Machine learning algorithms vs. thresholding to segment ischemic regions in patients with acute ischemic stroke

Tomasetti Luca, Liv Jorunn Høllesli, Kjersti Engan, *Senior Member, IEEE*, Kathinka Dæhli Kurz, Martin Wilhelm Kurz, and Mahdieh Khanmohammadi

***Abstract*— *Objective:* Computed tomography (CT) scan is a fast and widely used modality for early assessment in patients with symptoms of a cerebral ischemic stroke. CT perfusion (CTP) is often added to the protocol and is used by radiologists for assessing the severity of the stroke. Standard parametric maps are calculated from the CTP datasets. Based on parametric value combinations, ischemic regions are separated into presumed infarct core (irreversibly damaged tissue) and penumbra (tissue-at-risk). Different thresholding approaches have been suggested to segment the parametric maps into these areas. The purpose of this study is to compare fully-automated methods based on machine learning and thresholding approaches to segment the hypoperfused regions in patients with ischemic stroke. *Methods:* We test two different architectures with three mainstream machine learning algorithms. We use parametric maps as input features, and manual annotations made by two expert neuroradiologists as ground truth. *Results:* The best results are produced with random forest (RF) and *Single-Step* approach; we achieve an average Dice coefficient of 0.68 and 0.26, respectively for penumbra and core, for the three groups analysed. We also achieve an average in volume difference of 25.1ml for penumbra and 7.8ml for core. *Conclusions:* Our best RF-based method outperforms the classical thresholding approaches, to segment both the ischemic regions in a group of patients regardless of the severity of vessel occlusion. *Significance:* A correct visualization of the ischemic regions will guide treatment decisions better.**

***Index Terms*—Computed tomography perfusion; Ischemic stroke; Machine learning; Thresholding.**

I. INTRODUCTION

CEREBRAL stroke is the second leading cause of death and the third leading cause of disability worldwide [1]. Despite significantly reduced incidence over the past years in the entire world, the worldwide prevalence of cerebral stroke is estimated to be 17 million strokes causing 6.5 million deaths per year [1], [2]. In Norway, acute cerebral stroke is the third leading cause of death in adults and the leading cause of disability and admission to nursing homes [3], [4]. Changes in demography will result in a predicted 34% increase in stroke incidence in Europe between 2015 and 2035, which is likely to be mirrored in other parts of the world [2]. Thus, cerebral stroke has a huge socio-economic impact on society and a tremendous impact on the quality of life for every single patient [5].

There are two broad categories of cerebral stroke; hemorrhagic and ischemic stroke. Approximately 20% of all strokes are due to hemorrhage, while approximately 80% are due to ischemia [6]. Both groups can further be divided into different subtypes. Ischemic stroke may be caused by arteriosclerosis, thrombi, emboli, dissections, or systemic hypoperfusion, all of them leading to ischemia due to reduced blood flow in regions of the brain.

The severity of ischemia usually varies within the area of reduced blood flow, and for clinical use, the area is divided into two distinct regions: ischemic core and penumbra. The ischemic core is defined as irreversibly damaged brain tissue [7]. The tissue within the penumbra is critically hypoperfused and is located around and adjacent to the infarct core. If blood flow is restored timely, this tissue may regain neurological function [7]. If the blood flow remains low, however, the area of penumbra will transfer into an irreversibly damaged infarct core. The ischemic penumbra was introduced by Astrup et al. as "a region of hypoperfused, electrically silent, and functionally impaired but viable tissue" [8]. Restoring blood flow and thereby preventing the penumbra from proceeding to irreversibly damaged infarct core, is the main treatment goal in patients with acute ischemic stroke (AIS). Penumbra may change into infarct core rapidly in AIS patients. Therefore, rapid recognition of stroke symptoms and acute treatment in a stroke center are of vital importance.

According to the European Stroke Organization guidelines, Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) are the two modalities recommended for diagnostic imaging in acute stroke patients [9]. MRI with diffusion-weighted imaging (DWI) is superior to CT scans for detection of small acute infarctions and identification of some stroke mimics. Nevertheless, CT is the preferred imaging modality in many centers for acute stroke patients due to its widespread availability, rapid scan times, and its high sensitivity for detecting hemorrhage. DWI has been considered the gold standard for ischemic core estimation [10]–[14]; however, there are very few hospitals where MRI are used as the first imaging tool in acute stroke patients, since it is not always timely available on a 24/7 basis, plus, some patients have contraindications for this type of modality. MRI is usually performed within the first days after an AIS. Treatment, timing of treatment, and other variables will affect further development of the penumbra. Hence, any core of follow-up MRI might have developed after the acute imaging and might not be comparable with the imaging results in the acute setting. In the last years, DWI has been contested as the de-facto gold standard since it cannot accurately differentiate irreversibly ischemic tissue from salvageable tissue [15], [16], and it has been shown that the detected ischemic regions can be partially reverse, especially if DWI is performed in the early window time [16]–[18].

The study is approved by the Regional ethic committee project 2012/1499. (Luca Tomasetti and Liv Jorunn Høllesli contributed equally to this work.) *(Corresponding author: Luca Tomasetti)*

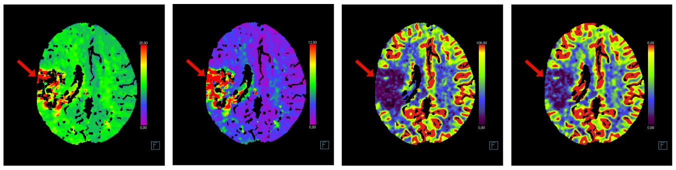
Tomasetti Luca, Kjersti Engan, and Mahdieh Khanmohammadi are with the Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway (e-mail: luca.tomasetti@uis.no; kjersti.engan@uis.no; mahdieh.khanmohammadi@uis.no)

Liv Jorunn Høllesli and Kathinka Dæhli Kurz are with the Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway and also with Stavanger Medical Imaging Laboratory (SMIL), Department of Radiology, Stavanger University Hospital, 4019 Stavanger, Norway (email: liv.jorunn.hollesli@sus.no; kathinka.dehli.kurz@sus.no)

Martin Wilhelm Kurz is with the Neuroscience Research Group, Stavanger University Hospital, 4019 Stavanger, Norway, and also the Department of Neurology, Stavanger University Hospital, 4019 Stavanger, Norway and also with the Department of Clinical Medicine, University of Bergen, 5007 Bergen, Norway (email: friedrich.martin.wilhelm.kurz@sus.no)

At Stavanger University Hospital (SUS), patients with suspected acute stroke are routinely investigated with non-contrast computed tomography (NCCT) of the head, CT angiography (CTA) of the precerebral and cerebral arteries, i.e. arch to vertex angiogram, and CT Perfusion (CTP) immediately after hospital admission. In most cases MRI including DWI is performed during the next days. In patients with suspected stroke with unknown time of symptom onset, MRI with DWI is used as a first-line diagnostic tool upon hospital admission.

Whether treatment is applied depends on time from symptom onset to hospital admission, but also largely depends on imaging results with CT Perfusion being the key-modality for patient selection. In CTP a time series of three-dimensional (3D) datasets are acquired during contrast agent injection. Based on the changes in the tissue density over time, color-coded parametric maps are calculated. The different parametric maps highlight spatio-temporal information from the passage of the contrast agent within the brain tissue. Generally, parametric maps based on CTP are generated in two steps: the first step acquires a time-density curve for each pixel based on the track of the contrast agent. The second step consists of extracting specific information from the generated time-density curves. Cerebral blood flow (CBF), cerebral blood volume (CBV), time-to-peak (TTP), mean transit time (MTT) and time-to-maximum (T_{Max}) are all examples of parametric maps [7]. Radiologists use parametric maps for diagnosis and treatment planning and are indirectly assessing penumbra and core by evaluating such parametric maps. An example of the parametric maps of a single brain slice, involved in this study, is given in Fig. 1.



(a) TTP. (b) T_{Max}. (c) CBF. (d) CBV.

Fig. 1. Parametric maps of a single slide of a patient's brain. In this patient there is an ischemic area on the right side in the vascular territory of the middle cerebral artery (pointed by a red arrow). TTP = time-to-peak; T_{Max} = time-to-maximum; CBF = relative cerebral blood flow; CBV = relative cerebral blood volume. Color in the online version.

Time is a fundamental factor for patients affected by an ischemic stroke. Automation of the recognition process for the ischemic regions, penumbra and core, can be immensely helpful for medical doctors for treatment decisions. Over the last decades, different methods and parameters were tested to find the most suitable approach to segment the ischemic regions using parametric maps as input.

Region growing is a technique to extract connected areas in an image based on pixel information; this method is defined as semi-automatic because the user manually selects a seed for the growing region algorithm. This technique was used by Matesin et al. [19] in relation with CT head images of stroke lesions, and by Dastidar et al. [20], for measuring the volumetric infarction using 3D T2 Fast Spin Echo MRI in patients affected by stroke. Their goal was to delineate ischemic areas, and not to distinct the core from the penumbra. The first delineation of both areas, using a region growing technique in combination with the parametric maps acquired by CTP analyses, was implemented by Contin et al. [21].

A series of studies have proposed experiments with threshold values on the derived parametric maps to improve the results achieved by the region's growing approaches. Different thresholds have been proposed for different parametric maps, generated from different vendors, and applied to various datasets [10]–[12], [22] to estimate both the ischemic regions, or the infarct core, or penumbra. These studies have used follow-up images (such as DWI or NCCT), acquired hours later after the stroke onset, to delineate the ground truth of the infarct regions and used them as a comparison for their predictions. For this reason, studies using DWI as follow-up imaging present some limitations: they only included patients who were later identified with infarct lesions in follow-up images, excluding the ones who underwent the same routine at the time of hospital admission but did not show any lesion in the follow-up DWI; they also excluded patients with contraindication for MRI. Moreover, since the threshold values were compared with final infarctions, assessed after the patient's treatment, they do not present a perfect estimation of the infarctions before treatment decision; thus, they are not the best candidates to help medical doctors during the treatment making decision. Furthermore, the studies have proposed quite distinct thresholding values due to the different vendors used for post-processing evaluation and the distinct window of time (≤ 1 hour to 7 days) used for follow-up images to evaluate the ground truth for the ischemic regions. Thus, there is no real consensus to properly define the ischemic regions based on threshold values on the parametric maps derived from CTP.

