ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI

Oskar Maier[a,1,*], Björn Menze[f], Matthias Liebrandt[c], Stefan Winzeck[f], Levin Häni[e], Mattias Heinrich[a], Janina von der Gablentz[c], Abdul Basit[m], Paul Bentley[h], Liang Chen[g,h], Daan Christiaens[q,s], Francis Dutil[v], Karl Egger[j], Chaolu Feng[k], Ben Glocker[g], Michael Götz[p], Tom Haeck[q,s], Hanna-Leena Halme[n,o], Mohammad Havaei[v], Khan M Iftekharuddin[t], Pierre-Marc Jodoin[v], Konstantinos Kamnitsas[g], Elias Kellner[i], Antti Korvenoja[n], Hugo Larochelle[v], Christian Ledig[g], Jia-Hong Lee[u], Frederik Maes[q,s], Qaiser Mahmood[l,m], Klaus H Maier-Hein[p], Richard McKinley[x], Chris Pal[w], Linmin Pei[t], Janaki R Rangarajan[q,s], Syed M S Reza[t], David Robben[q,s], Daniel Rueckert[g], Eero Salli[n], Paul Suetens[q,r,s], Ching-Wei Wang[u], Matthias Wilms[a], Peter Schramm[d], Thomas F Münte[c], Ulrike M Krämer[c], Mauricio Reyes[e], Heinz Handels[a]

[a] *Institut for Medical Informatics, University of Lübeck, Lübeck, Germany*
[b] *Graduate School for Computing in Medicine and Live Science, University of Lübeck, Germany*
[c] *Department of Neurology, University of Lübeck, Germany*
[d] *Institute of Neuroradiology, University Medical Center Lübeck*
[e] *Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland*
[f] *Institute for Advanced Study and Department of Computer Science, Technische Universität München, Munich, Germany*
[g] *Biomedical Image Analysis Group, Department of Computing, Imperial College London, UK*
[h] *Division of Brain Sciences, Department of Medicine, Imperial College London, UK*
[i] *Department of Radiology, Medical Physics, University Medical Center Freiburg, Germany*
[j] *Department of Neuroradiology, University Medical Center Freiburg, Germany*
[k] *College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, China*
[l] *Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden*
[m] *Pakistan Institute of Nuclear Science and Technology, Islamabad, Pakistan*
[n] *HUS Medical Imaging Center, Radiology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*
[o] *Department of Neuroscience and Biomedical Engineering NBE, Aalto University School of Science, Aalto, Finland*
[p] *Junior Group Medical Image Computing, German Cancer Research Center, Heidelberg, Germany*
[q] *ESAT/PSI, Department of Electrical Engineering, KU Leuven, Belgium*
[r] *iMinds, Medical IT Department, KU Leuven, Belgium*
[s] *Medical Imaging Research Center, UZ Leuven, Belgium*
[t] *Vision Lab, Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, USA*
[u] *Graduate Institute of Biomedical Engineering, National Taiwan University of Science and Technology, Taipei City, Taiwan*
[v] *Université de Sherbrooke, Sherbrooke, Qc, Canada*
[w] *Ecole Polytechnique de Montréal, Canada*
[x] *Department of Diagnostic and Interventional Neuroradiology, Inselspital Bern, Switzerland*

Abstract

Ischemic stroke is the most common cerebrovascular disease and its diagnosis, treatment and study relies on non-invasive imaging. Algorithms for stroke lesion segmentation from magnetic resonance imaging (MRI) volumes are intensely researched, but the reported results are largely incomparable due to different datasets and evaluation schemes. We approach this urgent problem of comparability with the Ischemic Stroke Lesion Segmentation (ISLES) challenge organized in conjunction with the MICCAI 2015 conference. In this paper we propose a common evaluation framework, describe the publicly available datasets and present the results of the two sub-taks: Sub-Acute Stroke Lesion Segmentation (SISS) and Stroke Perfusion Estimation (SPES). A total of 16 research groups participated with a wide range of state-of-the-art automatic segmentation algorithms and a thorough analysis of the obtained data allows us to determine the current state, to derive recommendations and to identify the remaining challenges. The segmentation of acute perfusion lesions addressed in SPES was found to be feasible, but SISS's sub-acute lesion segmentation methods still lack accuracy. Overall, no type of method was found to perform superior to the others. The annotated ISLES image datasets continue to be publicly available through an online evaluation system to serve as an ongoing benchmarking resource (`www.isles-challenge.org`).

Keywords: ischemic stroke, segmentation, MRI, challenge, benchmark, comparison

1. Introduction

Ischemic stroke is the most common cerebrovascular disease and one of the most common causes of death and disability worldwide (WHO, 2012). In ischemic stroke an obstruction

*Corresponding author(s)
Email address: `maier@imi.uni-luebeck.de` (Oskar Maier)
URL: `isles-challenge.org` (Oskar Maier)

of the cerebral blood supply causes underperfusion and, within minutes up to hours, tissue death. The affected area of the brain, the stroke lesion, undergoes a number of diseases stages that can be subdivided into *acute* (0-24h), *sub-acute* (24h-2w) and *chronic* (>2w) according to the time passed since stroke onset (González et al., 2011). Magnetic resonance imaging (MRI) of the brain is often used to assess a stroke lesion's presence, location, extend, age and other factor, as this modality is highly sensitive for many of the critical tissue changes observed in stroke.

Time is brain is the watchword of stroke units worldwide. Possible treatment options are largely restricted to reperfusion therapies (thrombolysis, thromboectomy), which have to be administered not later than four to six hours after the onset of symptoms. Unfortunately, these interventions are associated with considerable risk of bleeding if the lesion was underperfused for a too long time. To this end, considerable effort has gone into finding image descriptors that predict stroke outcome (Wheeler et al., 2013) or the response to treatment (Albers et al., 2006; Lansberg et al., 2012), and to identify those patients that would benefit from a treatment even beyond the regular treatment window (Kemmling et al., 2015).

At present, lesions are only evaluated in a qualitative fashion in the clinical setting. When used in clinical research studies, the manual segmentation of the lesion remains a tedious and time consuming task, taking up to 15 minutes per case (Martel et al., 1999), and with low inter-rater agreement (Neumann et al., 2009). Developing automated methods that locate, segment and quantify the stroke lesion area from MRI scans remains an open task. Such image processing algorithms would be expected to have a broad impact by supporting the clinicians' decisions and render their predictions more robust and reproducible.

Still, segmenting stroke lesions from MRI images poses a challenging problem. First, the stroke lesions' appearance varies significantly over time, not only between, but even within the clinical phases of stroke development. This holds especially true for the sub-acute phase: At the beginning of this interval, the lesion usually shows strongly hyperintense in the diffusion weighted imaging (DWI) sequence and moderately hyperintense in fluid attenuation inversion recovery (FLAIR). Towards the second week, the hyperintensity in the FLAIR sequence increases, while the DWI appearance converges towards isointensity (González et al., 2011). Additionally, a ring of edema can build up and disappear again. In the acute phase, the DWI denotes the infarcted region as hyperintensity. The magnitude of the actual underperfusion shows up on perfusion maps. The mismatch between these two is often considered the potentially salvageable tissue, termed *penumbra* (González et al., 2011). Second, stroke lesions can appear at any location in the brain and take on any shape. They may or may not align with the vascular supply territories and multiple lesions can appear at the same time (e.g. caused by an embolic shower). Some lesions may have radii of few milliliters while others encompass almost a complete hemisphere. Third, lesion structures may not appear as homogeneous regions, but their intensity can vary significantly within the lesion territory. In addition, automatic

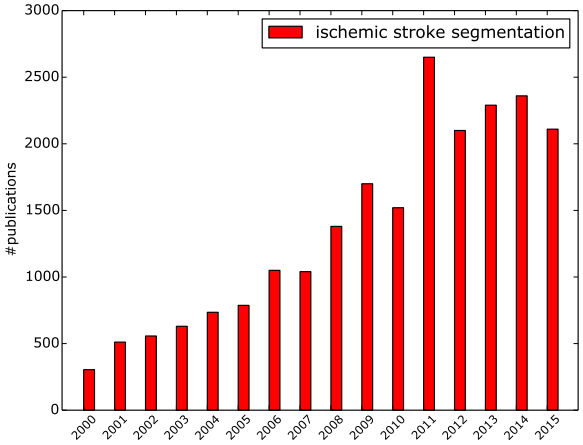


Figure 1: Increasing count of publications over the years as returned by Google scholar for the search terms *ischemic stroke segmentation* on 2015-11-05. Note that the year 2015 is still ongoing.

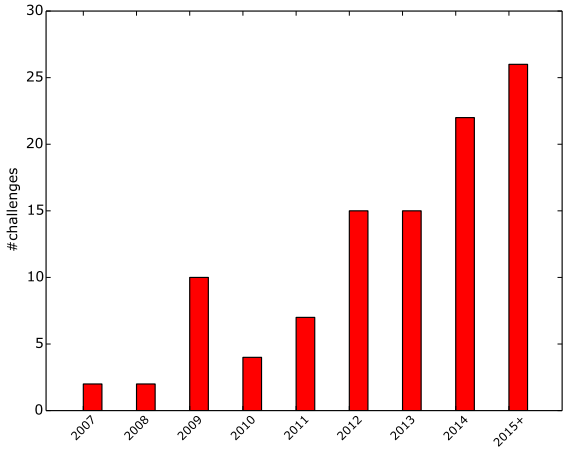


Figure 2: Increasing count of challenges over the years as collected on `http://grand-challenge.org` on 2015-08-14. Note that the year 2015 is still ongoing.

stroke lesion segmentation is complicated by the possible presence of other stroke-similar pathologies, such as older stroke or white matter hyperintensities (WMHs). The latter is especially prevalent in older patients which constitute the highest risk group for stroke. Finally, a good solution should work with routinely acquired MRI scans of clinical quality. That means coping with movement artifacts, imaging artifacts and the effects of varying scanning parameters and machines.

1.1. Current methods

The quantification of stroke lesions has gained interest during the past years (Fig. 1). Still, only few groups have approached towards automating the image segmentation task, and most of them only very recently. A recent review of non-chronic stroke lesion segmentation (Rekik et al., 2012) summarizes the most important works until 2008, reporting as few as 5 automated stroke lesion segmentation algorithms. A collection of more re-

cent approaches not included in Rekik et al. (2012) are reported in Tab. 1. While more and more automatic solutions are presented, there are also a number of semi-automatic methods indicating the difficulty of the task. Of the automatic algorithms, only a few employ pattern classification techniques to learn segmentation (Prakash et al., 2006; Maier et al., 2014, 2015c) or design probabilistic generative models of the lesion formation (Derntl et al., 2015; Menze et al., 2015; Forbes et al., 2010; Kabir et al., 2007; Martel et al., 1999).

While all approaches make an effort to quantify segmentation accuracies, most lack detailed descriptions of the dataset which is a critical matter as stroke lesion shape and appearance changes rapidly during the first hours and days, significantly altering the difficulty of the segmentation task. Information about the stroke evolution phase is sometimes omitted (Seghier et al., 2008; Forbes et al., 2010) or, where mentioned, not clearly defined (Saad et al., 2011; Muda et al., 2015). Where provided, the definition of acute stroke often mixes with the sub-acute phase (Ghosh et al., 2014; Mah et al., 2014; Tsai et al., 2014). Only few studies detail on pathological inclusion and exclusion criteria, except in a few rare cases (James et al., 2006; Maier et al., 2015c): Results obtained on right-hemispheric stroke only (Dastidar et al., 2000) are not comparable to ones omitting small lesions (Mah et al., 2014) nor to those obtained from two central axial slices of each volume (Li et al., 2004). Comparability is further impeded by a wide range of dataset sizes ($N \in [2, 57]$), and the set of MRI sequences used in the different centers varies as do the quantitative evaluation measures. All this renders the interpretation of the results difficult and explains the wide range of segmentation accuracies reported over the years. A very recent work (Maier et al., 2015b) compares a number of classification algorithms on a common dataset, but these do not represent the state-of-the-art nor are they implemented by their respective authors.

In the present benchmark study we approach the urgent problem of comparability. To this end, we planned, organized and pursued the *I*schemic *S*troke *LE*sion *S*egmentation (ISLES) challenge: A direct, fair and independently controlled comparison of automatic methods on a carefully selected public dataset. ISLES 2015 was organized as a satellite event of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2015, held in Munich, Germany. It was posing two sub-challenges dealing with different phases of the stroke lesion evolution: First, the *S*troke *P*erfusion *ES*timation (SPES) challenge dealing with the image interpretation of the early stroke and, second, the *S*ub-acute *I*schemic *S*troke lesion *S*egmentation (SISS) challenge dealing with the later stroke image patterns. In both tasks we aim at answering a number of open questions: What is the current state-of-the-art performance of automatic methods for ischemic stroke lesion segmentation? Which type or class of algorithm is most suited for the task? Which difficulties are overcome and which challenges are still faced? And what are the recommendations we can give researchers in the field after the extensive evaluation conducted?

2. Setup of ISLES

Image segmentation challenges aim at an independent and fair comparison of various segmentation methods for a given segmentation task. In these de-facto benchmarks participants are first provided with representative training data with associated ground truth, on which the participants can adjust their algorithms. Later, a testing dataset without ground truth is distributed and the participants submit their results to the organizers, who score and rank the submissions.

