasetti, K. Engan, M. Khanmohammadi, and K. D. Kurz. Cnn based segmentation of infarcted regions in acute cerebral stroke patients from computed tomography perfusion imaging. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–8, 2020. doi: 10.1145/3388440.3412470.

[20] L. Tomasetti, L. J. Høllesli, K. Engan, K. D. Kurz, M. W. Kurz, and M. Khanmohammadi. Machine learning algorithms vs. thresholding to segment ischemic regions in patients with acute ischemic stroke. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2021. doi: 10.1109/JBHI.2021.3097591.

[21] H. Wang, M. Naghavi, C. Allen, R. M. Barber, Z. A. Bhutta, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen, M. M. Coates, et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. *The lancet*. doi: 10.1016/S0140-6736(16)31012-1.

[22] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. A. Janssen. A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. *Technology in Cancer Research & Treatment*, 19, 2020. doi: 10.1177/1533033820946787.