In recent years, Machine Learning (ML) and neural network algorithms have achieved promising results in a large number of medical image analysis applications, and have also made their way into the stroke application [23]–[27]. Kemmling et al. proposed a generalized linear model using the parametric maps as input and clinical data to quantify changes of tissue infarction [23]. Qiu et al. implemented a ML-based algorithm to detect early infarction in patients with AIS using NCCT as input and follow-up DWI as ground truth [24]. Kasasbeh et al. used a semi-automatic approach based on a convolutional neural network (CNN) with the entire set of parametric maps as input to classify the infarct core using follow-up DWI as ground truth [26]. However, these ML and CNN based methods were only trained to classify the infarct core regions and did not find the penumbra areas. Differently, Qiu et al. developed two distinct ML models, using a multiphase CTA as input and DWI/NCCT follow-up images as ground truth, to predict core and penumbra [25]. Their primary goal was to demonstrate the validity of using multiphase CTA in comparison to CTP imaging for evaluating ischemic regions, but they stated limitations in their data material. Nevertheless, using follow-up images for delineating the ischemic regions limits the usability for medical doctors since they might not be helpful for treatment decisions but just for comparison with the clinical outcome. Our research group was, to the best of our knowledge, the first using the entire 4D CTP data as input to a neural network to segment both penumbra and core simultaneously. A modified U-Net model was used in a small pilot study to segment both penumbra and core regions using the entire 4D CTP volume as input and with ground truth generated with manual expert assessment directly from the parametric maps [27]. The results were promising, but they were based on a very small pilot study and need to be validated on a larger sample size.

Before continuing to use the entire 4D dataset as input, we wish to study the utility of automatically segmenting the penumbra and core based on the parametric maps that are already calculated in the standard software used in clinical practice. Based on the ideas and the shortcomings of the published methods, we propose in this paper a ML-based method using the parametric maps as input and both core and penumbra regions as output, in addition to healthy tissue. One can argue that CNN naturally fits this type of problem; nevertheless, several examples of classical ML methods with this application can be found in the literature [23]–[25] using follow-up images as ground truth, bearing with them the same issues mentioned earlier. Moreover, learning good CNN models usually require large datasets, and/or transfer-learning, and we have a limited dataset to work with. Thus, we aim to properly understand if well-established ML models, less data-hungry and complex than CNN models, can help to predict both the ischemic regions and have the potential to assist medical doctors during treatment decisions. We give a comparison of the proposed method with different parameters and with thresholding methods from the literature. This paper contributes with the following:

- Proposing a fully-automatic ML-based algorithm to segment both penumbra *and* infarct core regions in patients affected by AIS, since a correct visualization of the salvageable tissue will guide treatment decision better,
- Using the parametric maps as input, due to their wide usage by medical doctors for early assessment of ischemic strokes,
- Training the models using a dataset with different groups of patients based on their level of vessel occlusion, generalizing the models and the training data and not restricting the type of patients that can be tested,
- Adopting as ground truth, images annotated by expert neuroradiologists directly from the parametric maps based on CTP,
- And finally, testing different ML algorithms and parameters to find the most suitable approach. Both a single-step approach, segmenting normal brain, penumbra, and core in one go; and a two-step approach, segmenting penumbra and core individually before combining them, were tested. This was further compared to thresholding approaches.

II. Data Material

A. Dataset and ground truth

1) Context: Stavanger University Hospital (SUS) serves a population of 365.000. Close to 450 patients with AIS are annually admitted to the hospital. All consecutive patients with suspected AIS having received intravenous thrombolytic therapy are prospectively listed in a population-based database. Information about clinical severity measured by the National Institutes of Health Stroke Scale (NIHSS, scoring scale assessing neurological deficit) on admission, and at discharge are available. Long term functional outcome measured by the modified Rankin scale (mRS, scoring scale assessing long term functional outcome) at 90 days are also registered, in addition to mRS on hospital admission.

2) Dataset: The dataset in this study comprises CTP scans from 152 patients between January 2014 and August 2020. 137 of these patients had an AIS with visible perfusion deficit. Patients with AIS were divided into the following groups: 77 patients with large vessel occlusion (LVO), and 60 patients with non-large vessel occlusion (Non-LVO) Additionally, 15 patients without ischemic stroke (WIS) who were admitted with suspicion of stroke, but turned out not to have a stroke in the diagnostic workup, were included in the dataset. Age, gender, and NIHSS score for the groups are shown in Table I.

TABLE I
PATIENT CHARACTERISTICS.

		LVO	**Non-LVO**	**WIS**
Age (average/range)		72 (39-94) years	75 (41-94) years	60 (27-85) years
Gender	Male	49 (64%)	37 (62%)	8 (53%)
	Female	28 (36%)	23 (38%)	7 (47%)
NIHSS score (maximum /minimum /average)	On hospital admission	38/0/13	19/0/6	14/1/3
	On hospital discharge	25/0/5	10/0/2	1/0/0

LVO was defined using CT angiography; occlusion of the internal carotid artery, M1 and proximal M2 segment of the middle cerebral artery, A1 segment of the anterior cerebral artery, P1 segment of the posterior cerebral artery, basilar artery, and vertebral artery occlusion were regarded LVO. Non-LVO was defined as patients with perfusion deficits and affection of more distal arteries or with perfusion deficits without visible proximal artery occlusion.

3) Ground truth: Ground truth images are manually annotated by two expert neuroradiologists. The manual annotations are done using the entire set of the CT examination including the parametric maps from the CTP (CBV, CBF, TTP, T_{Max}), the maximum intensity projection (MIP) images, calculated as the maximum Hounsfield unit value over the time sequence of the CTP, providing a 3D volume from the 4D acquisition of CTP. Furthermore, the MRI examination performed within 1 to 3 days after the CT examination was used in assistance to generate the ground truth images. In-house developed software was used for the annotations.

B. Imaging protocol and Analysis

The CT scanners used for image acquisition were Siemens Somatom Definition Flash (installed in 2012) and a Siemens Somatom Definition Edge (installed in 2014), Erlangen, Germany.

Patients with suspected acute cerebral stroke with symptom onset within 4,5 hours prior to hospital admission were routinely investigated by NCCT of the head. If contraindications were excluded, intravenous thrombolysis bolus-dose was administered in the CT lab. Then CTA and CTP were performed. Technical details about the protocols are shown in Table II. Further, the CTP images were analyzed using the software "syngo.via" from Siemens Healthineers with manufacturer default settings to generate color-coded parametric maps (CBF, CBV, TTP, MTT, and T_{Max}).

TABLE II
COMPUTED TOMOGRAPHY TECHNICAL PROTOCOL FOR ACUTE ISCHEMIC STROKE.

	NCCT of the head	**CTA of the cerebral arteries**	**CT perfusion**
Patient position	Head first, supine	Head first, supine	Head first, supine
Spiral/sequence	Spiral	Spiral	Spiral
kV	120	100	80
mAs	280	160	200
Rotation time (s)	1	0.28	0.28
Slice collimation	3 mm c 20 x 0.6 mm	0.6 mm c 128 x 0.6 mm	5 mm c 32 x 1.2 mm
Pitch	0.55	1.0	-
X-care	Yes	No	No
IV contrast	No	60 ml Omnipaque 350 mg I/ml + 40 ml NaCl	40 ml Omnipaque 350 mg I/ml + 40 ml NaCl
Flow rate	-	5 ml/second	6 ml/second
Start delay	-	4 seconds	4 seconds, $\geq$60 seconds after CTA
Scan direction	Caudocranial		

III. Ischemic segmentation by Thresholding

Several studies define threshold values on some of the parametric calculations or on a combination of them to segment the ischemic stroke regions. The variability in the chosen thresholding value(s) is mainly due to the various vendors used for post-processing the parametric maps, the different definitions of the ground truth for the ischemic regions. It also lies in the decision of using the entire brain or just the ipsilesional hemisphere in statistical evaluations. Table III lists some of them in addition to information about their dataset, the number of patients, NIHSS score, time of stroke onset, vendor used, and their defined threshold values on different parametric maps. It also shows the different optimal thresholds that are proposed in each of these studies to segment either core, penumbra, or both.

Most of the listed studies evaluated their method by testing the mismatch between values from parametric maps derived from CTP images and the corresponding follow-up DWI, as the gold standard. The only study which did not use DWI as ground truth for the ischemic regions is Murphy et al. [22]. They defined the core region 5 to 7 days after the onset of stroke in the NCCT images, while the penumbra was the difference between the infarct and ischemic

TABLE III
INFORMATION ABOUT THE DATASET AND THE THRESHOLD VALUE(S) OF THE VARIOUS RESEARCH METHOD ANALYZED.

Article	Patients	NIHSS (mean)	Vendor	Stroke onset	Follow-up Images	Threshold	
						Penumbra	Core
Bathla et al. [28]	39	7	Siemens	N.A.	$\leq$ 24h	$T_{Max}>6s$	CBF<20%
Wintermark et al. [11]	130	15.3	Philips	$\leq$ 12h	$\leq$7d	MTT $>$ 145%	CBV $\leq$ 2.0ml/100g
Campbell et al. [12]	49	16.5	Philips	$\leq$ 6h	$\leq$1h	$T_{Max}>6s$	CBF<31% (with TTP >4s)
Cereda et al. [10]	103	16	In-house	$\leq$ 8h	$\leq$3h	N.A.	CBF<38% (with $T_{Max}>4s$)
Bivard et al. [13]	180	12	Toshiba	$\leq$ 6h	$\leq$24h	TTP>+5s	CBF<50%
Murphy et al. [22]	25	15.1	General Electric	$\leq$ 7h	N.A.	CBF$\leq$ 25ml/100g CBV$\leq$ 2.15ml/100g	CBF $\leq$ 13.3ml/100g CBV$\leq$ 1.12ml/100g
Schaefer et al. [14]	55	14	General Electric	$\leq$ 9h	$\leq$3h	N.A.	CBF$\leq$15% + CBV$\leq$30%

region. Nevertheless, they state that this difference “could lead to an underestimation of the final infarct size”. All the approaches displayed in Table III, with differences in their chosen parametric maps and the optimal values, demonstrate the lack of a consensus to define the ischemic regions based on thresholds.

Only the default setting used by “syngo.via” to define the ischemic regions after the parametric maps generation (CBF<27ml/100ml/min to define tissue at risk and CBV<1.2ml/100ml for non-viable tissue) and the thresholds proposed by Bathla et al. [28] were implemented for comparison with our best method due to the usage of the same vendor and software system as our input. We compare with a gold standard based on expert assessment of the parametric maps and manual delineation of the regions since these expert assessments are used normally for treatment decisions and are clinically relevant.