Previous challenges in the medical image processing communities deal with the segmentation of tumors (Menze et al., 2015) or multiple sclerosis lesions (Styner et al., 2008) in MRI of the brain; complete lungs (Murphy, 2011) or their vessels (Rudyanto et al., 2014) in computed tomography scans; 4D ventricle extraction (Petitjean et al., 2015) as well as myocardial tracking and deformation (Tobon-Gomez et al., 2013); prostate segmentation from MRI (Litjens et al., 2014); and brain extraction in adults (Shattuck et al., 2009) and neonatal (Išgum et al., 2015).

The number of challenges has been steadily increasing over the past years (Fig.2), as visible from the events listed on `http://grand-challenge.org`. Many of these have become the de-facto evaluation standards for new algorithms in particular when adhering to some standards also listed at `http://grand-challenge.org`: Both training and testing datasets should be representative for the task, well described and large enough to draw significant conclusions from the results; the associated ground truth is created by experts following a clearly defined set of rules; the evaluation metrics chosen capture all aspects relevant for the task; and ideally, challenges remain open for future contestants and serve as an ongoing benchmark for algorithms in the field.

With ISLES 2015, we introduce a benchmark for the growing but inaccessible collection of stroke lesion segmentation algorithms. The challenge was launched in February 2015 to provide a fair and direct comparison benchmark for ischemic stroke lesion segmentation. Potential participants were contacted directly following an extensive literature review on stroke segmentation or via suitable mailing lists. The training dataset of 28/30 (SISS/SPES) was released in April 2015 using the The SICAS Medical Image Repository (SMIR) platform[1] (Kistler et al., 2013). The participants could download the testing dataset of 36/20 cases from the 14th of September 2015 and had to submit their results within a week. The ground truth for this second set is kept private with the organizers. Repeated submissions were allowed, but only the last counted. The organizer evaluated the submitted results and presented them during a final workshop at the international MICCAI conference 2015 in Munich, Germany. All conclusions presented in this paper are drawn from these testing results.

We refrained from an on-site evaluation, as previous attempts (Murphy et al., 2011; Menze et al., 2015; Petitjean et al., 2015) have shown such an endeavor to be prone to complications unrelated to the actual algorithms' performances. Instead,

[1] `www.smir.ch`

Table 1: Listing of publications describing non-chronic stroke lesion segmentation in MRI with evaluation on human image data since Rekik et al. (2012). Column **A** denotes the lesion phase i.e. (A)cute, (S)ub-acute or (C)hronic. Column **T** denotes the method type i.e. (A)utomatic or (S)emi-automatic. Column **N** denotes the number of testing cases (mostly leave-one-out evaluation scheme is employed). Column **Sequences** denotes the used MRI sequences. Column **DC** denotes the reported Dice's coefficient score if available. Column **Metrics** denotes the metrics used in the evaluation. Abbreviations are: V=visual evaluation,VE=volume error,PPV=positive prediction value,+=other metrics,^{m}=median reported. Note that the lesion phases were adapted to our definition if sufficient information was available.

Method	A	T	N	Sequences	DC	Metrics
Prakash et al. (2006)	A	A	57	DWI	0.72	DC,+
Soltanian-Zadeh et al. (2007)	ASC	A	2	T1,T2,DWI,PD		+
Seghier et al. (2008)	SC	A	8	T1	0.64	DC
Forbes et al. (2010)	?	A	3	T2,FLAIR,DWI	0.63	DC
Saad et al. (2011)	AC	A	?	DWI		V
S et al. (2012)	A	S	41	DWI,ADC	0.81	DC
Artzi et al. (2013)	AS	S	10	FLAIR,DWI		ASSD,HD,VE
Maier et al. (2014)	S	A	8	T1,T2,FLAIR,DWI,ADC	0.74	DC,ASSD,HD
Tsai et al. (2014)	AS	A	22	DWI,ADC	0.9	DC,PPV
Mah et al. (2014)	S	A	38	T2,DWI	0.73	DC^m,+
Nabizadeh et al. (2014)	AS	S	6	DWI	0.80	DC,+
Ghosh et al. (2014)	S	A	2	ADC		VE
Maier et al. (2015c)	S	A	37	T1,T2,FLAIR,DW,ADC	0.63	DC,ASSD,HD
Muda et al. (2015)	AC	A	20	DWI	0.73	DC
Derntl et al. (2015)	S	A	13	T1,T1c,T2,FLAIR	0.42	DC
Menze et al. (2015)	AS	A	18	T1,T1c,T2,FLAIR,DWI	0.78	DC
Maier et al. (2015b)	S	A	37	FLAIR	0.44-0.67	DC,ASSD,HD
Maier et al. (2015b)	S	A	37	T1,T2,FLAIR,DW,ADC	0.54-0.73	DC,ASSD,HD

the results obtained on the evaluation set were hidden from the participants to avoid tuning on the testing dataset. It might be argued that such an approach is vulnerable to cheating, but even during on-site evaluations the annotation of the images could be guided by user interaction and a certain amount of trust should be granted.

After the challenge workshop, the benchmark phase is continued and new methods can enter the competition by downloading image data from and uploading results to the SMIR evaluation platform. The results and rankings of the initial participants remain as a frozen table a the challenge web page[2], while the SMIR platform supplies an automatically generated listing of these and all future results.

2.1. Participants

Interested research teams could register for one or both of the tasks. All submitted algorithms were required to be fully automatic, no other restrictions were imposed. Until the day of the challenge, the SMIR platform listed over 120 registered users for the ISLES 2015 challenge and a similar count of training dataset downloads. Of these, 14 teams provided testing dataset results for SISS and 7 algorithms participates in SPES. Their affiliations and methods can be found in Tab. 2. For a detailed description algorithm, please refer to Appendix A.

3. Data and methods

3.1. SISS image data and ground truth

We gathered 64 sub-acute ischemic stroke cases for the training and testing sets of the SISS challenge. A total of 56 cases were supplied by the Stroke Unit at the University Medical Center Schleswig-Holstein in Lübeck, Germany and acquired in diagnostic routine with varying resolutions, views and imaging artifact load. Another eight cases came from the Clinic for Neuroradiology at the Klinikum rechts der Isar in Munich, Germany. The local ethics committee approved their release under Az.14-256A. Full data anonymization was ensured by removing all patient information from the files and the facial bone structure from the images.

Considered for inclusion were all cases with a diagnosis of ischemic stroke for which at least the set of T1-weighted (T1), T2-weighted (T2), DWI ($b = 1000$) and FLAIR MRI sequences had been acquired. Additional pathological deformation, such as e.g. non-stroke WMHs, haemorrhages or previous strokes, did not lead to the exclusion of a case. Scans performed outside of the sub-acute stroke development phase were reject. Due to the lack of the exact time passed since stroke onset in most cases, lesions were visually classified as sub-acute infarct if pathologic signal was found concomitantly in FLAIR and DWI images (presence of vasogenic and cytotoxic edema with evidence of swelling due to increased water content).

Acquired in a routine diagnostic setting and representing the clinical reality, these data with secondary pathologies, stroke similar deformations, older stroke, imaging artifacts, varying acquisition axes, differing resolutions, movement artifacts, etc. Hence, some display acquisition artifacts such as movement or Gibbs, some where recorded in axial, others in sagittal direction and with different voxel sizes. To allow for an evaluation of the participating algorithms rather than the preprocessing methods employed by each team, the cases have been consistently preprocessed by the organizers: All MRI sequences are skull-stripped, re-sampled to an isotropic spacing of $1^3 mm$ and rigidly co-registered to the FLAIR sequences.

[2] `www.isles-challenge.org`

Table 2: List of all participants in the ISLES challenge. All teams are color coded for easier reference in all further listings. The ML column denotes whether the submitted algorithm is based on machine learning. Refer to the SISS and SPES columns for the sub-challenges each team participated in. Additionally, a very short summary of each method is provided. For a detailed description of each algorithm and used abbreviations, see Appendix A.

	Team	FN	SN	ML	SISS	SPES
	UK-Imp1	Liang	Chen	Y	Y	
		Regional RFs (dorsal, medial, ventral)				
	DE-Dkfz	Michael	Goetz	Y	Y	
		Image selector RF + online lesion ET				
	FI-Hus	Hanna	Halme	Y	Y	
		RF (deviation from global average) + Contextual Clustering (CC)				
	CA-McGill	Andrew	Jesson	Y	Y	
		Local classifiers (554 GMM) + regional RF				
	UK-Imp2	Konstantinos	Kamnitsas	Y	Y	
		2-path 3D CNN				
	US-Jhu	John	Muschelli	Y	Y	
		RF (e.g. SD, skew, kurtosis)				
	SE-Cth	Qaiser	Mahmood	Y	Y	
		RF (e.g. gradient, entropy)				
	US-Odu	Syed	Reza	Y	Y	
		RF (many features e.g. texture)				
	TW-Ntust	Ching-Wei	Wang	Y	Y	
		RF (many features e.g. edge)				
	CN-Neu	Chaolu	Feng	N	Y	Y
		Bias-correcting Fuzzy C-Means + Level Set				
	BE-Kul1	Tom	Haeck	N	Y	Y
		Tissue priors + EM-opt MRF + Level Set on sequence subset				
	CA-USher	Francis	Dutil	Y	Y	Y
		2-path 2D CNN				
	DE-UzL	Oskar	Maier	Y	Y	Y
		RF (anatomically and appearance motivated features)				
	BE-Kul2	David	Robben	Y	Y	Y
		2-step cascaded ETs				
	DE-Ukf	Elias	Kellner	N		Y
		Rule-based hemisphere-comparing approach				
	CH-Insel	Richard	McKinley	Y		Y
		RF (case bootstrapped forest of forests)				

Table 4: Details of the SISS data.

number of cases	28 training and 36 testing
number of medical centres	1 (train), 2 (test)
number of expert segmentations for each case	1 (train), 2 (test)
MRI sequences	FLAIR, T2w TSE, T1w TFE/TSE, DWI

In addition to the wide range of acquisition and clinical related variety, the sub-acute lesions themselves display a wide range of variability (Tab. 3). Great care has been taken to preserve the diversity of the stroke cases when splitting the data into testing and training datasets: both contain single- and multi-focal cases, small and large lesions, etc. Main difference is the addition of the additional 8 cases from Munich to the testing dataset only, hence no training could be performed on this second center data. The testing dataset is therefore slightly larger than the training dataset (Tab. 4).

All expert segmentation used in ISLES were prepared by experienced raters. For SISS, two ground truth sets (GT01 and GT02) were created and the segmentations were performed on the FLAIR sequence, which is known to exhibit lower inter-rater differences as e.g. T2 (Neumann et al., 2009). The guidelines for the expert raters were as follows:

1. The segmentation is performed on the FLAIR sequence
2. Other sequences provide additional information
3. Only sub-acute ischemic stroke lesion are segmented
4. Partially surrounded sulci/fissures are not included
5. Very thin/small or largely surrounded sulci/fissures are included
6. Surrounded hemorrhagic transformations are included
7. Segmentation contains no holes
8. Segmentation is exact, but spatially consistent (no sudden spikes or notches)

Infarct lesions with signal change due to hemorrhagic transformation were included. Acute infarct lesions (DWI signal for cytotoxic edema only, no FLAIR signal for vasogenic edema) or residual infarct lesions with gliosis and scarring after infarction (no DWI signal for cytotoxic edema, no evidence of swelling) were not included.

3.2. SPES image data and ground truth

All patients included in the SPES dataset were treated for acute ischemic stroke at the University Hospital of Berne between 2005 and 2013. Patients included in the dataset received a diagnosis of ischemic stroke by MRI imaging with an identifiable lesion on DWI as well as on perfusion weighted imaging (PWI), with a proximal occlusion of the middle cerebral artery

Table 3: Stroke lesion characteristics of the 64 SISS cases. The strong diversity is representative for stroke lesions and emphasizes the difficulty of the task. μ denotes the mean value, $[min, max]$ the interval and n the total count. Abbreviations are: anterior cerebral artery (ACA), middle cerebral artery (MCA), posterior cerebral artery (PCA) and basilar artery (BA).

Lesion count	$\mu = 2.46$
	$[1, 14]$
Lesion volume	$\mu = 17.59ml$
	$[1.00, 346.06]$
Haemorrhage present	$n_1 = 12$
	0=no,1=yes
Non-stroke WMH load	$\mu = 1.34$
	0=none, 1=small, 2=medium, 3=large
Lesion localization (lobes)	$n_1 = 11, n_2 = 24, n_3 = 42, n_4 = 17, n_5 = 2, n_6 = 6$
	1=frontal, 2=temporal, 3=parietal, 4=occipital, 5=midbrain, 6=cerebellum
Lesion localization	$n_1 = 36, n_2 = 49$
	1=cortical, 2=subcortical
Affected artery	$n_1 = 6, n_2 = 45, n_3 = 11, n_4 = 5, n_5 = 0$
	1=ACA, 2=MCA, 3=PCA, 4=BA, 5=other
Midline shift	$n_0 = 51, n_1 = 5, n_2 = 0$
	0=none, 1=slight, 2=strong
Ventricular enhancement	$n_0 = 38, n_1 = 15, n_2 = 3$
	0=none, 1=slight, 2=strong
Laterality	$n_1 = 18, n_2 = 35, n_3 = 3$
	1=left, 2=right, 3=both

(MCA) (M1 or M2 segment) documented on digital subtraction angiography. Further requirements were that an attempt at endovascular therapy was undertaken, either by intra-arterial thrombolysis (before 2010) or by mechanical thrombectomy (since 2010), a minimum patient age of 18 and no motion artifacts.