IV. Machine Learning Approaches

Applying ML algorithms in the field of medical image analysis is rapidly growing [29]. To train state-of-the-art ML models, patient data sets that have the necessary size and quality of samples are needed. Given that the patient data is protected by strict privacy and security rules this can be a challenge, however, if the necessary training set is available to train appropriate ML algorithms, good prediction models can be obtained. The ML models tested in this study include Support Vector Machine, Decision Tree learning, and Random Forest. Each ML algorithm uses in input a training set $T = \{(x_1, y_1), \ldots, (x_T, y_T)\}$, composed of x_i features vectors and the relative y_i class label.

Support Vector Machine (SVM) is an algorithm used for binary classification that creates a line or a hyperplane, which separates the features from the input data into classes. In 1992, Boser et al. [30] proposed a supervised classification algorithm that has evolved into SVM as we know it today.

Decision Tree learning (DT), firstly introduced by Breiman et al. [31], is an efficient classification technique that creates a tree-like structure by computing the relationship between independent features and a target. DT covers both binary and multi-class classification. The tree splits into branches by using conditions at each internal node and the end of the branch that does not split anymore is the decision (leaf).

Random Forest (RF) is a supervised learning algorithm and the “forest” consists of an ensemble of decision trees. To classify a new object from an input vector, the input vector is fed to each tree in the forest and each tree casts a unit vote for the most popular class at the input vector. Finally, the forest chooses the classification having the most votes. Breiman proposed this algorithm to minimize a possible overfitting problem generated by the usage of a single DT [32].

V. Proposed Method

In this paper, we test a single and a two-step method for segmenting core and penumbra in patients suspected of AIS using machine learning based on the parametric maps (CBF, CBV, TTP, and T_{Max}), derived from CTP datasets acquired at admission, the MIP map, and the NIHSS score. Various stages are performed during the proposed methods: (1) *Brain extraction and data imbalance:* extracting the brain tissue from the parametric maps to use only the pixel values inside the brain as input features, (2) *SLIC:* obtaining the 3D superpixel version of the parametric maps (CBF, CBV, T_{Max}, and TTP), (3) *Machine Learning algorithm:* Feeding the features from the parametric maps and their generated superpixel to our implemented machine learning algorithms to predict the ischemic regions.

Fig. 2 shows the flowchart of our proposed methods. In the reminder of the paper, we call them ***Single-Step*** and ***Two-Step*** approaches. The features used for the proposed methods are the four parametric maps, the MIP map, and the NIHSS score. The input to the *Single-Step* method is all the aforementioned features (top part of Fig. 2) and it classifies both core and penumbra simultaneously. The *Single-Step* approach was tested with the DT and RF algorithms, but not with the SVM model since our implemented SVM model performs only binary classifications. In addition to the *Single-Step* method, we test another multi-stage classification method, which is simply adapted from the way neuroradiologists at SUS perform during the treatment decision process. The *Two-Step* approach is based on:

Step1: Takes as input the MIP, TTP, and T_{Max} maps, plus the NIHSS score; it performs a prediction of the penumbra region and outputs a binary image showing the predicted penumbra.

Step2: CBV and CBF parametric maps are used as input; it predicts the ischemic core resulting in a binary image.

A. Brain extraction and Data imbalance

We introduce a preprocessing step to extract the brain tissue from the whole image and work with pixels within the brain tissue (BT). In the reminder of the paper, the set of pixels belonging to the brain tissue for all patients p is called $\mathrm{BT} = \bigcup \mathrm{BT}^p$, while the various parametric maps are called CBF ^{p}, CBV ^{p}, TTP ^{p}, and $\mathrm{T_{Max}}^p$. This step helps to balance the classes inside the dataset. Moreover, we convert the pixel values into a $[0, 1]$ interval for each input feature based on the color bar on the right of each corresponding parametric map. Each input feature is mapped with the corresponding color in the bar and transformed into a value in the $[0, 1]$ interval, where the value 0 corresponds to the bottom value in the bar, while the value 1 indicates the top value. This was performed to reduce each input feature into a single value instead of keeping all three color channels.

B. Superpixel (SLIC)

A modified version of the Simple Linear Iterative Clustering (SLIC) algorithm [33] is employed to generate superpixel regions in the parametric maps. The regions are based on the initial segmentation of the intensity values of the maps. Using SLIC, we stacked each slice to obtain a 3D superpixel version for each parametric map and used it as extra features as input to the model. These new features should help the models to consider the adjacent pixels along the third dimension (z-axis). In the reminder of the paper, the superpixel version of the parametric maps for a patient p are called: CBF $^p_{\mathrm{SLIC}}$, CBV $^p_{\mathrm{SLIC}}$, TTP $^p_{\mathrm{SLIC}}$, and $\mathrm{T_{Max}}^p_{\mathrm{SLIC}}$. SLIC generates superpixel regions by clustering pixels utilizing their proximity and similarity in the image plane. An example of a normalized TTP map from one of the patients analyzed and the generated superpixel image is given in Fig. 3.

C. Machine Learning for core and penumbra

We implement three mainstream classical ML methods including, support vector machines, decision tree, and random forest. To the best

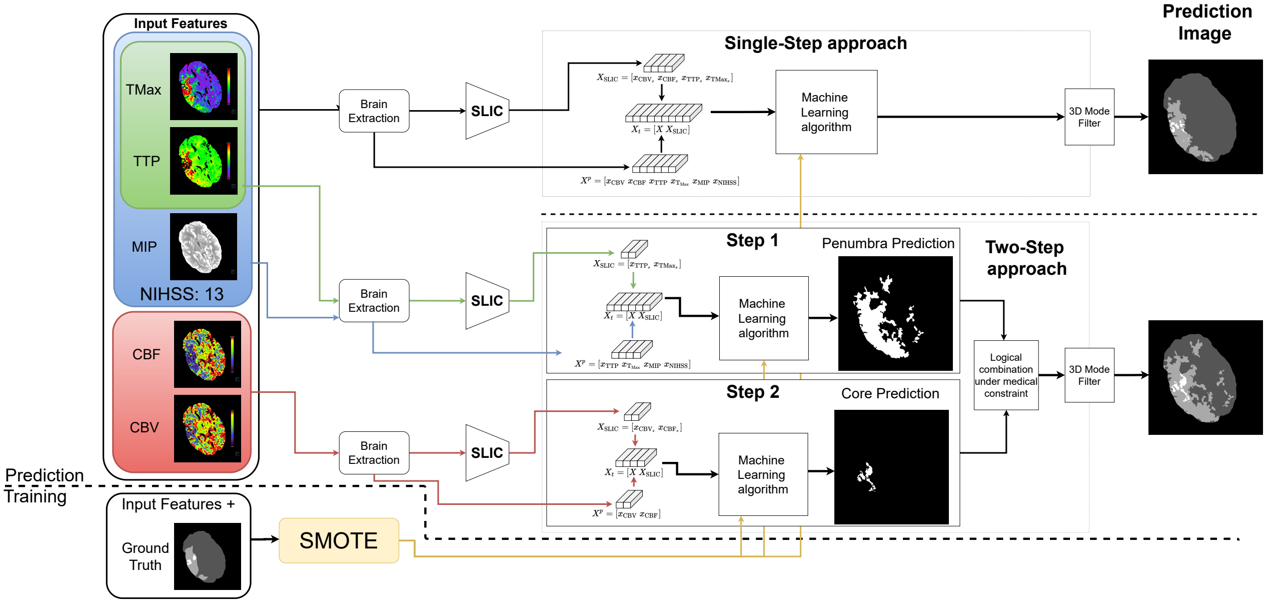

Input Features
TMax
TTP
MIP
NIHSS: 13
CBF
CBV
Prediction
Training
Input Features +
Ground Truth
SMOTE
Brain Extraction
SLIC
Brain Extraction
SLIC
Brain Extraction
SLIC
Single-Step approach
$X_{\text{SLIC}} = [x_{\text{CBV}_s}, x_{\text{CBF}_s}, x_{\text{TTP}_s}, x_{\text{TMax}_s}]$
$X_t = [X\ X_{\text{SLIC}}]$
$X^p = [x_{\text{CBV}}\ x_{\text{CBF}}\ x_{\text{TTP}}\ x_{\text{T}_{\text{Max}}}\ x_{\text{MIP}}\ x_{\text{NIHSS}}]$
Machine Learning algorithm
3D Mode Filter
Prediction Image
Step 1
Penumbra Prediction
$X_{\text{SLIC}} = [x_{\text{TTP}_s}, x_{\text{TMax}_s}]$
$X_t = [X\ X_{\text{SLIC}}]$
$X^p = [x_{\text{TTP}}\ x_{\text{T}_{\text{Max}}}\ x_{\text{MIP}}\ x_{\text{NIHSS}}]$
Machine Learning algorithm
Two-Step approach
Logical combination under medical constraint
3D Mode Filter
Step 2
Core Prediction
$X_{\text{SLIC}} = [x_{\text{CBV}_s}, x_{\text{CBF}_s}]$
$X_t = [X\ X_{\text{SLIC}}]$
$X^p = [x_{\text{CBV}}\ x_{\text{CBF}}]$
Machine Learning algorithm

Fig. 2. Visual description of the proposed multi-classification methods: for the *Single-Step* approach, all the parametric maps are adopted as input features for the ML algorithm to generate a final prediction image. The *Two-Step* approach work in a different way: Step1 takes in input six features for each pixel inside the brain and generates a binary map to classify the penumbra region(s) in a brain slice; Step2 takes in input 4 features, for each pixel, from different parametric maps and returns as output a binary map containing the predicted core region(s) if any. The final prediction combines the two binary maps only including the core regions that are inside the penumbra regions. A final post-processing step using a 3D mode filter is implemented. SLIC refers to the algorithm to extract superpixel regions. Color in the online version.