The stroke MRI was performed on either a 1.5T (Siemens Magnetom Avanto) or 3T MRI system (Siemens Magnetom Trio). The stroke protocol encompassed whole brain DWI (24 slices, thickness 5mm, repetition time 3200ms, echo time 87ms, number of averages 2, matrix 256 * 256) yielding isotropic b1000 images. For PWI the standard dynamic-susceptibility contrast enhanced perfusion MRI (gradient-echo echo-planar imaging sequence, repetition time 1410ms, echo time 30 ms, field of view 230 * 230 mm, voxel size: 1.8 * 1.8 * 5.0 mm, slice thickness 5.0 mm, 19 slices, 80 acquisitions) was acquired. PWI images were recorded during the first pass of a standard bolus of 0.1mmol/kg gadobutrol (Gadovist, Bayer Healthcare). Contrast medium was injected at a rate of 5 ml/s followed by a 20ml bolus of saline at a rate of 5ml/s. Perfusion maps were obtained by block-circular singular value decomposition using the Perfusion Mismatch Analyzer (PMA, from Acute Stroke Imaging Standardization Group ASIST) Ver.3.4.0.6.

Sequences and derived maps made available to the participants are T1 contrast enhanced (T1c), T2, DWI, cerebral blood flow (CBF), cerebral blood volume (CBV), time-to-peak (TTP) and time-to-max (Tmax) (Tab. 5).

For preprocessing, all were rigidly registered to the T1c with constant resolution of $2 \times 2 \times 2$ and automatically skull-stripped (Bauer et al., 2013). Together with the removal of all patient data from the files full anonymization was achieved.

To determine the eligibility of a patient for treatment or to assess a treatment response in clinical trials, the pretreatment estimation of the potentially salvageable penumbral area is crucial. A 6 second threshold applied to the Tmax map has been suggested (Straka et al., 2010) and successfully applied in large multi-center trials (Lansberg et al., 2012) to determine the area of hypoperfusion (i.e. penumbra + core). But this approach requires the manual setting of a region of interest as well as considerable manual postprocessing. For SPES, we are interested in whether advanced segmentation algorithms could replace the human intervention and render the estimation simultaneously reproducible.

Table 5: Details of the SPES data.

number of cases	30 training and 20 testing
number of medical centres	1
number of expert segmentations for each case	1
MRI sequences	T1c, T2, DWI, CBF, CBV, TTP, Tmax

The hypoperfused tissue was segmented semi-manually with Slicer 3D Version 4.3.1 by a medical doctor with a preadjusted threshold for Tmax of 6 seconds applied to regions of interest as described in Straka et al. (2010); Lansberg et al. (2012). The label represents the stroke-affected regions with restricted perfusion, which is the first requirement to determine the penumbral area via a perfusion-diffusion mismatch approach.

The collected data hence consists of a variety of acute MCA cases (Tab. 6). The split into training and testing cases was performed by an experienced neuroradiologist by visual inspection of cases and using as criteria the complexity in visually defining the extent of the penumbra area.

The training dataset is additionally equipped with a manually created DWI segmentation ground truth set denoting roughly the stroke's core area. These are not evaluated in the challenge.

3.3. Evaluation metrics

As measures, we employ (1) Dice's coefficient (DC), which describes the volume overlap between two segmentations and

Table 6: Stroke lesion characteristics of the 50 SPES cases. The cases are restricted to MCA stroke eligible for cerebrovascular treatment. μ denotes the mean value, $[min, max]$ the interval and n the total count.

Lesion count	$\mu = 1$
	Not always connected, but single occlusion as source.
Lesion volume	$\mu = 133.21ml$
	[45.62, 252.20]
Affected artery	all MCA
Laterality	$n_1 = 22, n_2 = 28, n_3 = 0$
	1=left, 2=right, 3=both

is sensitive to the lesion size, (2) the average symmetric surface distance (ASSD), which denotes the average surface distance between two segmentations, and (3) the Hausdorff distance (HD), which is a measure of the maximum surface distance and hence especially sensitive to outliers. Additionally, precision and recall values are reported for each classifier to assess over- and under-segmentation, respectively.

The DC is defined as

$$DC = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

with A and B denoting the set of all voxels of ground truth and segmentation respectively. To compute the ASSD, we first define the average surface distance (ASD), a directed metric, as

$$ASD(A_S, B_S) = \frac{\sum_{a \in A_S} min_{b \in B_S} d(a, b)}{|A_S|} \quad (2)$$

and then average over both directions to obtain the ASSD

$$ASSD(A_S, B_S) = \frac{ASD(A_S, B_S) + ASD(B_S, A_S)}{2} \quad (3)$$

Here A_S and B_S denote the surface voxels of ground truth and segmentation respectively. Similar, the HD is defined as the maximum of all surface distances with

$$HD(A_S, B_S) = \max\{\max_{a \in A_S} \min_{b \in B_S} d(a, b), \max_{b \in B_S} \min_{a \in A_S} d(b, a)\} \quad (4)$$

The distance measure $d(\cdot)$ employed in both cases is the Euclidean distance, computed taking the voxel size into account. Finally, precision and recall are calculated from true positive (TP), false positive (FP) and false negative (FN) voxels as

$$precision = \frac{TP}{TP + FP} \quad (5)$$

and

$$recall = \frac{TP}{TP + FN} \quad (6)$$

3.4. Ranking

After selecting suitable evaluation metrics, we face the problem of establishing a meaningful ranking for the competing algorithms: The different measures are neither in the same range nor direction.

In the simplest case, a single of the above metrics can be chosen and used in the ranking (Menze et al., 2015). But this would mean neglecting the aspects revealed by the remaining measures and is hence a bad choice for most challenges.

A second approach taken by some challenges (Styner et al., 2008) is to compare two expert segmentations against each other. The resulting evaluation values are then assumed to indicate the upper limit and hence denote the 100 percent mark of each measure. New segmentations are then evaluated and the values compared to their respective 100 percent mark, resulting in a percentage rating for each measure. Drawback is that for not range restricted measures, such as the ASSD, one has to define an arbitrary 0 percent mark.

We chose a third approach based on the ideas of Murphy et al. (2011), which builds on the concept that a ranking reveals only the direction of a relationship between two items (i.e. higher, lower, equal), but not its magnitude. Basically, each participant's results are ranked per case according to each measure and then the obtained ranked averaged. For a more detailed account, see Appendix B.

4. Results: SISS

4.1. Inter-observer variance

Comparing the ground truths of SISS against each other provides the baseline above which an automatic method can be considered to produce results superior to a human rater and a measure of the task's difficulty (Tab. 7, last row). The two expert segmentations overlap at least partially for all cases. Compared to similar tasks, such as e.g. brain tumor segmentation, for which inter-observer DC values of 0.74 ± 13 to 0.85 ± 0.08 are reported (Menze et al., 2015), the ischemic stroke lesion segmentations problem can be considered difficult with a DC score of 0.70 ± 0.20.

4.2. Leaderboard

The main result of the SISS challenge is a leaderboard for state-of-the-art methods in sub-acute ischemic stroke lesion segmentation (Tab. 7, averages and standard deviations (STD)). The evaluation measures and ranking system employed are described in the method part of this article (Sec. 3.4). No participating method managed to segment all 36 testing cases successfully and the best scores are still far from the human rater performance. Note that for further experiments, we will focus on DC averages only, as the ASSD and HD values cannot be readily computed for the failed cases and are thus not suitable for a direct comparison of methods with non-matching successfully segmented cases.

4.3. Statistical analaysis

We performed a statistical analysis of the results to rule out random influences on the leaderboard ranking. Each pair of methods is compared with the two-sided Wilcoxon signed-rank test (Wilcoxon, 1945), a nonparametric test of the null hypothesis that two samples come from the same population against an alternative hypothesis (Fig. 3).

The two highest ranking methods show no statistically significant differences with a confidence of 95% (i.e. $p < 0.025$).

Table 7: SISS challenge leaderboard after evaluating the 14 participating methods on the testing dataset. The *rank* is the final measure for ranking the algorithms' performances relative to each other. The *cases* column denotes the number of successfully (DC> 0) segmented cases. All evaluation measures are given in mean±STD. Please note that the ASSD and HD values were computed excluding the failed cases. The last row shows the inter-observer results for comparison.

Rank	Method	Cases	ASSD (mm)	DC [0,1]	HD (mm)
3.25	UK-Imp2	34/36	05.96 ± 09.38	0.59 ± 0.31	37.88 ± 30.06
3.82	CN-Neu	32/36	03.27 ± 03.62	0.55 ± 0.30	19.78 ± 15.65
5.63	FI-Hus	31/36	08.05 ± 09.57	0.47 ± 0.32	40.23 ± 33.17
6.40	US-Odu	33/36	06.24 ± 05.21	0.43 ± 0.27	41.76 ± 25.11
6.67	BE-Kul2	33/36	11.27 ± 10.17	0.43 ± 0.30	60.79 ± 31.14
6.70	DE-UzL	31/36	10.21 ± 09.44	0.42 ± 0.33	49.17 ± 29.6
7.07	US-Jhu	33/36	11.54 ± 11.14	0.42 ± 0.32	62.43 ± 28.64
7.54	UK-Imp1	34/36	11.71 ± 10.12	0.44 ± 0.30	70.61 ± 24.59
7.66	CA-USher	27/36	09.25 ± 09.79	0.35 ± 0.32	44.91 ± 32.53
7.92	BE-Kul1	30/36	12.24 ± 13.49	0.37 ± 0.33	58.65 ± 29.99
7.97	CA-McGill	31/36	11.04 ± 13.68	0.32 ± 0.26	40.42 ± 26.98
9.18	SE-Cth	30/36	10.00 ± 06.61	0.38 ± 0.28	72.16 ± 17.32
9.21	DE-Dkfz	35/36	14.20 ± 10.41	0.33 ± 0.28	77.95 ± 22.13
10.99	TW-Ntust	15/36	07.59 ± 06.24	0.16 ± 0.26	38.54 ± 20.36
	inter-observer	36/36	02.02 ± 02.17	0.70 ± 0.20	15.46 ± 13.56

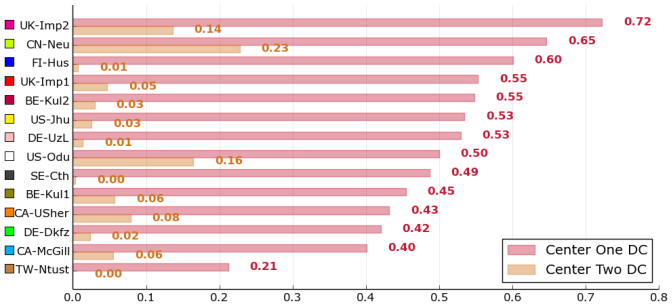


Figure 4: Adaptation to the data from the second medical center. The graph shows each methods average DC scores on the 28 cases from the first and the eight cases from the second medical center.

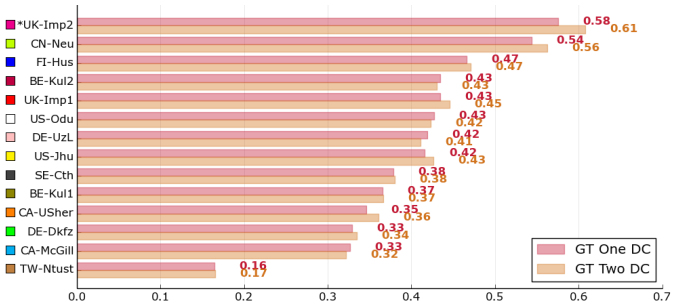


Figure 5: Differences in performance on the two ground truth sets. The graph shows each methods average DC scores on the 36 testing dataset cases broken down by ground truth set. A star (*) before a team's name denotes statistical significant difference according to a paired Student's t-test with $p < 0.05$.

No other algorithm performs better than them and they both are better than the 12 remaining ones. Next comes a group of four methods to which only the two winners prove superior. But among these, FI-Hus takes the highest position, as it is statistically better than eight other methods, while the other three only prove superior to at most four competitors. The established leaderboard ranking is largely confirmed by the statistical analysis.

4.4. Impact of multi-center data

Cases acquired at different medical centers can differ greatly in appearance. A good automatic stroke lesion segmentation method should be able to cope with these variations. We broke down each methods' results by medical center (Fig. 4) to test whether this holds true for the participating algorithms.

Since the training dataset contained only cases from the first center, the difference in performance should reveal the methods' generalization abilities. We observed that not a single algorithm reached second center scores comparable to its first center scores. This is a strong hint towards a difficult adaptation problem.

4.5. Dependency on observer variations

A good segmentation method does not only adapt well to second center data, but equally to another observer's ground truth. Only the GT01 ground truth set was made available to the participating teams during the training/tuning phase. Hence particularly machine learning solutions could be expected to show deficits on the second rater ground truth GT02. To test how well the methods generalize, we compared their performance on the testing sets GT01 ground truth against their performance on the formerly unseen GT02 set (Fig. 5).

The average DC scores of each method differed only slightly over the ground truth sets. Only in a single case, UK-Imp2, the difference was significant (paired Student's t-test with $p < 0.05$), but the higher results were obtained for the, formerly unseen, GT02 set. We can hence conclude that all algorithms generalized well in the sense of ground truth sets. An additional data analysis showed the ranking of the methods doe snot change if only one or the other of the ground truth sets is employed for evaluation.