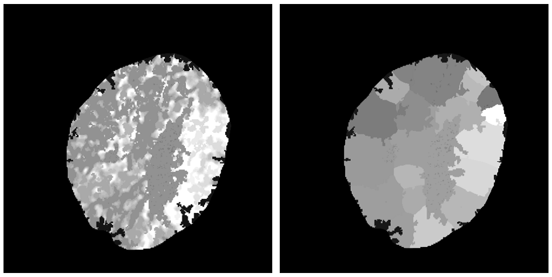


Fig. 3. Visual comparison of a TTP map in grayscale (left) and the generated superpixel image (right) after the brain extraction.

of our knowledge, there exists no defined convention on which of the parametric maps should be used to detect core and which shows penumbra better. Let L^p be the number of pixels in BT^p. In the training phase, the totality of input features to these ML approaches are defined as a matrix. For a patient p, let the input features vector for CBV be:

$$x^p_{\text{CBV}} = \text{stack}(\text{CBV}^p(i,j))_{\forall (i,j) \in \text{BT}^p}$$

where x^p_{CBV} is a vector of size L^p. The stack function concatenates all the pixels in an image, row-by-row, into a vector. The input features totality of the parametric maps for a patient p is given by the matrix X^p. For simplicity we omit the p in the following notation where all definition are on a single patient:

$$X = [x_{\text{CBV}}\ x_{\text{CBF}}\ x_{\text{TTP}}\ x_{\text{T}_{\text{Max}}}\ x_{\text{MIP}}\ x_{\text{NIHSS}}]$$

Defining $[\mathbf{1}]$ as a all-ones vector of length L^p, x_{NIHSS} is defined as $x_{\text{NIHSS}} = \text{NIHSS} \cdot [\mathbf{1}]$. In the same way, the input features totality for the superpixel version of the parametric maps is given by the matrix X^p_{SLIC}, defined as:

$$X_{\text{SLIC}} = [x_{\text{CBV}_s}\ x_{\text{CBF}_s}\ x_{\text{TTP}_s}\ x_{\text{T}_{\text{Max}}s}]$$

where x_{CBV_s} is represented as a vector:

$$x_{\text{CBV}_s} = \text{stack}(\text{CBV}^p_{\text{SLIC}}(i,j))_{\forall (i,j) \in \text{BT}^p}$$

The total matrix X_T is given by the combination of the two input features matrices depending on the model trained: $X_T = [X\ X_{\text{SLIC}}]$.

In the prediction phase, as shown in Fig. 2, the input features matrix X^p used for Step1 has 6 columns since the CBF and CBV parametric maps are excluded. Then, the model generates a binary map for the penumbra region over the entire image. Subsequently, the input feature matrix for the second step is derived only from CBV and CBF parametric maps plus their corresponding superpixel versions. This matrix has 4 columns as illustrated in Fig. 2. The selection of parametric maps is also in line with proposed methods in the literature [12]–[14] since TTP and T_{Max} are often used for detecting penumbra, while the other parametric maps are used for segmenting core regions.

To create the final prediction image, in the *Two-Step* approach, the binary predictions of core and penumbra are logically combined so the common white areas in both predictions indicate ischemic core in the final result. The logical AND combination simulate the medical constraint, where the ischemic core is limited to be inside the penumbra since the hypoperfused tissue always contains the dead tissue. For both the approaches (*Single-Step* and *Two-Step*), the patient's predictions pass through a 3D mode filter. This post-processing step helps to reduce unwanted noise and it also allows the predictions from a ML method to rely on the adjacent voxels in the z-axis, i.e. between adjacent slices.

VI. Experiments and Results

A. Dataset division

In this paper data from 152 patients were used, 137 from AIS patients divided into two groups (LVO and Non-LVO) and 15 patients WIS but who were admitted with suspicion of stroke. The dataset was randomly split into a training, validation, and holdout set, as described in Table IV, carefully dividing the LVO, Non-LVO, and WIS patients over the sets. The idea behind this division is to create a model that generalizes the classification of the ischemic regions working for all.

TABLE IV
DIVISION IN TRAINING, VALIDATION, AND HOLDOUT DATASET.

	Training (#; %)	Validation (#; %)	Holdout (#; %)	Tot. (#; %)
LVO	29; 37.7	29; 37.7	19; 24.6	77; 50.6
Non-LVO	24; 40	25; 41.7	11; 18.3	60; 30.5
WIS	6; 40	6; 40	3; 20	15; 9.8
Total	59; 38.8	60; 39.5	33; 21.7	152; 100

As many have reported, DWI is a questionable measure to describe the ischemic core [10], [15]–[18], thus we propose to use manual annotations made by two expert neuroradiologists as the golden ground truth to assess both the ischemic regions during early stages and with different level of severity.

Even with removing the background and only considering the pixels inside the BT, the core and penumbra classes are still undersampled, leading to a class imbalance problem in the dataset. To overcome this problem, during the training phase we implement the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [34] to over-sample the classes with a minor number of occurrences. SMOTE relies on the generation of synthetic examples on the difference between the feature vector under construction and its nearest neighbor. We over-sample the penumbra by a maximum of 5 times its standard amount and the core by a maximum of 20 times. These maximum values were chosen for their class importance and amounts. Before applying the SMOTE algorithm, the core and penumbra classes represent only 0.5% and 9.4% of the entire set respectively. After the application of the algorithm, they represent 7.6% and 36.5% of the dataset respectively.

B. Evaluation metrics

In all the experiments the predictions are compared with the ground truth and multi-class confusion matrices are generated. Our dataset is composed of three classes $\mathbf{C} \in \{\text{core}, \text{penumbra}, \text{healthy brain}\}$.

TABLE V
EXAMPLE OF MULTI-CLASS CONFUSION MATRIX FOR THE CORE CLASS. TP =TRUE POSITIVE, FP = FALSE POSITIVE, FN = FALSE NEGATIVE, AND TN = TRUE NEGATIVE.

		Predicted class		
		Core	**Penumbra**	**Healthy Brain**
Actual class	**Core**	TP_c	FN_c	FN_c
	Penumbra	FP_c	TN_c	TN_c
	Healthy Brain	FP_c	TN_c	TN_c

Table V presents a multi-class confusion matrix example for the core class: TP_c (True Positive) indicates the number of pixels predicted correctly as the core; FP_c (False Positive) represents the number of pixels classified as core class but belonging to a different class; FN_c (False Negative) is the number of pixels predicted as a different class but labeled as the core in a ground truth image; TN_c (True Negative) displays the number of pixels that are classified as not core and belonging to one of the other classes. All the values in each multi-class confusion matrix are calculated based only on the number of voxels inside the BT, excluding all non-brain tissue voxels as the binary mask of brain vs background is found during pre-processing. From each confusion matrix of class $c \in \mathbf{C}$, we calculate the recall $\text{rec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}$, the precision $\text{prec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}$, and the Dice coefficient (equivalent to the F1-score) $\text{Dice}_c = \frac{2 \cdot \text{prec}_c \cdot \text{red}_c}{\text{prec}_c + \text{rec}_c} = \frac{2 \cdot \text{TP}_c}{2 \cdot \text{TP}_c + \text{FP}_c + \text{FN}_c}$. The range for these values is $[0, 1]$. We also consider the Hausdorff distance between predictions and ground truth regions [35], and the absolute difference in the volume among the predictions (V_p [ml]) and the ground truth (V_g [ml]): $\Delta V = |V_g - V_p|$. The range value for the Hausdorff distance and ΔV is $[0, \infty]$ Bland-Altman plots were used to illustrate mean differences and limit of agreement between predicted volume and volume calculated from ground truth images.

C. Hyper-parameter optimization of ML algorithms

Before evaluating our methods, a series of hyper-parameter optimizations on the ML algorithms were performed using a Bayesian optimization. The input features for these optimizations, for a patient p, were solely based on X^p, without the usage of SLIC nor SMOTE algorithms. For DT and RF models, the hyper-parameters taken into consideration during the optimization were:

- the minimum number of leaf, with a range $[1, L^p/2]$,
- the maximum number of decision splits, in the range $[1, L^p - 1]$,
- Gini's diversity index, Twoing rule, and Cross-entropy for the split criterion to use,
- the number of decision trees in the model (1 for the DT algorithm, a range of $[1, 500]$ for the RF).

TABLE VI
OPTIMAL HYPER-PARAMETERS FOR THE DT AND RF ALGORITHMS DIVIDED BY *Single-Step* AND *Two-Step* APPROACHES.

Method			# DT	Split criterion	Min # Leaf	Max # Split
DT	*Single-Step*		1	Cross-entropy	138	22489
	Two-Step	Step1		Cross-entropy	153	358000
		Step2		Cross-entropy	10	34427
RF	*Single-Step*		4	Gini	345	5535
	Two-Step	Step1	10	Cross-entropy	384	1400500
		Step2	10	Gini	2	20979

Differently, for the SVM model, we considered the following:

- Gaussian, Linear, and Polynomial kernel functions,
- the maximum penalty on the observations with a range of $[0.001, 1000]$,
- standardized vs not standardized features.

The values display in Table VI show the best hyper-parameters for the DT and RF algorithms divided by *Single-Step* and *Two-Step* approaches, after an exhaustive set of experiments. Table VII presents the optimal hyper-parameters for the SVM model. All the experiments described in the next sections use the same set of hyper-parameters defined in Table VI and Table VII.

TABLE VII
OPTIMAL HYPER-PARAMETERS FOR THE SVM MODEL WITH THE *Two-Step* APPROACH.

Method			Kernel Function	Max penalty	Standardize
SVM	*Two-Step*	Step1	Gaussian	993.73	No
		Step2	Gaussian	0.487	No

D. Experiment 1 - ML algorithms and feature combination

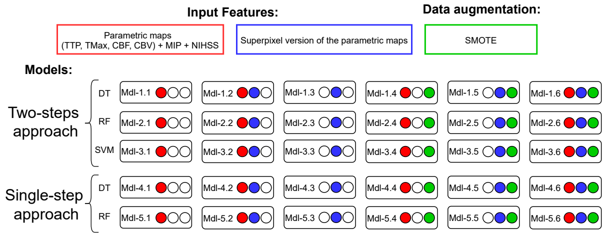


Fig. 4. Description of the models implemented to test the two approaches, the input features used (the parametric maps with or without the superpixel regions), the usage of data augmentation (SMOTE). Experiments' names are included in the reminder of the paper. DT = Decision Tree; RF = Random Forest; SVM = Support Vector Machine. Color in the online version.