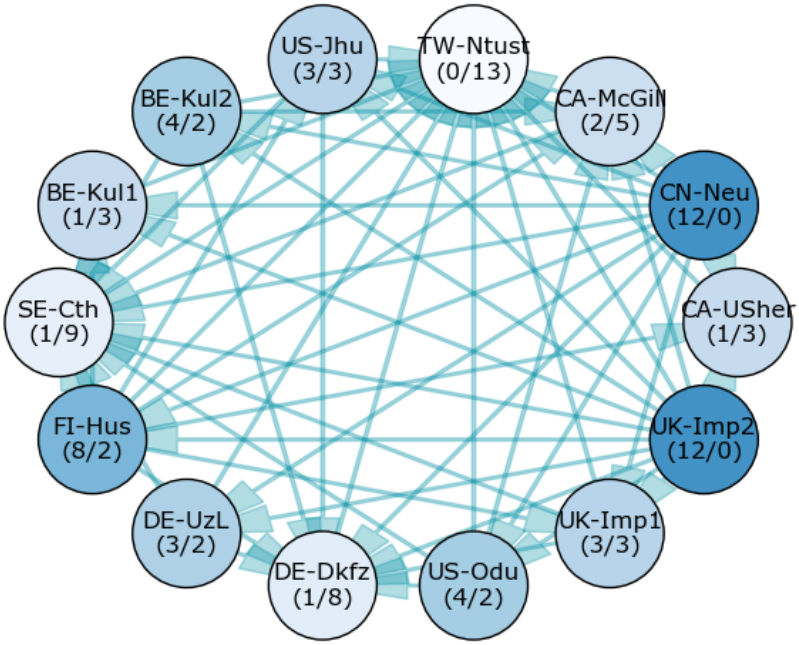


Figure 3: Significant differences between the 14 participating methods' case ranks according to a two-sided Wilcoxon signed-rank test ($p < 0.025$). Each node represents a team, each edge a significant difference of the tail side team over the head side team. Therefore, the less outgoing and the more incoming edges a team has (denoted by numbers in brackets (*#out*/*#in*) for easier interpretation), the weaker its method compared to the others. The saturation of the node colors indicates the strength of a method, where better methods are highlighted with more saturated colors. Note that all teams with the same number of incoming and outgoing edges perform, statistically spoken, equally well. A higher importance of incoming over outgoing edges or vice-versa cannot be readily established.

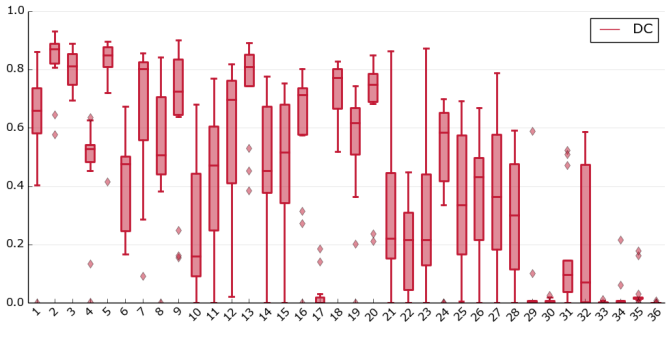


Figure 6: Box plots of the 14 teams' DC results on all testing dataset cases. E.g. the first box was computed from all teams' results on the first case. The band in the box denotes the median, the upper and lower limits the first and third quartile. Outliers are plotted as diamonds.

4.6. Outlier cases

A benchmark is only as good as its data. The average scores obtained on the different cases of the testing dataset differed widely and some proved especially difficult or easy to segment (Fig. 6). For all cases starting from 29, this variation can be explained through the different acquisition parameters at the second medical center. But the weak performance of all methods on cases such as 10, 17 and 23 must have other reasons. We therefore took a closer look at these three as well as the overall most successful cases 2, 5 and 13 (Fig. 7).

The three cases that were successfully processed by nearly all algorithms show large, clearly outlined lesions with a strongly hyperintense FLAIR signal. In two of these cases, the DWI signal is relatively weak, in some areas nearly isointense. Still, for these cases the algorithms displayed the highest confidence. One of the difficult cases contains only a single small lesion with marginal FLAIR and strong DWI hyperintensities. Another case, equally showing a small lesion, has a stronger FLAIR support, but also displays large periventricular WMHs that seem to confuse most algorithms, despite missing DWI hyperintensities. This behavior was also visible for the third of the failed cases. Here, the actual lesion is correctly segmented by most methods, as it is clearly outlined with strong FLAIR and DWI support. But many algorithms additionally delineated parts of the periventricular WMHs, which again only show up in the FLAIR sequence.

4.7. Correlation with lesion characteristics

The properties of the cases might have an influence on the segmentation quality, as some are clearly easier to segment than others. To find such correlations, we related various lesion characteristics to the average DC scores obtained over all teams using suitable statistics (Tab. 8).

Significant moderate correlation was found between the lesion volume and the average DC values. Statistically significant difference of means was found comparing cases with haemorrhage present versus cases without and between left hemispheric and right hemispheric lesions. Since the characteristics

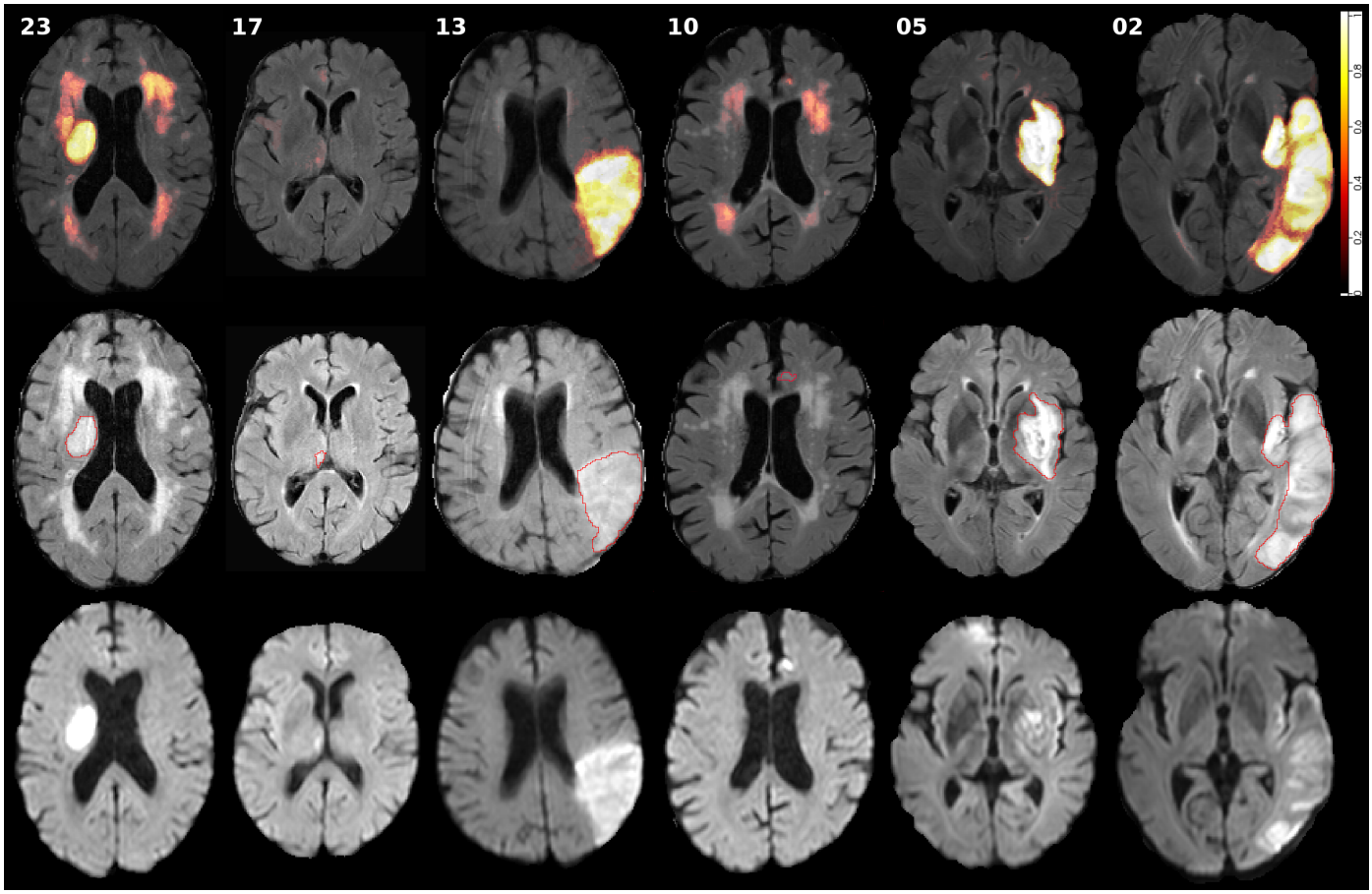


Figure 7: Visual results for selected difficult (10, 17, 23) and easy (2, 5, 13) cases from the SISS testing dataset. The first row shows the distribution of all 14 submitted results on a slice of the FLAIR volume. The second row shows the same image with the ground truth (GT01) outlined in red. And the third row shows the corresponding DWI sequence. Please refer to the online version for colors.

Table 8: Correlation between the SISS case characteristics and the average DC values over all teams. A ρ denotes a Spearman correlation, a t a Student's t-test. All p-values are two tailed (p_2). Significant results according to a 95% confidence interval are denoted by a $*$. Secondary tests appearing in the table were performed against the lesion volume rather than the average DC values.

Characteristic	Test	p_2
Lesion count	$\rho = -0.21$	0.23
Lesion volume	$\rho = +0.76$	0.00$*$
Haemorrhage present	$t = +2.29$	0.03$*$
vs. lesion volume	$t = +4.33$	0.00$*$
Non-stroke WMH load	$\rho = -0.01$	0.97
Midline shift	$t = +0.51$	0.62
Ventricular enhancement	$t = +1.56$	0.13
Laterality	$t = +2.66$	0.01$*$
vs. lesion volume	$t = +2.12$	0.03$*$
Movement artifacts	$\rho = -0.30$	0.08
Imaging artifacts	$\rho = +0.24$	0.15

can not be assumed independent, we furthermore tested the last two groupings for significant differences in lesion volumes between the groups. This was found in both cases (see secondary test for each of these two characteristics). We could not establish a reliable significant influence on the results for any single parameter. Even the lesion volume's influence is not ensured, as we will detail in the discussion.

5. Results: SPES

5.1. Leaderboard

To establish an overall leaderboard for state-of-the-art methods in automatic acute ischemic stroke lesion segmentation, all submitted results were ranked relatively as described in Sec.3.4 (Tab. 9).

This ranking is the outcome of the opening event and was used to determine the competition winners. No completely failed segmentation was handed in for any of the algorithms and the evaluation measures of the highest ranking teams denote a high segmentation accuracy.

5.2. Statistical analysis

A strict ranking is suited to determine the winners of a competition, but the differences between method performances are not necessarily reliable. Therefore, we performed a statistical analysis of the results to rule out random influences on the leaderboard. Each pair of methods was compared with the two-sided Wilcoxon signed-rank test (Wilcoxon, 1945), a nonparametric test of the null hypothesis that two samples come from the same population against an alternative hypothesis (Fig. 8).

Statistically spoken, no significant difference can be observed between the two first ranked methods, nor between the third and fourth place. Hence, SPES has two first ranked, two second ranked and one third ranked method according to the statistical analysis.

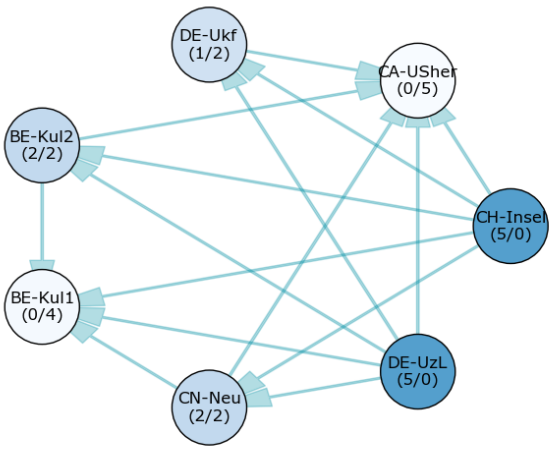


Figure 8: Visualization of significant differences between the 7 participating methods' case ranks. Each node represents a team, each edge a significant difference of the tail side team over the head side team according to a two-sided Wilcoxon signed-rank test ($p < 0.025$). Therefore, the less outgoing and the more incoming edges a team has (denoted by numbers in brackets (*#out*/*#in*) for easier interpretation), the weaker its method compared to the others. The saturation of the node colors roughly denotes the strength of a method, where better methods are depicted with stronger colors. Note that all teams with the same number of incoming and outgoing edges perform, statistically spoken, equally well.

5.3. Results per case and method

An average and a statistical similarity between the first two and second two methods was already established. To test whether these pairs behave similar for all of the testing dataset cases, we plotted the DC scores of each team against the cases (Fig. 9).