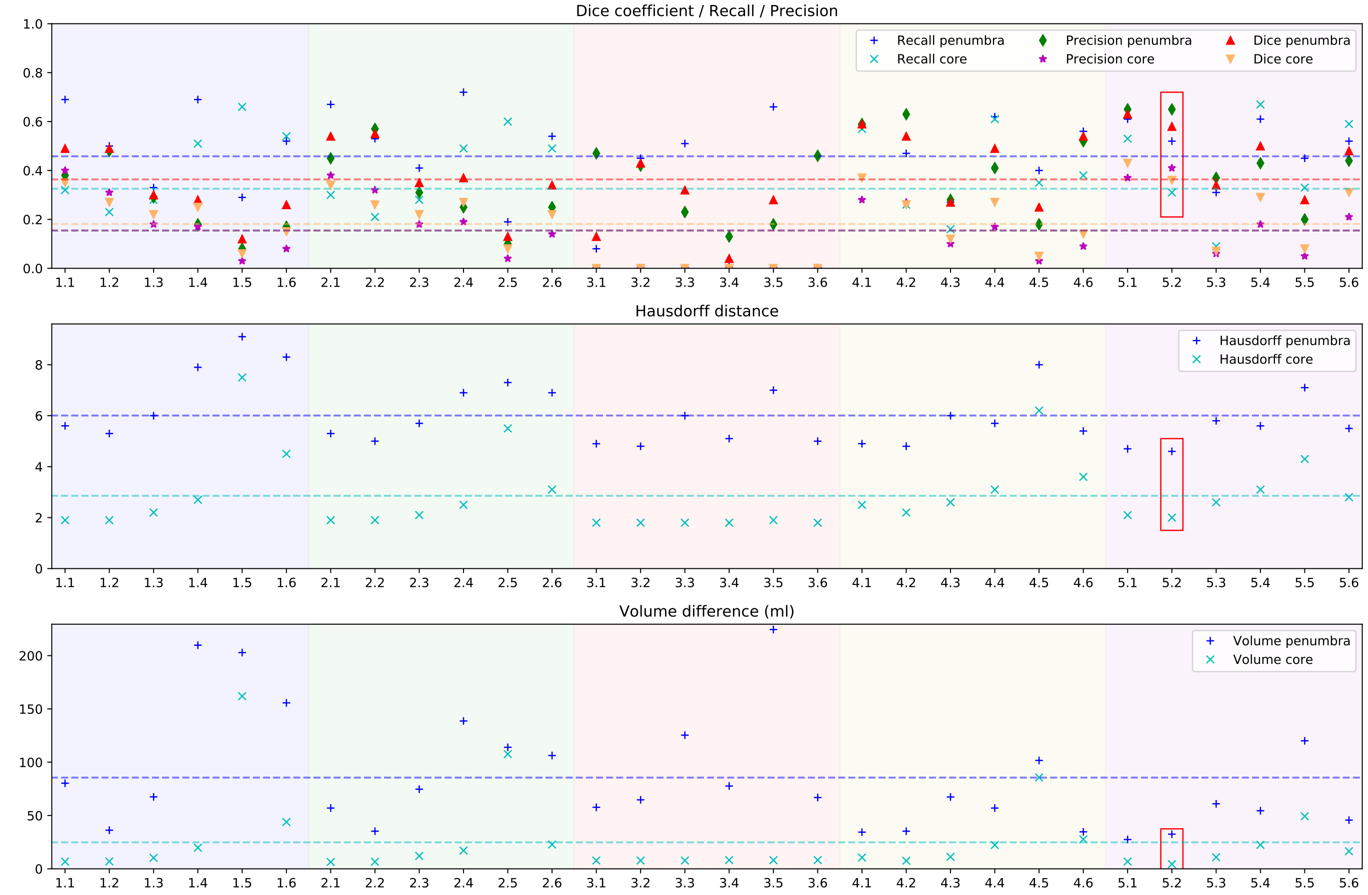


Fig. 5. Results generated with the validation set on the 30 experiments described in Fig. 4. The x-axis contains the experiment IDs, while the y-axis refers to the statistic values. Each value represents the average of the patients in the validation set, including all the different severities. Note that for the top subplot we want high values, but for the mid and bottom subplots we want low values. All the experiments were tested with a number of superpixel regions equal to 10. The colored regions in the plot represent the division of the various experiments: blue, green, and red contain the experiments with the two steps approach using DT, RF, and SVM models respectively; yellow and purple have the experiments for DT and RF with the *Single-Step* approach. The colored horizontal lines display the average for the corresponding statistical measures. With the only exception of *Mdl-5.1*, *Mdl-5.2* (inside a red rectangle) is the one that presents the best tradeoff for all the evaluation metrics among the set of experiments. Color in the online version.

For both the *Two-Step* and the *Single-Step* approaches, a series of six experiments were conducted to determine whether the inclusion of superpixels as extra features is beneficial and to see if using SMOTE to balance the classes during training gives better models. These six experiments were repeated for the different ML algorithms except SVM for *Single-Step* approach, due to our implementation of the approach which performs only binary classification.

Fig. 4 illustrates the 30 conducted experiments: (*Two-Step*×3 ML algorithms)×6 + (*Single-Step* ×2 ML algorithms)×6. The number of superpixel regions used for this set of experiments is 10. Fig. 5 shows the results for all models during the first experiment set taking into account all the various groups (LVO, Non-LVO, and WIS) together. The best model was selected mainly based on the averaging metrics in Fig. 5 for both the classes. Looking at Fig. 5, *Mdl-5.1* shows the best performances both for core and penumbra regardless of the group. Nevertheless, *Mdl-5.2* offers comparable results to *Mdl-5.1* in the majority of the metrics. Moreover, *Mdl-5.2* uses the superpixel regions as input features, on the contrary of Mdl-5.1, and the best number of superpixel regions should be investigated further.

E. Experiment 2 - Number of superpixels

After the first experiment set, we performed a series of empirical analyses on *Mdl-5.2* to choose the most adequate number of super-

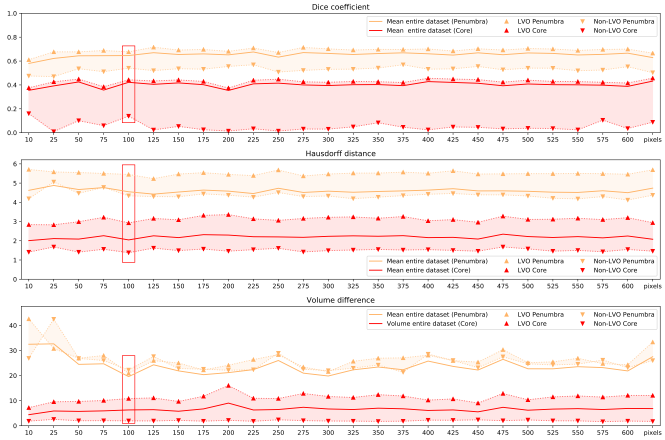


Fig. 6. Various plots (Dice coeff., Hausdorff dist., ΔV) achieved with the validation set for selecting the best number of superpixel regions for *Mdl-5.2*. The x-axis indicates the number of superpixel regions. Results achieved by the best-performed model are highlighted with a red rectangle. Solid lines represent the average of the patients in the validation set, including all the different severities. Color in the online version.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3097591, IEEE Journal of Biomedical and Health Informatics

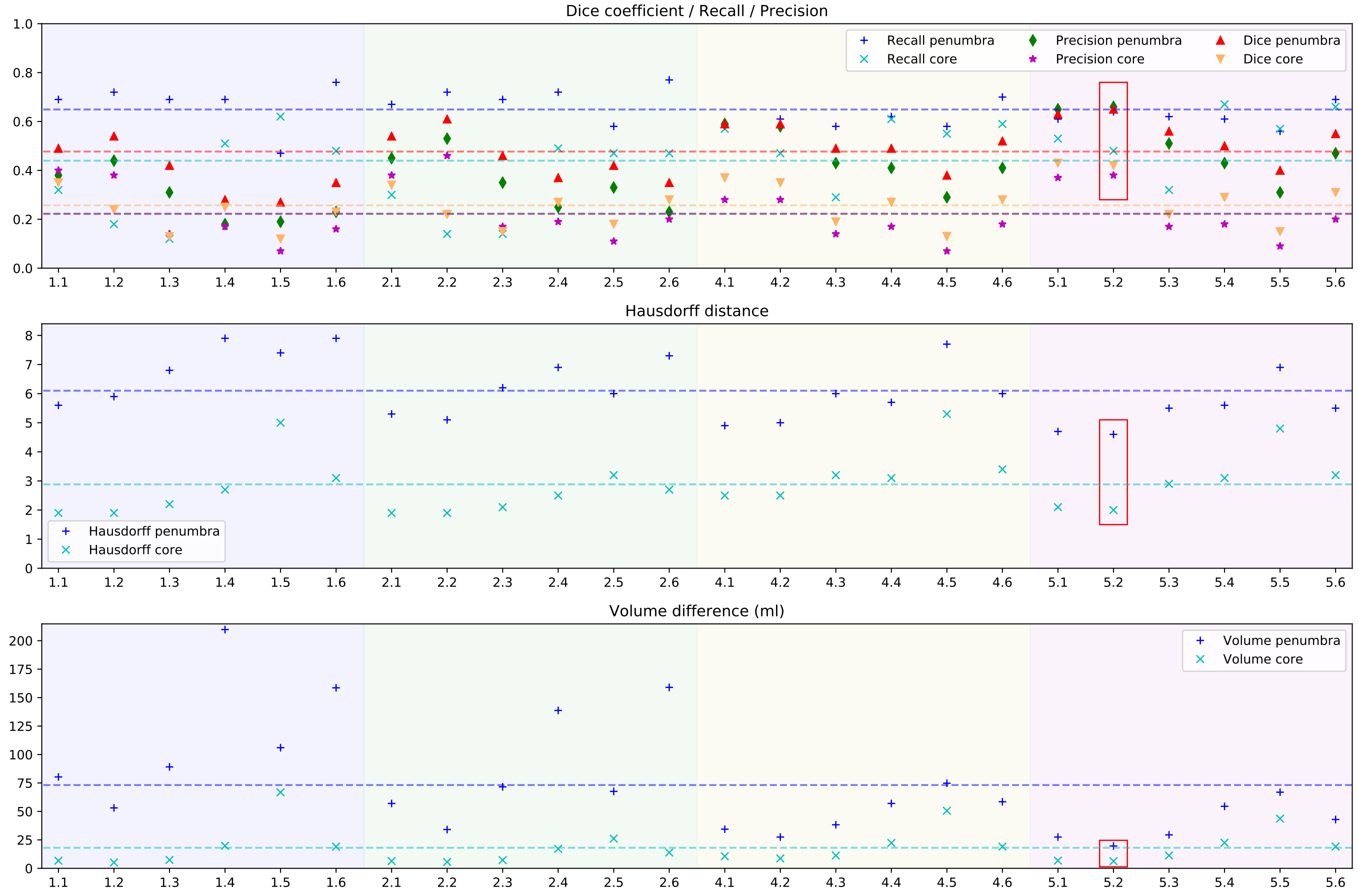


Fig. 7. Statistical measures to select the best input data combination to use. All the methods were tested with the best number of superpixel regions (100). The best model (*Mdl-5.2*) is highlighted inside a red rectangle. Color in the online version.

pixel regions for the SLIC algorithm that produces the best results. We repeat a series of experiments using the *Mdl-5.2* starting with 25 total number of 3D superpixel regions and continue by increasing the number until 600. The increment is 25 for each iteration. Fig. 6 presents the results obtained with different numbers of superpixel regions including 10 and also the total number of pixels in the image for *Mdl-5.2*. Fig. 6 shows the average metrics for the LVO and Non-LVO groups, and the average for the entire validation set (LVO, Non-LVO, and WIS). The combination of statistical metrics for both penumbra and core classes shows a clear difference when superpixel is used as shown in Fig. 6. As highlighted in the figure, 100 superpixel regions give slightly better results compared to the others. It is noticeable that 100 superpixel regions yield the lowest volume difference for the penumbra class, which highly influenced the selection decision, the highest Dice coefficient for the core class on average, and significant results for the other metrics, and as such we propose to use 100 in further experiments.