The performance lines of the highest ranked methods, CH-Insel and DE-UzL, display a very similar pattern and, except for some small variation, reach mostly very similar DC values. It seems like both methods are doing roughly the same. This observation does not hold true for the two runner-ups, BE-Kul2 and CN-Neu. Both methods display outliers towards the lower end and their performances for the testing dataset cases are not as near to each other as for the first two methods. I.e., while similar in performance, the methods seem to represent differ-

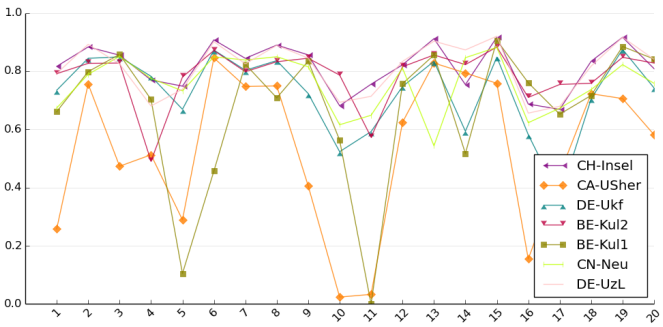


Figure 9: DC score result of all 7 SPES teams for each of the testing dataset cases. Most methods show a similar pattern. Please refer to the online version for color.

Table 9: SPES challenge leaderboard after evaluating the 7 participating methods on the testing dataset. The *rank* is the final measure for ranking the algorithms' performances relative to each other. The *cases* column denotes the number of successfully (DC> 0) segmented cases. All evaluation measures are given in mean±STD. Since no method failed completely on a single case, the reported ASSD values are suitable for a direct comparison between methods. The last two rows denote thresholding methods employed in clinical studies.

rank	method	cases	ASSD (mm)	DC [0,1]
2.02	CH-Insel	20/20	1.65 ± 1.40	0.82 ± 0.08
2.20	DE-UzL	20/20	1.36 ± 0.74	0.81 ± 0.09
3.92	BE-Kul2	20/20	2.77 ± 3.27	0.78 ± 0.09
4.05	CN-Neu	20/20	2.29 ± 1.76	0.76 ± 0.09
4.60	DE-Ukf	20/20	2.44 ± 1.93	0.73 ± 0.13
5.15	BE-Kul1	20/20	4.00 ± 3.39	0.67 ± 0.24
6.05	CA-USher	20/20	5.53 ± 7.59	0.54 ± 0.26
Tmax> 6*s* (Christensen et al., 2010)		20/20	13.02 ± 4.15	0.27 ± 0.10
Tmax> 6*s* & size< 3*ml* (Straka et al., 2010)		20/20	7.04 ± 4.99	0.32 ± 0.13

ent segmentation functions. The lowest ranked methods mainly differ from the others in the sense that they fail to cope with the more difficult cases.

Overall, most algorithms exhibit the same tendencies i.e. imaging and/or pathological differences between the cases seem to influence all methods in a similar fashion. In other words, the methods agree largely on what could be considered difficult and easy cases.

5.4. Outlier cases

We took a close look at the cases with the highest (no.15) and lowest (no.11) average DC scores (Fig. 10) to establish a rationale between the difference in obtained average DC scores.

The difficult case in the top row displays high TTP and Tmax delays in some areas of the lesion, paired with a reduced CBV. The remaining parts of the lesion show moderately increased TTP and Tmax times and a reduced CBF, but no markedly lower CBV. This hints towards a reduced, but still functioning collateral blood supply into the latter areas. The TTP interval of the easier case in the bottom row is markedly more limited, containing largely only values between 250 and 350, except in the ventricles. The lesion area is outlined relatively homogeneous in the TTP, albeit with lower values. The CBF and CBV display no clear signs. The Tmax shows a slight delay.

One could deduce that the intervals and values of the perfusion maps seem to have a great influence on the segmentation results. But a short survey of the remaining cases did show a wide range of values, seemingly uncorrelated with the average DC scores (data not shown).

6. Discussion: SISS

With the SISS challenge, we provided a public dataset with a fair and independent automatic evaluation system to serve as a general benchmark for automatic sub-acute ischemic stroke lesion segmentation methods. As main result of the opening event, we were able to assess the current state-of-the-art performance in automatic sub-acute ischemic stroke lesion segmentation and to give well-founded recommendations for future developments. In this section, we review the results of the experiments conducted, discuss their implications and try to answer the questions posed in the introduction.

Foremost, we aimed to establish if the task can be considered solved: The answer is a clear no. Even the best methods are still far from human rater performance as set by the inter-rater results. And while the observers agreed at least partially in all cases, no automatic method segmented all cases successfully. Many issues remain and a target-oriented community effort is required to improve the situation.

The best accuracy reached is an average DC of 0.6 with an ASSD of 4*mm*. The high average HD of at least 20*mm* reveals many outliers and/or missed lesions. A STD of 0.3 DC denotes high variations and indeed we observe many completely or largely failed cases for each method.

Previously published DC results on sub-acute data (Tab. 1) are all slightly to considerably better. This underlines the need for a public dataset for stroke segmentation evaluation that encompasses the whole complexity of the task: Private datasets are often too selective and the reported results differ greatly without any possibility to identify the causes behind.

The low scores obtained by all participating algorithms show that sub-acute ischemic stroke lesion segmentation is a very difficult task. This is furthermore supported by the inter-rater variations obtained, an observation that has been made before: Neumann et al. (2009) report median inter-rater agreement of $DC = 0.78$ and $HD = 23.4mm$ over 14 subjects and 9 raters and Fiez et al. (2000) volume differences of 18 ± 16%.

6.1. The most suitable algorithm and the main challenges of the task

The benchmark results were reviewed to identify the type of algorithm most suitable for sub-acute ischemic stroke lesion segmentation, but no definite winner could be determined. While there are clear differences between the submitted methods, the specific algorithms can be found over all ranks. We were not even able to determine clear performance differences between types of approaches: The two statistically equally well performing winners, include one highly trained machine learning algorithm (UK-Imp2 with a convolutional neural network (CNN)) and one non-machine learning approach (CN-Neu with fuzzy C-means). We have to conclude that many of the participating algorithms are equally suited and that the devil's in the detail. This finding is supported by the wide spread distribu-

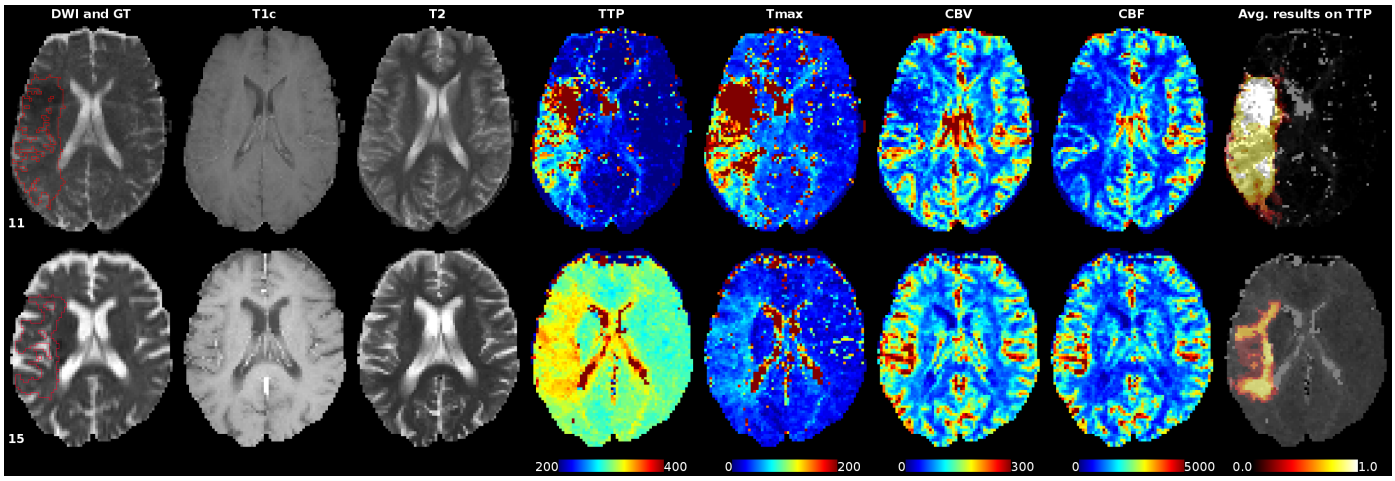


Figure 10: Sequence of the cases with the lowest (top-row) and highest (bottom-row) average DC score over all 7 teams participating in SPES. The ground truth (GT01) is painted red into the DWI sequence slices in the first row. The last row shows the distribution of the resulting segmentations on the gray-scale version of the TTP. All perfusion maps are windowed equally for direct comparison.

tion of random forest (RF) methods over all ranks, including the third and the next to last. Adaptation to the task and tuning of the hyperparameters is the key to goods results. An observation made is that the three winners all use a combination of two algorithms, possibly compensating the weak points of one with the other.

All participating methods showed good generalization abilities regarding the second rater. Since the inter-rater variability is high, we can assume that even the machine learning algorithms did not suffer from overfitting or, in other words, managed to avoid the inter-rater idiosyncrasies. Another explanation could be that the differences between the two raters fall into regions where little image information supports the presence of lesions.

Quite contrary, not a single algorithm adapted well to the second medical center data. Differences in MRI acquisition parameters and machine dependent intensity variations are known to pose a challenge for all automatic image processing methods (Han et al., 2006). Seemingly, the center-dependent differences are difficult to learn or model. Regrettably, we did not have enough second center data in the testing dataset to draw a conclusive picture: The observed high variations might equally be caused by the considerably smaller lesion sizes in the second center dataset or other factors not attributable to multi-center variations (Jovicich et al., 2009). Special attention should be paid to this point when developing applications.

Cases for which all methods obtained good results show mostly large and well delineated lesions with a strong FLAIR signal, while small lesions with only a slightly hyperintense FLAIR support posed difficulties. Surprisingly, quite a number of algorithms have troubles to differ between sub-acute stroke lesions and periventricular WMHs, despite the fact that the latter show an isointense DWI signal. This might be attributable to the strongly hyperintense DWI artifacts and often inhomogeneous lesion appearance, reducing the methods' confidence in the DWI signal. We can not tell tell whether our observations are supported by other researchers' findings, as, unfortunately, most publications fail to discuss the particularities of their results.

None of our collected lesion characteristics was found to to exhibit a significant influence on the results (Tab. 8): The lesion volume correlates significantly with the scores, but the DC is known to reach higher values for larger volumes. The apparent performance differences in the presence of haemorrhages and the dependency on laterality could both be explained by differences in the respective groups lesion sizes. To investigate combinations of characteristics, e.g. with multifactorial ANOVAs, a larger number of cases would be required.

The conclusions drawn here are meant to be general and valid for most of the participating methods. A method-wise discussion is out of the scope of this article. Any interested reader is invited to download the participants training dataset results and perform his/her own analysis to test, whether these findings hold true for a particular algorithm.

6.2. Recommendations and limitations

When developing new methods, no particular algorithms should be excluded a-priori. Instead, the characteristics of stroke lesion appearances, their evolution and the observed challenges should be studied in detail. Based on these information, new solutions targeting the specific problems can be developed. The specific algorithm can then be selected depending on how well the envisioned solutions can be integrated. Where possible, the strength of different approaches should be combined to counterbalance their weaknesses.

Evaluation should never be solely conducted on a private dataset: The variation between the cases is to large for a small set to compensate for all of them and hence renders any fair comparison impossible. We believe that with SISS we supplied a testing dataset which suitably reflects the high variation in case characteristics and the difficulty of the problem.

Special attention should be put on the adaptation to second center data, which proved to be especially difficult. One could either concentrate on single-center solutions, try to develop a

method that can encompass the large inter-center variations or aim for an approach that can be specifically adapted. The hole subject requires further investigation and should not be handled lightly.

Considering on how many cases the methods failed completely, it would be interesting to develop solutions that allow automatic segmentation algorithms to signal a warning, when they assume to have failed on a segmentation. This problem is related to multi-classifier competence, which few publications have dealt with to date (Woloszynski and Kurzynski, 2011; Galar et al., 2013).

Physicians and clinical researchers should not expect a fully automatic, reliable and precise solution in the near future. The task is simply to complex and variable for current algorithms to solve. Instead, the findings of this investigation can help them to identify suitable solutions that can serve as support tools: Especially clearly outlined, large lesions are already segmented with good results, which are usually tedious to outline by hand. For smaller and less pronounced lesions, the manual approach is still recommended. Furthermore, they should be aware that individual adaptations to each data source are most likely required - either by tuning the hyperparameters or through machine learning.

The testing dataset for the SISS challenge did not contain data from more than two centers and for the second of these only eight cases. Considering the large variability of stroke lesions, this prevented us to draw a general conclusion about second center adaptation, as other influence could not be ruled out. Should a new stroke lesion segmentation challenge be devised, the organizers might want to pay special attention to this point.

7. Discussion: SPES

All the best ranking methods show high average DC, low ASSD and only minimal STD, denoting accurate and robust results. A linear regression analysis furthermore revealed a good volume fit for the best methods (CH-Insel: $r = 0.87$ and DE-UzL: $r = 0.93$). We can say that reliable and robust perfusion lesion estimation from acute stroke MRI is in reach. For a final answer, a thorough investigation of the inter- and intra-rater scores would be required, which lies out of the scope of this work.

In clinical context a Tmax thresholding at $> 6s$ was established to correlate best with other cerebral blood flow measures (Takasawa et al., 2008; Olivot et al., 2009b) and final lesion outcome (Olivot et al., 2009a; Christensen et al., 2010; Forkert et al., 2013) and already used in large studies (Lansberg et al., 2012). We started out with the same method when creating the ground truth for SPES, but followed by considerable human correction. The comparison against the simple thresholding (Tab. 9, second to last row) hence gives an idea of the intervention in creating the ground truth. Compared against the participating methods, it becomes clear that these managed to capture the physicians intention when segmenting the perfusion lesion quite well and that simple thresholding might not suffice.

An improved version proposed by Straka et al. (2010), where binary objects smaller than $3ml$ are additionally removed, leads to better results (Tab. 9, last row) than simple thresholding, but still far from SPES' algorithms.

The discrepancy between the relatively good results reported by Olivot et al. (2009a); Christensen et al. (2010) and Straka et al. (2010) and the poor performance observed in this study can be partially explained by the different end-points (expert segmentation on PWI-MRI vs. follow-up FLAIR/T2), the different evaluation measures used (DC/ASSD vs. volume similarity) and the different data. This only serves to highlight the need for a public evaluation dataset. From a image processing point of view, the volume correlation is not a suitable measure to evaluate segmentations: It can lead to good results despite completely missed lesions.

7.1. The most suitable algorithm and the main challenges of the task

Both of the winning methods are based on machine learning (RFs) and both additionally employ expert knowledge (e.g. a prior thresholding of the Tmax map). Their obtained results are significantly better than those of all other teams. The runner-ups are another RF method and a modeling approach, followed in order by a rule based approach, another model and a CNN.

Although the number of participating methods is to small to draw a general conclusion, the results indicate that RFs in their various configuration are highly suitable algorithms for the task of stroke penumbra estimation. Furthermore, they are known to be robust and allow for a fast application, both strong requirements to use them in clinical context.

An automated method has to fulfill the strict requirements of clinical routine. Since for stroke *time is brain*, it has to fit tightly into the stroke protocol i.e. is restricted to a few minutes of runtime (Straka et al. (2010) name $5min$ as upper limit). With $6min$ (CH-Insel) and $20sec$ (DE-UzL), including all pre- and post-processing steps, the two winning methods fit the requirements, DE-UzL even leaving room for overhead.

7.2. Recommendations and limitations

New approaches for perfusion estimation should move away from simple methods (e.g. rule-based or thresholding). These are easy to apply, but our results indicate that they cannot capture the whole complexity of the problem. Machine learning, especially RFs, seem to be more suitable for the task: They can detect non-linear functional relationship between data and desired results that a simpler approach cannot. Domain knowledge is likely required to achieve state-of-the-art results, as the Tmax map thresholding of the two winning methods indicates. Evaluation should in any case be performed via a combination of suitable, quantitative measures. Simple volume difference or qualitative evaluation are of limited expressiveness and render the presented results incomparable. Where possible, the evaluation and training data should be publicly released. Finally, it has to be kept in mind that the segmentation task is a time-critical one and application times are always to be reported alongside the quantitative results.

The presented algorithms appear largely ready for clinical and/or research use, where they could prove their suitability in

practice. It remains to be shown that the diffusion lesion can be segmented equally well and whether the resulting perfusion-diffusion mismatch agrees with follow-up lesions.

SPES suffers from a few limitations: While MCA strokes are most common and well suited for mechanical reperfusion therapies (Kemmling et al., 2015), the restriction to low-noise MCA cases limits the result transfer to clinical routine. The generality of the results is additionally reduced by providing only single-center, single-ground truth data. Finally, voxel-sized errors in the ground truth prevented the evaluation of the HD, which would have provided additional information.

8. Conclusion

With ISLES, we provide an evaluation framework for the fair and direct comparison of current and future ischemic stroke lesion segmentation algorithms. To this end, we prepared and released well described, carefully selected annotated multi-spectral MRI datasets under a research license, developed a suitable ranking system and invited research groups from all over the world to participate. An extensive analysis of 21 state-of-the-art methods' results presented in this work allowed us to derive recommendations and to identify remaining challenges. We have shown that segmentation of acute perfusion lesions in MRI is feasible. The best methods for sub-acute lesion segmentation, on the other hand, still lack the accuracy and robustness required for an immediate employment. Second-center acquisition parameters and small lesions with weak FLAIR-support proved the main challenges. Overall, no type of segmentation algorithm was found to perform superior to the others, rather approaches using combinations of multiple methods and/or domain knowledge faired best.

The results of ISLES would have been more conclusive with a larger number of second- or even third-center cases and additional expert ground truth sets. For organizers of future challenges, we can formulate a number of recommendations: keep the set-up simple and well described to render participation as easy as possible; make aggressive advertisement on all available channels to reach interested groups; data preparation is the main work; defined ground truth creation rules carefully and a-priori; do not expect to release more than 50% of the acquired cases; use multi-center data; and avoid last minute changes if possible.

For the next version of ISLES, we would like to focus on the acute segmentation problem from a therapeutical point of view. By modeling a benchmark reflecting the time-critical decision making processes for cerebrovascular therapies, we hope to promote the transfer from methods to clinical routine and further the exchange between the disciplines. A multi-center dataset with hundreds of cases will allow the participants to develop complex solutions.

Appendix A. Participating algorithms

This section includes short descriptions of the participating algorithms. For a more detailed description of the participating Methods, please refer to the workshop's postproceeding volume (Crimi et al., 2016) or the challenge proceedings (Maier et al., 2015a).

Use abbreviations are: white matter (WM), gray matter (GM), cerebral spinal fluid (CSF), random forest (RF), extremely randomized trees (ET), contextual clustering (CC), gaussian mixture models (GMM), convolutional neural network (CNN), Markov Random Field (MRF), Conditional Random Field (CRF) and expectation maximization (EM).

Appendix A.1. *UK-Imp1*

We propose a multi-scale patch-based random forest algorithm for sub-acute stroke lesion segmentation. In the first step, we perform an intensity normalization under the exclusion of outliers. Secondly, we extract features from all images: Patch-wise intensities of each modality are extracted at multiple scales obtained with Gaussian smoothing. We parcellate the whole brain into three parts, including top, middle, and bottom. To keep an equilibrated class balance in the training set, only a subset of background patches is samples from locations all over the brain. Subsequently, we train three standard RF (Breiman, 2001) classifiers based on the patches selected from three parts of the brain. Finally, we perform some postprocessing operations, including smoothing the outputs of the RFs, applying a threshold, and performing some morphological operations to obtain the binary lesion map.

Appendix A.2. *DE-Dkfz*

The basic idea of this approach is that a single classifier might not be able to learn all possible appearances of stroke lesions. We therefore use 'Input-Data Adaptive Learning' to train an individual classifier for every input image. The learning is done in two steps: First, we learn the similarity between two images to be able to find similar images for unseen data. We define the similarity between two images as the DC that can be achieved by a classifier trained on the first image with the second image. Neighborhood Approximation Forests (NAF) (Konukoglu et al., 2013) are used to predict similar images for images without a ground-truth label (e.g. without the possibility to calculate the DC). We use first-order statistic description of the complete images as features for the learning algorithm. While the first step is done offline, the second step is done online, when a new and unlabeled image should be segmented. A specific, voxel-wise classifier is trained from the closest three images, selected by the previous trained NAF. For the voxel classifier we use ETs (Geurts et al., 2006) which incorporate DALSA to show the general applicability of our approach (Goetz et al., 2016). In addition to the intensity values we use Gaussian, Difference of Gaussian, Laplacian of Gaussian (3 directions), and Hessian of Gaussian with Gaussian sigmas of $1, 2, 3mm$ for every modality, leading to 82 features per voxel.

Appendix A.3. *FI-Hus*

The method performs lesion segmentation with a RF algorithm and subsequent CC (Salli et al., 2001). We utilize the

training data to build statistical templates and use them for calculation of individual voxel-wise differences from the voxel-wise cross-subject mean. First, all image volumes are warped to a common template space using Advanced Normalization Tools (ANTS). Mean and standard deviation over subjects are calculated voxel-by-voxel, separately for T1, T2, FLAIR and DWI images; these constitute the statistical templates. The initial lesion segmentation is calculated using RF classification and 16 image features. The features include normalized voxel intensity, spatially filtered voxel intensity, intensity deviation from the mean specified by the template, and voxel-wise asymmetry in intensities across hemispheres, calculated separately for each imaging sequence. For RF training, we only use a random subset of voxels in order to decrease computational time and avoid classifier overfitting, As a last phase, the lesion probability maps given by the RF classifier are subjected to CC to spatially regularize the segmentation. The CC algorithm takes the neighborhood of each voxel into account by using a Markov random field prior and iterated conditional modes algorithm.

Appendix A.4. *CA-McGill*

Description not provided.

Appendix A.5. *UK-Imp2*

We developed an automatic segmentation system, based on a 11-layers deep, multi-scale, 3D CNN. The network classifies voxels after processing a multi-modal 3D patch around them. To achieve efficient processing of greater image context, we developed a network architecture with two parallel convolutional pathways that processes the image at different scales. To train our system we build upon the work in Urban et al. (2014) and form batches with large image segments, equally sampled from the two classes. We exploit our network's fully convolutional nature to densely train on multiple voxels in the central part of the segments. By utilizing small 3^3 kernels that lead to deeper architectures with less trainable parameters, as well as adopting Dropout, Batch Normalization (Ioffe and Szegedy, 2015) and augmenting the database using reflection along the sagittal axis, we heavily regularize our network and show that it is possible to train such a deep and wide network on a limited database. Training our CNN takes approximately one day on a GeForce GTX Titan Black, while inference on a brain volume requires 3 minutes. We applied only minimum preprocessing, normalizing the modalities of each patient to zero mean and unit variance. For our final submission in the testing phase of the challenge, the outputs of 3 similar CNNs were averaged, to reduce noise caused by randomness during training. Additionally, we implemented a 3D, densely connected CRF by extending the work of Krähenbühl and Koltun (2012), which can efficiently postprocess a multi-modal scan in 2 minutes. Finally, connected components smaller than 20 voxels are eliminated.

Appendix A.6. *US-Jhu*

As rigid registration may not correct local differences between spatial locations across sequences, we re-register images to the FLAIR using Symmetric Normalization (Avants et al., 2008). We normalize the voxel intensities to a z-score using the 20% trimmed mean and standard deviation from each image. To train an algorithm, we create a series of predictors, including the x-y flipped voxel intensity, local moments (mean, sd, skew, kurtosis), and the images smoothed with large Gaussian filters. We trained a RF from 9 images, downsampled to 300,000 voxels, with the manual segmentation as the outcome (Breiman, 2001). From the RF, we obtained the probability of lesion and determined the threshold for these probabilities using the out-of-sample voxels from the training images, optimizing for the DC.

Appendix A.7. *SE-Cth*

The proposed framework takes the multi-spectral MRI brain images as input and includes two preprocessing steps: (1) Correction of bias field using the N3 bias field correction algorithm (Sled et al., 1998) and (2) normalization of the intensity values of each MRI modality to the interval $[0, 1]$, done by applying linear histogram stretching. For each voxel of multi-spectral MRI images, the following set of meaningful features is extracted: intensities, smooth intensities, median intensities, gradient, magnitude of the gradient and local entropy. All these features were normalized to zero mean and unit deviation. These features are then employed to train the RF (Criminisi and Shotton, 2013) classifier and segment the sub-acute ischemic stroke lesion. In this work, we set the RF parameters to: number of trees=150 and depth of each tree=50. A total of 999,000 data samples (i.e. 37,000 randomly selected from each training case) is used to train the RF classifier. Finally, the postprocessing is performed using dilation and erosion operations in order to remove small objects falsely classified as stroke lesion.

Appendix A.8. *US-Odu*

This work proposes fully automatic ischemic stroke lesion segmentation in multispectral brain MRI by innovating on our prior brain tumor segmentation work (Reza and Iftekharuddin, 2014). The method starts with the standard MRI preprocessing steps: intensity inhomogeneity correction and normalization. Next step involves two primary sets of feature extraction from T1, T2, FLAIR and DWI imaging sequences. The first set of features includes the pixel intensities ($I_{FL}, I_{T1}, I_{T2}, I_{DWI}$) and differences of intensities ($d_1 = I_{FL} - I_{T1}, d_2 = I_{FL} - I_{T2}, d_3 = I_{FL} - I_{DWI}$) that represents the global characteristics of brain tissues. In the second set, local texture features such as piecewise triangular prism surface area, multi-fractal Brownian motion (Islam et al., 2013) and structure tensor based local gradients are extracted to capture the surface variation of the brain tissues. We use a mutual information based implementation of minimum redundancy maximum relevance feature ranking technique and choose the 19 top ranked features. A classical RF classifier is employed to classify the brain tissues as lesion or background. Finally, a binary morphological filter is used to reduce the false positives from the original detections. We observe a few remaining false positives that compromise the overall performance. Our future works will include the study of more effective features, sophisticated feature selection techniques and an effective false positive reduction technique.

Appendix A.9. *TW-Ntust*

A fully automatic machine learning based stroke lesion three-dimensions segmentation system is built, which consists of a feature selection method, a multi-level RF model and a simple 3D registration approach. Only the FLAIR sequence was used and 275 features, which can be categorized into 24 types, are extracted for building RF models. To deal with the three dimensional data, a multi-RF model is developed and for stacks of five slices in the Z direction, a random forest model is built. The RF model generates probability maps. After obtaining the potential candidates from the RFs, we build a three-dimensional registration framework with backward and forward searching (Wang et al., 2015). It is applied to generate optimal three-dimensional predictions and too remove larger outliers. The system finds the largest object among all stacks and uses the stack with the largest object as the referenced stack. Then, the system performs backward and forward registration to maintain spatial consistency and remove the objects with no overlap to the detected objects in the neighboring stacks.