F. Experiment 3 - Validate the superpixel result

The chosen number of superpixels, 100, was validated by repeating all the thirty experiments described in Fig. 4 and using 100 superpixel regions instead of 10 regions, which was used during the first evaluation round. Due to ineffective performance, SVM has been exempt from this validation step. Results are depicted in Fig. 7 showing the overall metrics for both the two ischemic regions. The model *Mdl-5.2* still shows the most promising results even compared with *Mdl-5.1*. It achieves the highest Dice coefficient and precision values for both the classes, and excellend ΔV results.

G. Hyper-parameter optimization on the best model

The selected *Mdl-5.2* model went under a final step of performing optimization of its hyper-parameters with the current setting (100 superpixel regions) validated in the previous experiment set. We have taken into consideration the same hyper-parameters defined in Sec. VI-C for RF. The new optimal hyper-parameters for *Mdl-5.2* are 48 number of DT, cross-entropy as the selected split criterion, 68982 as the maximum number of decision splits, and 315 as the minimum number of leaves.

H. Final test of the best model

We test the holdout set proposing *Mdl-5.2* as the best model, 100 as the most efficient number of superpixel regions, with the hyper-parameters defined in Sec. VI-G. A visual result of two sample predicted images along with ground truth and their corresponding parametric maps are shown in Fig. 8.

Furthermore, we remove the post-processing step (3D mode filter) and predict the regions to understand how the results are influenced by this step. Table VIII presents the results of the proposed best model, i.e. the *Single-Step* method with RF, *Mdl-5.2*, and 100 as the number of superpixels, in comparison with the same model without any post-processing step, the “syngo.via” default setting to define the ischemic regions, and the thresholding values proposed by Bathla et al. [28], since it is, to the best of our knowledge, the only research using “syngo.via” as vendor. Table VIII also depicts reported results from other thresholding methods ([11]–[13], [22]) which used other vendors for parametric maps acquisition and post-processing steps, thus a direct comparison with our model is not possible. Bland-Altman plots are used to visualize the predicted volume in comparison with the ground truth volume between the four methods compared in Table VIII, shown in Fig. 9. For all rows, the statistical results are based solely on our holdout set to establish fair comparability with the other approaches. The results for two subsets of the data (LVO, Non-LVO) are presented separately, while for the WIS subset only ΔV is displayed.

Inter-observer variability

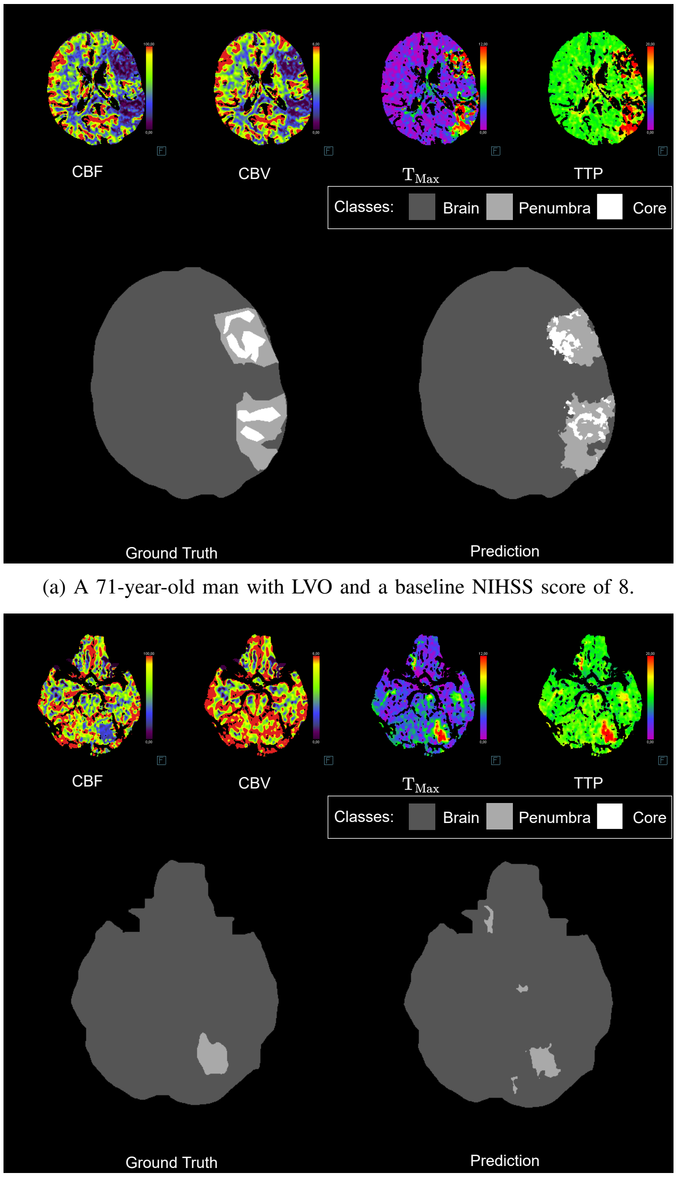
33 randomly selected patients (19 from the LVO subset, 11 from the Non-LVO subset, and 3 from the WIS group) were manually annotated by two different neuroradiologists, using the same criteria adopted for the creation of the ground truth images. The aim is to understand the inter-observer variability between two neuroradiologists. We investigate the inter-observer variability and compare it with the metrics of the automated method. Table VIII shows the inter-observer variability in the measurements of the ischemic regions for the two subsets of the data, LVO and Non-LVO, in comparison with the results achieved with our best method *Mdl-5.2*.

VII. Discussion

We have proposed a multi-stage algorithm based on ML that automatically classifies ischemic core and penumbra regions in parametric maps generated from CTP images. The CTP scans were acquired from patients with AIS and WIS. In a real-life situation, medical doctors need to decide the treatment for a patient in a small time window; thus, an automatic approach can be valuable. Expert assessments used as ground truth are commonly implemented in clinical use in many applications. We consider it to be a good method to interpret the ischemic regions, due to the lack of consensus on thresholding methods and the recent oppositions over the de-facto DWI as the gold standard [15]–[18]. Nevertheless, these assessments present some variability among the experts (Table VIII), thus an automatic approach might present some advantages during analysis and can aid medical doctors in rapid recognition of ischemic regions. We have trained our method with ground truth images directly acquired from the CTP parametric maps, MIP, and follow-up images. This results in better and more precise visualization of the two ischemic regions in the brain: the salvageable (penumbra) and the irreversibly damaged tissue (infarct core). Fast and correct visualization of the penumbra will guide the treatment better since it is fundamental to treat patients where relevant tissue can be saved, and not invest a lot of resources and time in trying to save tissue that is already irreversibly damaged and where the treatment might even harm the patient due to the risk of hemorrhage.

The criteria to select the best method was based on a study of various implemented experiments and their relative statistical results. First, we performed a set of thirty experiments described in Sec. VI-D and in Fig. 4, to select the right features and model. From the relative outcomes in Fig. 5, the results provided by *Mdl-5.2* (RF with *Single-Step* approach using all parametric maps at once) produces considerable statistical measures in the majority of the metrics, regardless of the severity group or class. It is interesting to notice that the *Single-Step* approach generates better results or all metrics but the *Two-Step* approach with RF producs slightly better results in the Hausdorff distance for the core class. Results for irreversibly damaged tissue for SVM models were not taken into consideration since these models fail to predict the mentioned class.

Subsequently, we applied a different number of superpixel regions to *Mdl-5.2* to find one that gives the best prediction results (Sec. VI-E, and in Fig. 6). It is clear that the results are not the best without applying superpixel, however, there is not a clear difference between different numbers of superpixel regions; Dice coefficient and Hausdorff distance outcomes do not present large discrepancies during the increment of superpixel regions; the metric influencing the final decision was the volume difference due to its drastic drop for the selected number of superpixels for the penumbra class and a significantly low value for the core class. Another important factor that helped to select the best number of superpixel regions was how the performances of the models differ with the various stroke severity groups. From Fig. 6 it is clear to notice that, among all the experiments in this set, *Mdl-5.2* presented the best tradeoff between the difference in volume and Dice coefficient for both the classes. One can argue that 125 superpixel regions give more or less similar results as 100 regions, however, ΔV is higher especially for



(a) A 71-year-old man with LVO and a baseline NIHSS score of 8.

(b) A 60-year-old woman with Non-LVO and a baseline NIHSS score of 1.

Fig. 8. Visual comparison with four parametric maps (top), ground truth images (left), and the corresponding predicted image with the best method (right) of one slice for two patients included in the testset, one labeled as LVO (a), the other as Non-LVO (b). The dark grey area is healthy brain tissue, the light gray area represents the penumbra, and the white region indicates the ischemic core. Color in the online version.