Appendix A.10. *CN-Neu*

We propose a framework to automatically extract ischemic lesions from multi-spectral MRI images. We suppose that the input images of different modalities have already been rigidly registered in the same coordinate system and non-brain tissues have already been removed from the images (Gao et al., 2014). Lesion segmentation is then performed by the proposed framework in three major steps: 1) preliminary segmentation, 2) segmentation fusion, and 3) boundary refinement. No training data is needed and no preprocessing and postprocessing steps involved. In the proposed framework, MRI images of each modality are first segmented into brain tissues (WM, GM and CSF) and ischemic lesions by weighting suppressed fuzzy c-means. Preliminary lesion segmentation results are then fused among all the imaging modalities by majority voting. The judge rule is that candidate voxels are regarded as lesions only if 1) they are considered as brain lesions in FLAIR images, and 2) they are viewed as brain lesions in more than 1 imaging modality beside FLAIR. The fused segmentation results are finally refined by a three phase level set method. The level set formulation is defined on multi-spectral images with the capability of dealing with intensity inhomogeneities (Feng et al., 2013).

Appendix A.11. *BE-Kul1*

We present a fully-automated generative method that can be applied to individual patient images without need for a training data set. An EM-approach is used for estimating intensity models (GMMs) for both normal and pathological tissue. The segmentation is represented by a level-set that is iteratively updated to label voxels as either normal or pathological, based on which intensity model explains the voxels' intensity the best. A convex level-set formulation is adopted (Goldstein et al., 2009), that eliminates the need for manual initialization of the the level-set. For each iteration to update the level-set, a full EM-estimation of the GMM parameters is done.

As a preprocessing step, spatial priors of WM, GM and CSF are non-rigidly registered to the patient image. The prior information is relaxed by smoothing the spatial priors with a Gaussian kernel. For SPES, we make use of the T2-weighted and TTP-weighted MR images and for SISS the diffusion weighted and FLAIR-weighted MR images. For SPES, the modalities are used in a completely multivariate way, i.e. with bivariate Gaussian models. For SISS, the modalities are segmented separately and a voxel is only labeled as lesion if it is a lesion in both modalities.

Appendix A.12. *CA-USher*

We propose a fully-automatic CNN approach which is accurate while also being computationally efficient, a balance that existing methods have struggled to achieve. We approach the problem by solving it slice by slice from the axial view. The segmentation problem is then treated by predicting the label of the center of all the overlapping patches. We propose an architecture with two pathways: one which focuses on small details of the tissues and one focusing on the larger context. We also propose a two-phase patch-wise training procedure allowing us to train models in a few hours and to account for the imbalanced classes. We first train the model with a balanced dataset which allows us to learn features impartial to the distribution of classes. We then train the second phase by only training on the classification layer with a distribution closer to the ground truth's. This way we learn good features and introduce the correct class prior to the model. Fully exploiting the convolutional nature of our model also allows to segment a complete brain image in 25 seconds. To test the ability of CNNs to learn useful features from scratch, we employ only minimal preprocessing. We truncate the 1% highest and lowest intensities and applied N4ITK bias correction. The input data is then normalized by subtracting the channel mean and dividing by its standard deviation. A postprocessing method based on connected components is also implemented to remove small blobs which might appear in the predictions.

Appendix A.13. *DE-UzL*

We propose a novel voxel-wise RF classification method with features chosen to model a human observers discriminative criteria when segmenting a brain lesion. They are based on intensity, hemispheric difference, local histograms and center distances as detailed in (Maier et al., 2015c, 2016). First, the already co-registered, isotropic voxel-spacing and skull-stripped sequences are preprocessed with bias field correction and intensity range standardization (Maier, 2016) (SISS) resp. the Tmax capped at $10s$ (SPES). A total of $1,000,000$ voxels are randomly sampled, keeping each case's class ratio intact (i.e. imbalanced). With this training set, 50 trees are trained using Gini impurity and $\sqrt{163}$ features for node optimization. For SISS, the a-posteriori forest probability map is thresholded at 0.4 and objects smaller than $1ml$ removed. For SPES, the threshold is 0.35 and only the largest connected component is kept. Both are followed by an hole closing in sagittal slices. The proposed method was equally successfully applied to BRATS challenge

data (Maier et al., 2016), underlining the generality of our approach.

Appendix A.14. BE-Kul2

A single segmentation method for both the SISS and SPES sub-challenges is proposed (Robben et al., 2016). First, all data is preprocessed, including bias-field correction, linear intensity standardization, and affine registration to MNI space. Then, each voxel is probabilistically classified as lesion or background within the native image space. The classifier consists of 3 cascaded levels, in which each level extends the feature set and uses a more complex extremely randomized forest (Geurts et al., 2006). The first level only uses the T1 intensity. The second level uses all modalities, smoothed in a local neighborhood at different radii, as well as voxel coordinates in atlas space. The third level additionally uses the probabilities estimated in level 2, smoothed locally. Classifier hyperparameters were tuned using 5-fold cross-validation. Testing data is preprocessed similarly and the voxelwise probabilities are predicted by the classifier. A technique to select the threshold that optimizes the DC is presented and applied to the predicted probability map in order to obtain the final binary segmentation.

Appendix A.15. DE-Ukf

In almost all cases of acute embolic anterior circulation stroke only one hemisphere is affected. We exploit this fact to (i) restrict the segmentation to only the affected hemisphere and (ii) to preselect the potential lesion by comparing local histograms of the affected side with the contralateral counterpart used as reference. Our approach is based on the evaluation of just the Tmax and ADC-maps. First, we automatically find the plane which separates the left and right hemisphere by co-registration with a mirrored Tmax-image, and identify the affected hemisphere as the one with the higher median value. For each voxel at position $\vec{x}$, a normalized, regional histogram $H(\vec{x}, t_i)$ is calculated in a $20\times20\times12\text{mm}^3$ neighborhood with a bin-width of $t_{i+1}-t_i = 1.5$s. The difference to the corresponding contralateral histogram $\tilde{H}(\vec{x}, t_i)$, taken from the mirrored part of the brain is calculated via $D(\vec{x}) = 1/2\sum_i |H(\vec{x}, t_i) - \tilde{H}(\vec{x}, t_i)|$. The resulting map of histogram differences is thresholded by 0.5 to find the regions with unusual Tmax values. This preselection is thresholded with the generally accepted value of Tmax > 6s. The histogram neighborhood size and the morphological operation parameters are globally fine-tuned based on the training dataset. To clean the mask, morphological erosion and dilation is applied. Finally, the segmentation is multiplied with ADC > $1700\text{mm}^2/\text{s}$ to remove CSF voxels.

Appendix A.16. CH-Insel

The model is trained only using data from the SPES dataset; no additional data is used. The method makes use of all seven imaging modalities. Before learning takes place, the following preprocessing steps are employed: TMax values are censored below zero and above 100, and all imaging modalities are then scaled to lie in the interval $[0, 256]$. Simple image texture features, based on those first used in Porz et al. (2014) are extracted from each imaging modality. The resulting data points are used to train a decision forest model which assigns to each volume element a label indicating if it should be considered part of the perfusion lesion. The training algorithm is a modification of RF (Breiman, 2001), in which bootstrapping of the training data is performed first at the patient level, and only then at the voxel level. This avoids the effects of patient-level clustering and leads to out-of-sample patients. This out-of-sample data is then used to empirically discover a threshold at which the DC of the segmentation is maximized, avoiding the need for holding out training data to tune the classifier. After segmenting with this threshold, no further postprocessing was applied. The method takes approximately six minutes to segment a new case.

Appendix B. Ranking schema

Our ranking system builds on the concept that a rank reveals only the direction of a relationship between two items (i.e. higher, lower, equal), but not its magnitude. After obtaining from each participating team the segmentation results for each case, the following steps are executed:

1. Compute the DC, ASSD & HD values for each case
2. Establish each team's rank for DC, ASSD & HD separately for each case
3. Compute the mean rank over all three evaluation measures/case to obtain the team's rank for the case
4. Compute the mean over all case-specific ranks to obtain the team's final rank

Graphically, the schema looks like displayed in Fig. B.11.

The outcome of the procedure is a final rank (real number) for each participant, which defines its standing in the leaderboard relative to all others. For SISS, with two ground truth sets for the testing dataset, their respective final ranks are averaged. For SPES, only the DC and the ASSD were used.

This approach can be applied to any number of measures, independent of their range, type or direction. Its outcome denotes only the differences between algorithms and hence serves its purpose. For any interpretation of the results, the distinct evaluation measure values obtained have to be considered, too.

A challenge with winners requires an absolute ranking, an ongoing benchmark not. For the online, ongoing leaderboard, the rank is not computed, rather each user is invited to sort the result table according to their favorite evaluation measure.

Failed cases and resolving ties. In one step of our algorithm, we have to rank the performance of each team on one case regarding a single evaluation metric. Such a situation can lead to ties, which have to be handled specially. We chose to decorate both tied teams with the upper rank and leaving the following empty (see Tab. B.10 for an example).

This behavior has an interesting effect for very difficult cases, where most teams fail to produce a valid segmentation, as can be seen in the example of Tab. B.11.

Thus, difficult cases do not alter the mean, as they would do when simply averaging e.g. the DC values over all cases.

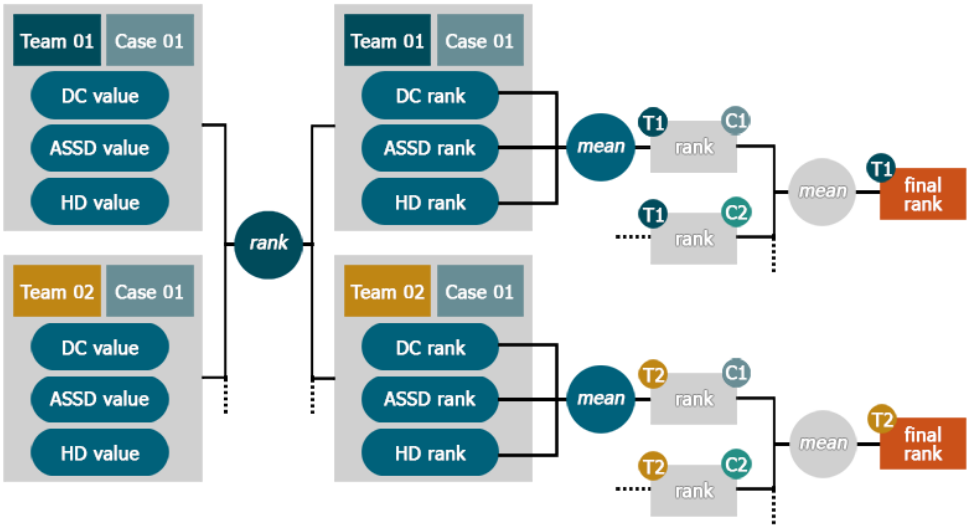


Figure B.11: Ranking schema as employed in the ISLES challenge.

Table B.10: Example of resolving ties for ISLES.

Team	DC
T-A	0.33
T-B	0.33
T-C	0.50
T-D	0.33
T-E	0.31

(a) Before...

Rank	Team
1	T-C
2	T-A, T-B, T-D
3	
4	
5	T-E

(b) ...after.

Table B.11: Tie resolving for difficult cases.

Team	DC
T-A	0.00
T-B	0.00
T-C	0.10
T-D	0.00
T-E	0.00

(a) Before...

Rank	Team
1	T-C
2	T-A, T-B, T-D, T-E
3	
4	
5	

(b) ...after.

Instead, only the performance relative to all other algorithms is compared, resulting in a more expressive ranking.

Beside resolving ties, we decided to introduce a concept of failed cases: When faced with (1) a missing segmentation mask or (2) a DC value of 0.00 (i.e. no overlap at all), the concerned case has been declared failed and all metric evaluation values subsequently set to infinity. Combined with the employed ranking approach and above described treatment of ties, this allows to incorporate missing segmentations in the ranking in a natural and fair manner. It could be argued that a DC of 0.00 could well mean that another part of the brain has been segmented. But the case has nevertheless to be considered a failed one, as the target structure has not been detected. Not declaring the case a failure would lead methods submitting a single random voxel segmentation to be ranked higher than an empty segmentation mask.

Acknowledgements

CN-Neu

This work was supported by the Fundamental Research Funds for the Central Universities of China under grant N140403006 and the Postdoctoral Scientific Research Funds of Northeastern University under grant No. 20150310.

References

Albers, G.W., Thijs, V.N., Wechsler, L.R., et al., 2006. Magnetic resonance imaging profiles predict clinical response to early reperfusion: the diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE) study. Annals of Neurology 60, 508–17.

Artzi, M., Aizenstein, O., Jonas-Kimchi, T., et al., 2013. FLAIR lesion segmentation: application in patients with brain tumors and acute ischemic stroke. European Journal of Radiology 82, 1512–8.

Avants, B.B., Epstein, C., Grossman, M., Gee, J., 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26–41.

Bauer, S., Fejes, T., Reyes, M., 2013. A Skull-Stripping Filter for ITK. Insight Journal .

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Christensen, S., Campbell, B.C., de la Ossa, N.P., et al., 2010. Optimal Perfusion Thresholds for Prediction of Tissue Destined for Infarction in the Combined EPITHET and DEFUSE Dataset, in: International Stroke Conference.

Crimi, A., Maier, O., Menze, B., Reyes, M., Handels, H. (Eds.), 2016. LNCS Brainlesion: Glioma, MS, Stroke and Traumatic Brain Injuries - First International BrainLes Workshop MICCAI 2015. Springer.

Criminisi, A., Shotton, J. (Eds.), 2013. Decision forests for computer vision and medical image analysis. Springer.

Dastidar, P., Heinonen, T., Ahonen, J.P., Jehkonen, M., Molnár, G., 2000. Volumetric measurements of right cerebral hemisphere infarction: use of a semiautomatic MRI segmentation technique. Computers in Biology and Medicine 30, 41–54.

Derntl, A., Plant, C., Gruber, P., et al., 2015. Stroke Lesion Segmentation using a Probabilistic Atlas of Cerebral Vascular Territories, in: Crimi, A., Maier, O., Menze, B., Reyes, M., Handels, H. (Eds.), LNCS Brainlesion: Glioma, MS, Stroke and Traumatic Brain Injuries - First International BrainLes Workshop MICCAI 2015, Springer Berlin Heidelberg. p. accepted.

Feng, C., Li, C., Zhao, D., Davatzikos, C., Litt, H., 2013. Segmentation of the left ventricle using distance regularized two-layer level set approach., in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 477–84.

Fiez, J.A., Damasio, H., Grabowski, T.J., 2000. Lesion segmentation and manual warping to a reference brain: intra- and interobserver reliability. Human brain mapping 9, 192–211.

Forbes, F., Doyle, S., Garcia-Lorenzo, D., Barillot, C., Dojat, M., 2010. Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation, in: IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), IEEE. pp. 69–72.

Forkert, N.D., Kaesemann, P., Treszl, A., et al., 2013. Comparison of 10 TTP and Tmax estimation techniques for MR perfusion-diffusion mismatch quantification in acute stroke. American Journal of Neuroradiology 34, 1697–703.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F., 2013. Dynamic classifier selection for One-vs-One strategy: Avoiding non-competent classifiers. Pattern Recognition 46, 3412–3424.

Gao, J., Li, C., Feng, C., et al., 2014. Non-locally regularized segmentation of multiple sclerosis lesion from multi-channel MRI data. Magnetic Resonance Imaging 32, 1058–66.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning 63, 3–42.

Ghosh, N., Sun, Y., Bhanu, B., Ashwal, S., Obenaus, A., 2014. Automated detection of brain abnormalities in neonatal hypoxia ischemic injury from MR images. Medical Image Analysis 18, 1059–69.

Goetz, M., Weber, C., Binczyk, F., et al., 2016. DALSA: Domain Adaptation for Supervised Learning From Sparsely Annotated MR Images. IEEE Transactions on Medical Imaging 35, 184–96.

Goldstein, T., Bresson, X., Osher, S., 2009. Geometric Applications of the Split Bregman Method: Segmentation and Surface Reconstruction. Journal of Scientific Computing 45, 272–293.

González, R.G., Hirsch, J.A., Lev, M.H., Schaefer, P.W., Schwamm, L.H. (Eds.), 2011. Acute Ischemic Stroke - Imaging and Intervention. Springer, Berlin Heidelberg. 2 edition.

Han, X., Jovicich, J., Salat, D., et al., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. NeuroImage 32, 180–94.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift 1502.03167.

Islam, A., Reza, S.M.S., Iftekharuddin, K.M., 2013. Multifractal texture estimation for detection and segmentation of brain tumors. IEEE Transactions on Biomedical Engineering 60, 3204–15.

Išgum, I., Benders, M.J.N.L., Avants, B.B., et al., 2015. Evaluation of automatic neonatal brain segmentation algorithms: the NeoBrainS12 challenge. Medical Image Analysis 20, 135–51.

James, J.R., Yoder, K.K., Osuntokun, O., et al., 2006. A supervised method for calculating perfusion/diffusion mismatch volume in acute ischemic stroke. Computers in Biology and Medicine 36, 1268–87.

Jovicich, J., Czanner, S., Han, X., et al., 2009. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. NeuroImage 46, 177–92.

Kabir, Y., Dojat, M., Scherrer, B., Forbes, F., Garbay, C., 2007. Multimodal MRI segmentation of ischemic stroke lesions., in: IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 1595–8.

Kemmling, A., Flottmann, F., Forkert, N.D., et al., 2015. Multivariate dynamic prediction of ischemic infarction and tissue salvage as a function of time and degree of recanalization. Journal of Cerebral Blood Flow and Metabolism 35, 1397–405.

Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P., 2013. The virtual skeleton database: an open access repository for biomedical research and collaboration. Journal of Medical Internet Research 15, e245.

Konukoglu, E., Glocker, B., Zikic, D., Criminisi, A., 2013. Neighbourhood approximation using randomized forests. Medical Image Analysis 17, 790–804.

Krähenbühl, P., Koltun, V., 2012. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials 1210.5644.

Lansberg, M.G., Straka, M., Kemp, S., et al., 2012. MRI profile and response to endovascular reperfusion after stroke (DEFUSE 2): a prospective cohort study. The Lancet. Neurology 11, 860–7.

Li, W., Tian, J., Li, E., Dai, J., 2004. Robust unsupervised segmentation of infarct lesion from diffusion tensor MR images using multiscale statistical classification and partial volume voxel reclassification. NeuroImage 23, 1507–18.

Litjens, G., Toth, R., van de Ven, W., et al., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Medical Image Analysis 18, 359–73.

Mah, Y.H., Jager, R., Kennard, C., Husain, M., Nachev, P., 2014. A new method for automated high-dimensional lesion segmentation evaluated in vascular injury and applied to the human occipital lobe. Cortex 56, 51–63.

Maier, O., 2016. MedPy - Medical image processing in Python.

Maier, O., Reyes, M., Menze, B., Handels, H. (Eds.), 2015a. ISLES 2015: Ischemic Stroke Lesion Segmentation - Proceedings.

Maier, O., Schröder, C., Forkert, N.D., Martinetz, T., Handels, H., 2015b. Classifiers for Ischemic Stroke Lesion Segmentation: A Comparison Study. PLOS ONE 10, e0145118.

Maier, O., Wilms, M., von der Gablentz, J., Krämer, U.M., Handels, H., 2014. Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers, in: Aylward, S., Hadjiiski, L.M. (Eds.), SPIE Medical Imaging, International Society for Optics and Photonics. p. 903504.

Maier, O., Wilms, M., von der Gablentz, J., et al., 2015c. Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. Journal of Neuroscience Methods 240, 89–100.

Maier, O., Wilms, M., Handels, H., 2016. Image Features for Brain Lesion Segmentation Using Random Forests, in: Crimi, A., Maier, O., Menze, B., Reyes, M., Handels, H. (Eds.), LNCS Brainlesion: Glioma, MS, Stroke and Traumatic Brain Injuries - First International BrainLes Workshop MICCAI 2015, Springer Berlin Heidelberg. p. accepted.

Martel, A.L., Allder, S.J., Delay, G.S., Morgan, P.S., Moody, A.R., 1999. Measurement of Infarct Volume in Stroke Patients Using Adaptive Segmentation of Diffusion Weighted MR Images, in: Taylor, C., Colchester, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 22–31.

Menze, B.H., Jakab, A., Bauer, S., et al., 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 1993–2024.

Muda, A.F., Saad, N.M., Abu-Bakar, S.A.R., Muda, A.S., Abdullah, A.R., 2015. Brain lesion segmentation using fuzzy C-means on diffusion-weighted imaging. ARPN Journal of Engineering and Applied Sciences 10.

Murphy, K., 2011. Development and evaluation of automated image analysis techniques in thoracic CT. Ph.D. thesis. Utrecht University.

Murphy, K., van Ginneken, B., Reinhardt, J.M., et al., 2011. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. IEEE Transactions on Medical Imaging 30, 1901–20.

Nabizadeh, N., John, N.M., Wright, C., 2014. Histogram-based gravitational optimization algorithm on single MR modality for automatic brain lesion detection and segmentation. Expert Systems with Applications 41, 7820–7836.

Neumann, A.B., Jonsdottir, K.Y., Mouridsen, K., et al., 2009. Interrater agreement for final infarct MRI lesion delineation. Stroke 40, 3768–71.

Olivot, J.M., Mlynash, M., Thijs, V.N., et al., 2009a. Optimal Tmax threshold for predicting penumbral tissue in acute stroke. Stroke 40, 469–75.

Olivot, J.M., Mlynash, M., Zaharchuk, G., et al., 2009b. Perfusion MRI (Tmax and MTT) correlation with xenon CT cerebral blood flow in stroke patients. Neurology 72, 1140–5.

Petitjean, C., Zuluaga, M.A., Bai, W., et al., 2015. Right ventricle segmentation from cardiac MRI: a collation study. Medical Image Analysis 19, 187–202.

Porz, N., Bauer, S., Pica, A., et al., 2014. Multi-modal glioblastoma segmentation: man versus machine. PLOS ONE 9, e96873.

Prakash, K.N.B., Gupta, V., Bilello, M., Beauchamp, N.J., Nowinski, W.L., 2006. Identification, segmentation, and image property study of acute infarcts in diffusion-weighted images by using a probabilistic neural network and adaptive Gaussian mixture model. Academic Radiology 13, 1474–84.

Rekik, I., Allassonnière, S., Carpenter, T.K., Wardlaw, J.M., 2012. Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal. NeuroImage: Clinical 1, 164–78.

Reza, S.M.S., Iftekharuddin, K.M., 2014. Multi-fractal texture features for brain tumor and edema segmentation, in: Aylward, S., Hadjiiski, L.M. (Eds.), SPIE Medical Imaging, International Society for Optics and Photonics. p. 903503.

Robben, D., Christiaens, D., Rangarajan, J.R., et al., 2016. A Voxel-wise, Cascaded Classification Approach to Ischemic Stroke Lesion Segmentation, in: Crimi, A., Maier, O., Menze, B., Reyes, M., Handels, H. (Eds.), LNCS Brainlesion: Glioma, MS, Stroke and Traumatic Brain Injuries - First Inter-

national BrainLes Workshop MICCAI 2015, Springer. p. accepted.
Rudyanto, R.D., Kerkstra, S., van Rikxoort, E.M., et al., 2014. Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the VESSEL12 study. Medical Image Analysis 18, 1217–32.
S, M., Varma, R., Kishore, L.T., 2012. A novel framework for segmentation of stroke lesions in Diffusion Weighted MRI using multiple b-value data, in: International Conference on Pattern Recognition (ICPR), IEEE. pp. 3762–3765.
Saad, N.M., Abu-Bakar, S.A.R., Muda, S., Mokji, M.M., Salahuddin, L., 2011. Brain lesion segmentation of Diffusion-weighted MRI using gray level co-occurrence matrix, in: IEEE International Conference on Imaging Systems and Techniques (IST), IEEE. pp. 284–289.
Salli, E., Aronen, H.J., Savolainen, S., Korvenoja, A., Visa, A., 2001. Contextual clustering for analysis of functional MRI data. IEEE Transactions on Medical Imaging 20, 403–14.
Seghier, M.L., Ramlackhansingh, A., Crinion, J., Leff, A.P., Price, C.J., 2008. Lesion identification using unified segmentation-normalisation models and fuzzy clustering. NeuroImage 41, 1253–66.
Shattuck, D.W., Prasad, G., Mirza, M., Narr, K.L., Toga, A.W., 2009. Online resource for validation of brain segmentation methods. NeuroImage 45, 431–9.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17, 87–97.
Soltanian-Zadeh, H., Bagher-Ebadian, H., Ewing, J.R., et al., 2007. Multiparametric iterative self-organizing data analysis of ischemic lesions using pre- or post-Gd T1 MRI. Cerebrovascular Diseases 23, 91–102.
Straka, M., Albers, G.W., Bammer, R., 2010. Real-time diffusion-perfusion mismatch analysis in acute stroke. Journal of Magnetic Resonance Imaging 32, 1024–37.
Styner, M., Lee, J., Chin, B., et al., 2008. 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. The MIDAS Journal .
Takasawa, M., Jones, P.S., Guadagno, J.V., et al., 2008. How reliable is perfusion MR in acute stroke? Validation and determination of the penumbra threshold against quantitative PET. Stroke 39, 870–7.
Tobon-Gomez, C., De Craene, M., McLeod, K., et al., 2013. Benchmarking framework for myocardial tracking and deformation algorithms: an open access database. Medical Image Analysis 17, 632–48.
Tsai, J.Z., Peng, S.J., Chen, Y.W., et al., 2014. Automatic detection and quantification of acute cerebral infarct by fuzzy clustering and histographic characterization on diffusion weighted MR imaging and apparent diffusion coefficient map. BioMed Research International 2014, 13.
Urban, G., Bendszus, M., Hamprecht, F.A., Kleesiek, J., 2014. Multi-modal Brain Tumor Segmentation using Deep Convolutional Neural Networks, in: MICCAI BraTS (Brain Tumor Segmentation) Challenge. Proceedings, winning contribution, pp. 31–35.
Wang, C.W., Budiman Gosno, E., Li, Y.S., 2015. Fully automatic and robust 3D registration of serial-section microscopic images. Scientific Reports 5, 15051.
Wheeler, H.M., Mlynash, M., Inoue, M., et al., 2013. Early diffusion-weighted imaging and perfusion-weighted imaging lesion volumes forecast final infarct size in DEFUSE 2. Stroke 44, 681–5.
WHO, 2012. Cause-specific mortality - estimates for 2000-2012. Technical Report. WHO.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.
Woloszynski, T., Kurzynski, M., 2011. A probabilistic model of classifier competence for dynamic ensemble selection. Pattern Recognition 44, 2656–2668.