TABLE VIII

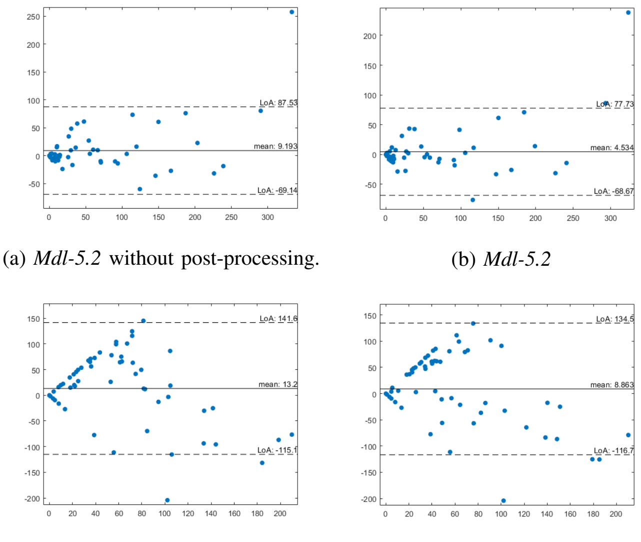
PATIENTS INCLUDED IN THIS TABLE ARE ALL PART OF THE HOLDOUT SET. THE RESULTS ARE PRESENTED FOR PENUMBRA (CORE) REGIONS. COMPARISON BETWEEN VARIOUS RESEARCHES USING THRESHOLDING VALUES WITH THE SAME VENDOR "SYNGO.VIA" (DEFAULT SETTING AND [28]) AND OUR BEST MODEL (*Mdl-5.2*). PREDICTIONS FROM [11]–[13], [22] ARE PRESENTED BUT THEY ARE NOT FOR COMPARISON DUE TO THE USAGE OF DIFFERENT VENDOR AND/OR POST-PROCESSING STEPS FOR GENERATING PARAMETRIC MAPS. ‡ MARKS THE RESULTS FOR THE *Mdl-5.2* METHOD WITHOUT USING ANY POST-PROCESSING STEP. INTER-OBSERVER VARIABILITY FOR TWO EXPERT NEURORADIOLOGISTS (NR_1, NR_2) AND THE SELECTED MODEL *Mdl-5.2* IS ALSO PRESENTED. NOTE THAT FOR THE DICE COEFFICIENT HIGHER VALUES ARE BETTER (⇑), WHILE FOR HAUSDORFF DISTANCE AND ΔV LOWER VALUES ARE PREFERABLE (⇓).

Method	Vendor	Dice Coefficient ⇑			Hausdorff Distance ⇓			ΔV (ml) ⇓			
		LVO	Non-LVO	All	LVO	Non-LVO	All	LVO	Non-LVO	WIS	All
		Penumbra (Core)									
Best Method (*Mdl-5.2*) ‡	Siemens "syngo.via"	0.66 (**0.26**)	0.51 (0.03)	0.66 (0.26)	6.9 (4.8)	3.5 (0.9)	5.2 (3.1)	44.2 (16.2)	6.9 (**0.8**)	1.4 (**0.0**)	27.9 (9.6)
Best Method (*Mdl-5.2*)		**0.69** (**0.27**)	**0.56** (0.03)	**0.68** (**0.26**)	**6.5** (**4.3**)	**3.0** (**0.7**)	**4.8** (**2.7**)	**40.7** (**12.9**)	**4.9** (1.0)	**0.9** (**0.0**)	**25.1** (**7.8**)
Default Setting		0.31 (0.25)	0.11 (**0.04**)	0.27 (0.20)	7.8 (6.2)	5.6 (4.4)	6.6 (5.2)	67.5 (48.2)	51.8 (37.4)	3.7 (12.1)	58.2 (40.8)
Bathla et al. [28]		0.47 (0.17)	0.22 (0.03)	0.45 (0.14)	6.9 (6.9)	4.5 (4.7)	5.6 (5.7)	65.2 (65.3)	16.5 (44.5)	22.5 (6.6)	43.3 (53.5)
Other thresholding methods presented but not used for comparison											
Bivard et al. [13]	Toshiba	0.42 (0.19)	0.16 (0.03)	0.39 (0.15)	7.3 (6.5)	4.6 (4.4)	5.8 (5.4)	70.6 (52.9)	30.2 (36.0)	1.5 (9.1)	50.8 (43.3)
Cambell et al. [12]	Philips	N.A. (0.22)	N.A. (0.04)	N.A. (0.18)	N.A. (5.9)	N.A. (3.9)	N.A. (4.9)	N.A. (35.2)	N.A. (24.9)	N.A. (5.6)	N.A. (29.1)
Murphy et al. [22]	General Electric	0.17 (0.27)	0.08 (0.05)	0.16 (0.23)	7.5 (5.0)	4.8 (3.1)	6.1 (4.0)	96.7 (13.4)	21.1 (13.3)	8.6 (2.1)	63.5 (12.3)
Wintermark et al. [11]	Philips	N.A. (0.19)	N.A. (0.02)	N.A. (0.14)	N.A. (7.5)	N.A. (5.5)	N.A. (6.4)	N.A. (90.8)	N.A. (71.4)	N.A. (20.1)	N.A. (77.9)
Inter-observer variability											
NR_1 vs NR_2	Siemens "syngo.via"	0.80 (0.55)	0.67 (0.33)	0.79 (0.54)	5.1 (3.2)	1.9 (0.5)	3.6 (2.0)	33.3 (5.6)	5.5 (0.7)	0.0 (0.0)	21.0 (3.5)
Mdl-5.2 vs NR_1		0.69 (0.25)	0.51 (0.01)	0.68 (0.25)	6.6 (4.2)	3.3 (0.5)	5.0 (2.6)	53.6 (12.4)	8.7 (0.3)	0.9 (0.0)	33.8 (7.2)
Mdl-5.2 vs NR_2		0.71 (0.30)	0.56 (0.03)	0.70 (0.30)	6.4 (4.2)	3.0 (0.7)	4.8 (2.9)	38.6 (12.2)	4.9 (1.0)	0.9 (0.0)	23.9 (7.3)

penumbra regions, meaning that 125 superpixel regions provide an overestimation of the tissue at risk, especially for the LVO group.

Finally, we validated the selected superpixel number by applying it to the other experiments (Sec. VI-F). SVM was excluded from this step as it performed poorly from the beginning (reference to Fig. 5). As shown in Fig. 7, increasing the number of superpixel regions slightly improved the statistical measures for both classes. Moreover, results achieved by *Mdl-5.2* present higher precision and lower ΔV in comparison with the other models. The proposed method can classify correctly both penumbra and core in patients affected by a large vessel occlusion. The differences between the healthy and the ischemic tissue are more noticeable, in contrast with ischemic regions in patients with Non-LVO; an example is given in Fig. 8 for two brain slices of two patients affected by LVO (Fig. 8 (a)) and Non-LVO (Fig. 8 (b)). From the examples in Fig. 8 and the results in Table VIII, our best method is shown to predict penumbra regions more precisely than core areas. In patients with LVO, the prediction of core regions achieved promising results. However, the detection of core regions in patients with Non-LVO is more challenging; the small core area can be difficult to classify correctly. This issue might be related to the limited number of samples for that particular class, since patients in the Non-LVO group does not always have a core region. We compared the performance of the proposed RF-based method with approaches based on thresholding suggested in the literature and the results are presented in Table VIII; comparison is only performed with the default setting and values from Bathla et al. [28] due to the usage of the same vendor. Predictions from the other methods [11]–[13], [22] are just presented for visualization purposes; a comparison does not apply to the utilization of different vendors to generate the parametric maps, but it illustrates an important limitation of thresholding.

The proposed method (*Mdl-5.2*) performs better than the thresholding approaches concerning the evaluation metrics. The use of a post-processing step slightly increment the performances of the best method, as it is possible to evince from Table VIII and Fig. 9. The *Mdl-5.2* method (using a 3D mode filter as a post-processing step) achieved the highest metrics for all the classes regardless of the stroke severity level. The sole exception where the model does not perform well is with the core class for Non-LVO group since, as it is possible to evince from Table VIII, it is the hardest class to predict correctly due to its limited number of samples and its narrow size in the BT.



(a) *Mdl-5.2* without post-processing.

(b) *Mdl-5.2*

(c) "syngo.via" default setting

(d) Bathla et al. [28]

Fig. 9. Bland-Altman plots of the volume calculated between the predictions and the ground truth images for model *Mdl-5.2* with (b), without (a) post-processing step, the "syngo.via" default setting (c), and the values presented by Bathla et al. [28] (d). Color in the online version.

Core predictions are slightly better than the one presented by the thresholding methods regardless of the group, while penumbra predictions are superior. This indicates a reliable understanding and agreement among ML predictions, threshold values, and neuroradiologists' annotations for the core regions. While, at the same time, it presents some uncertainties regarding the penumbra's definition. This might be related to the fact that the infarct core and penumbra are two dynamic regions inside the brain and highly dependent on the acquisition time of CTP and DWI. The perfusion examination shows the perfusion at that specific time, the penumbra and core size may change rapidly. In many studies, MRI is not performed immediately after CTP. DWI, often used as the gold standard for defining the ischemic core, cannot define penumbra. Our method, relying the ground truth on both CTP generated right after hospital admission (parametric maps derived from CTP, and MIP) and follow-up images, seems to provide a reliable method to predict both penumbra and core. Note that we propose to make predictions only based on data available right after hospital admission. The areas defined as ground truth from the DWI sequence can over- or underestimate the ischemic core in individual patients, making it unrealistic to expect perfect concordance between ischemic core measurements on CTP and DWI [10], [16]–[18]. Other reasons are: first, they are not taking into consideration any spatial characteristics of an image; second, the values are very sensitive to image artifacts. Third, patients with contraindication to MRI, i.e. heart pacemaker, metal foreign body, might be excluded from studies where MRI and DWI images are used. Moreover, it is complicated to find an optimal threshold value for any group of patients. All the methods rely on selected thresholds, which might produce good results for a particular and predefined group, but it might not be the best for a single case study or the entire dataset studied. Their validation method relies on the comparison of the thresholding values with the clinical outcome of the patient; however, this is not perfect as the patient might have received treatment or the symptoms might have changed. Nevertheless, the delineation of the core should not be smaller due to treatment if the model delineates the core region correctly.

Table VIII shows the inter-observer variability in thirty patients divided by stroke severity into LVO and the Non-LVO subsets. There is a discrepancy between the results for the LVO and the Non-LVO subsets. Results for the LVO group have some similarities between the manual annotations and the *Mdl-5.2*. Nevertheless, manual annotations present better results in all the statistic measurements in comparison with the *Mdl-5.2* method in the Non-LVO subset. However, results in Table VIII illustrate the difficulties of achieving a consensus even among neuroradiologists.

VIII. Conclusion

We proposed an automatic multi-classification approach for segmenting both ischemic core and penumbra based on random forest using the parametric maps as input features, *Mdl-5.2*. We implemented other approaches based on thresholding, proposed in the literature, and compared them with our proposed method considering manual annotations as the ground truth generated from parametric maps. The method was trained with patients, both with AIS and WIS, grouped by different stroke severities. It shows good results for patients with large vessel occlusions, but not very good for patients with non-large vessel occlusions. Our method generates more precise results than the thresholding approaches for the two regions, but there is still room for improvement. We achieve an average Dice coefficient of 0.68 and 0.26, respectively for penumbra and core, for the three groups analyzed. We also achieve an average in volume difference of 25.1ml for penumbra and 7.8ml for core. Detecting ischemic core and penumbra regions in patients with non-large vessel occlusion can be very complicated, as shown in Fig. 8. Therefore, in the future, we plan to use approaches based on deep neural networks with 4D CTP volume as input instead of the parametric maps to work with the original acquired data.

List of Abbreviations

BT Brain Tissue.
CBF Cerebral blood flow.
CBV Cerebral blood volume.
CNN Convolutional Neural Network.
CT Computed Tomography.
CTA Computed Tomography Angiography.
CTP Computed Tomography Perfusion.
DT Decision Tree.
DWI Diffusion-weighted Imaging.
LVO Large Vessel Occlusion.
MIP Maximum Intensity Projection.
ML Machine Learning.
MRI Magnetic Resonance Imaging.
MTT Mean transfer time.
NCCT Non-contrast Computed Tomography.
NIHSS National Institutes of Health Stroke Scale.
RF Random Forest.
SLIC Simple Linear Iterative Clustering.
SMOTE Synthetic Minority Over-sampling Technique.
SVM Support Vector Machine.
Non-LVO Non-Large Vessel Occlusion.
T_{Max} Time-to-maximum.
TTP Time-to-peak.
WIS Without Ischemic Stroke.

References

[1] V. L. Feigin, B. Norrving, and G. A. Mensah, “Global burden of stroke,” *Circulation research*, vol. 120, no. 3, pp. 439–448, 2017.

[2] E Stevens, E Emmett, Y Wang, C McKevitt, and C Wolfe, “The burden of stroke in europe, report,” *King’s College London for The Stroke Alliance for Europe (SAFE)*, 2017.

[3] B. Indredavik, R Salvesen, H Næss, and D Thorsvik, “Nasjonal retningslinje for behandling og rehabilitering ved hjerneslag,” *Helsedirektoratet: Oslo*, 2010.

[4] W. Johnson, O. Onuma, M. Owolabi, and S. Sachdev, “Stroke: A global response is needed,” *Bulletin of the World Health Organization*, vol. 94, no. 9, p. 634, 2016.

[5] E. J. Benjamin, S. S. Virani, C. W. Callaway, A. M. Chamberlain, A. R. Chang, S. Cheng, S. E. Chiuve, M. Cushman, F. N. Delling, R. Deo, *et al.*, “Heart disease and stroke statistics—2018 update: A report from the american heart association,” *Circulation*, 2018.

[6] 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*, vol. 8, no. 1, p. 34, 2017.

[7] K. Kurz, G Ringstad, A Odland, R Advani, E Farbu, and M. Kurz, “Radiological imaging in acute ischaemic stroke,” *European journal of neurology*, vol. 23, pp. 8–17, 2016.

[8] J. Astrup, B. K. Siesjö, and L. Symon, “Thresholds in cerebral ischemia-the ischemic penumbra.,” *Stroke*, vol. 12, no. 6, pp. 723–725, 1981.

[9] E. S. O. E. E. Committee, E. W. Committee, *et al.*, “Guidelines for management of ischaemic stroke and transient ischaemic attack 2008,” *Cerebrovascular diseases*, vol. 25, no. 5, pp. 457–507, 2008.

[10] C. W. Cereda, S. Christensen, B. C. Campbell, N. K. Mishra, M. Mlynash, C. Levi, M. Straka, M. Wintermark, R. Bammer, G. W. Albers, *et al.*, “A benchmarking tool to evaluate computer tomography perfusion infarct core predictions against a dwi standard,” *Journal of Cerebral Blood Flow & Metabolism*, vol. 36, no. 10, pp. 1780–1789, 2016.

[11] M. Wintermark, A. E. Flanders, B. Velthuis, R. Meuli, M. Van Leeuwen, D. Goldsher, C. Pineda, J. Serena, I. v. d. Schaaf, A. Waaijer, *et al.*, “Perfusion-ct assessment of infarct core and penumbra: Receiver operating characteristic curve analysis in 130 patients suspected of acute hemispheric stroke,” *Stroke*, vol. 37, no. 4, pp. 979–985, 2006.

[12] B. C. Campbell, S. Christensen, C. R. Levi, P. M. Desmond, G. A. Donnan, S. M. Davis, and M. W. Parsons, "Comparison of computed tomography perfusion and magnetic resonance imaging perfusion-diffusion mismatch in ischemic stroke," *Stroke*, vol. 43, no. 10, pp. 2648–2653, 2012.

[13] A Bivard, C Levi, V Krishnamurthy, J Hislop-Jambrich, P Salazar, B Jackson, S Davis, and M Parsons, "Defining acute ischemic stroke tissue pathophysiology with whole brain ct perfusion," *Journal of neuroradiology*, vol. 41, no. 5, pp. 307–315, 2014.

[14] P. W. Schaefer, L. Souza, S. Kamalian, J. A. Hirsch, A. J. Yoo, S. Kamalian, R. G. Gonzalez, and M. H. Lev, "Limited reliability of computed tomographic perfusion acute infarct volume measurements compared with diffusion-weighted imaging in anterior circulation stroke," *Stroke*, vol. 46, no. 2, pp. 419–424, 2015.

[15] P. Schellinger, R. Bryan, L. Caplan, J. Detre, R. Edelman, C Jaigobin, C. Kidwell, J. Mohr, M Sloan, A. Sorensen, *et al.*, "Evidence-based guideline: The role of diffusion and perfusion mri for the diagnosis of acute ischemic stroke: Report of the therapeutics and technology assessment subcommittee of the american academy of neurology," *Neurology*, vol. 75, no. 2, pp. 177–185, 2010.

[16] M. Goyal, J. M. Ospel, B. Menon, M. Almekhlafi, M. Jayaraman, J. Fiehler, M. Psychogios, R. Chapot, A. Van Der Lugt, J. Liu, *et al.*, "Challenging the ischemic core concept in acute ischemic stroke imaging," *Stroke*, vol. 51, no. 10, pp. 3147–3155, 2020.

[17] C. S. Kidwell, J. L. Saver, J. Mattiello, S. Starkman, F. Vinuela, G. Duckwiler, Y. P. Gobin, R. Jahan, P. Vespa, M. Kalafut, *et al.*, "Thrombolytic reversal of acute human cerebral ischemic injury shown by diffusion/perfusion magnetic resonance imaging," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 47, no. 4, pp. 462–469, 2000.

[18] M.-A. Labeyrie, G. Turc, A. Hess, P. Hervo, J.-L. Mas, J.-F. Meder, J.-C. Baron, E. Touzé, and C. Oppenheim, "Diffusion lesion reversal after thrombolysis: A mr correlate of early neurological improvement," *Stroke*, vol. 43, no. 11, pp. 2986–2991, 2012.

[19] M. Matesin, S. Loncaric, and D. Petravic, "A rule-based approach to stroke lesion analysis from ct brain images," in *ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces (IEEE Cat.*, IEEE, 2001, pp. 219–223.

[20] P. Dastidar, T. Heinonen, J.-P. Ahonen, M. Jehkonen, and G. Molnár, "Volumetric measurements of right cerebral hemisphere infarction: Use of a semiautomatic mri segmentation technique," *Computers in biology and medicine*, vol. 30, no. 1, pp. 41–54, 2000.

[21] L. Contin, C. Beer, M. Bynevelt, H Wittsack, and G Garrido, "Semi-automatic segmentation of core and penumbra regions in acute ischemic stroke: Preliminary results," in *IWSSIP International Conference*, 2010.

[22] 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*, vol. 37, no. 7, pp. 1771–1777, 2006.

[23] A. Kemmling, F. Flottmann, N. D. Forkert, J. Minnerup, W. Heindel, G. Thomalla, B. Eckert, M. Knauth, M. Psychogios, S. Langner, *et al.*, "Multivariate dynamic prediction of ischemic infarction and tissue salvage as a function of time and degree of recanalization," *Journal of Cerebral Blood Flow & Metabolism*, vol. 35, no. 9, pp. 1397–1405, 2015.

[24] W. Qiu, H. Kuang, E. Teleg, J. M. Ospel, S. I. Sohn, M. Almekhlafi, M. Goyal, M. D. Hill, A. M. Demchuk, and B. K. Menon, "Machine learning for detecting early infarction in acute stroke with non–contrast-enhanced ct," *Radiology*, vol. 294, no. 3, pp. 638–644, 2020.

[25] W. Qiu, H. Kuang, J. M. Ospel, M. D. Hill, A. M. Demchuk, M. Goyal, and B. K. Menon, "Automated prediction of ischemic brain tissue fate from MultiPhase CT-Angiography in patients with acute ischemic stroke using machine learning," *medRxiv*, 2020. DOI: 10.1101/2020.05.14.20101014.

[26] 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*, vol. 50, no. 6, pp. 1578–1581, 2019.

[27] 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*, 2020, pp. 1–8.

[28] G. Bathla, K. Limaye, B. Policeni, E. Klotz, M. Juergens, and C. Derdeyn, "Achieving comparable perfusion results across vendors. the next step in standardizing stroke care: A technical report," *Journal of neurointerventional surgery*, vol. 11, no. 12, pp. 1257–1260, 2019.

[29] M. Fatima, M. Pasha, *et al.*, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, p. 1, 2017.

[30] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[31] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[32] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[35] T Birsan and D. Tiba, "One hundred years since the introduction of the set distance by dimitrie pompeiu," in *IFIP Conference on System Modeling and Optimization*, Springer, 2005, pp. 35–39.