Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences

Oskar Maier[a,b,*], Matthias Wilms[a], Janina von der Gablentz[c], Ulrike M. Krämer[c], Thomas F. Münte[c], Heinz Handels[a]

[a] *Institute of Medical Informatics, University of Lübeck, Germany*
[b] *Graduate School for Computing in Medicine and Live Science, University of Lübeck, Germany*
[c] *Department of Neurology, University of Lübeck, Germany*

Abstract

Background: To analyse the relationship between structure and (dys-)function of the brain after stroke, accurate and repeatable segmentation of the lesion area in magnetic resonance (MR) images is required. Manual delineation, the current gold standard, is time consuming and suffers from high intra- and inter-observer differences.

New Method: A new approach is presented for the automatic and reproducible segmentation of sub-acute ischemic stroke lesions in MR images in the presence of other pathologies. The proposition is based on an Extra Tree forest framework for voxel-wise classification and mainly intensity derived image features are employed.

Results: A thorough investigation of multi-spectral variants, which combine the information from multiple MR sequences, finds the fluid attenuated inversion recovery sequence to be both required and sufficient for a good segmentation result. The accuracy can be further improved by adding features extracted from the T1-weighted and the diffusion weighted sequences. The use of other sequences is discouraged, as they impact negatively on the results.

Comparison with existing methods: Quantitative evaluation was carried out on 37 clinical cases. With a Dice coefficient of 0.65, the method outperforms

*Corresponding author. +49 451 500 5645
Email address: maier@imi.uni-luebeck.de (Oskar Maier)

earlier published methods.

Conclusions: The approach proves especially suitable to differentiate between new stroke and other white matter lesions based on the FLAIR sequence alone. This and the high overlap render it suitable for automatic screening of large databases of MR scans, e.g. for a subsequent neuropsychological investigation. Finally, each feature's importance is assessed in detail and the approach's statistical dependency on clinical and image characteristics is investigated.

Keywords: sub-acute ischemic stroke lesion, white matter lesion, segmentation, extra tree forest, random forest, multi-spectral MRI

---

1. Introduction

Cerebrovascular disease is the second most common cause of death in the world (Mathers et al., 2009) and even survivors often face grave physical and mental disabilities. Ischemic stroke is most common either due to local thrombosis, hemodynamic factors or embolic causes.

Diagnosis of stroke often involves the acquisition of brain magnetic resonance (MR) scans to assess the stroke's presence, location, extent, evolution and other factors. An automated method to locate, segment and quantify the lesion area could support the clinicians and render their findings more robust and reproducible.

Cognitive neuroscientists could equally benefit from automatic lesion segmentation. One neurosciencentific approach is based on lesion to symptom mapping, which is used to detect correlations between brain areas and cognitive functions by means of negative samples (Krämer et al., 2013). The approach often requires manual segmentation of pathologies such as stroke lesions in many MR volumes, a tedious and time consuming task, which is subject to high inter- and intra-observer variability (Fiez et al., 2000), diminishing the validity of any gained insight as well as the reproducibility.

Furthermore, a general applicable stroke lesion segmentation method would enable the fully automatic screening of large databases of routinely acquired

scans, possibly with an ensuing mapping of the lesion volume to functional and/or anatomical brain areas, to, for example, identify suitable study candidates.

Finally, automatic ischemic stroke lesion segmentation can help in quantitative studies aiming to answer the questions 'how an ischemic stroke evolves' and to discover the involved processes (Rekik et al., 2012).

1.1. Evolution and appearance of ischemic stroke lesions

Ischemic stroke lesion appearance in MR scans is not stable, as the lesion undergoes a number of phases. The hyper-acute phase commences with the total or partial loss of blood supply, followed by the acute (from 6 hours) and the sub-acute (from 24 hours) phases until the chronic phase (from 2 weeks) is reached (Gonzalez et al., 2006). During these phases a number of cell death mechanisms take place, changing the molecular composition of the lesion tissue and consequently its visibility in the MR sequences. Additionally, secondary effects such as swelling affect the anatomical layout of the brain. Our focus lies on the sub-acute phase, during which most MR scans are acquired in clinical context. During this phase, the lesions undergo changes and vary in appearance: They appear slightly to strongly hyperintense in fluid attenuated inversion recovery (FLAIR) and T2-weighted (T2w), hypointense in T1-weighted (T1w) and decreasingly hyperintense in diffusion weighted (DW) sequences. For a more detailed discussion of stroke evolution, appearance and the underlying processes, see Gonzalez et al. (2006) and Rekik et al. (2012).

1.2. Challenges of automatic ischemic stroke lesion segmentation

The gradual changes of stroke lesion appearance renders ischemic stroke lesion segmentation a challenging task. Furthermore, stroke lesions can appear in various sizes, shapes and in all areas of the brain, which makes it difficult to employ prior-based statistical methods. A crude description of the segmentation task would be 'to look for something unusual' in the images, but the anatomical brain layout differs between persons.

A recent overview by Rekik et al. (2012) identifies a number of common biological- and imaging-dependent challenges that have to be overcome: Any proposed method working with clinical routine scans has to be able to work with low quality images of varying resolutions and slice thicknesses, which might contain imaging artefacts (e.g. due to patient movement). Varying visibility of stroke lesions in MR sequences has to be considered, for example fogging in DW sequences (sometimes occurring isointense appearance of the lesions in DW scans around the second week) and the T2 shine through effect. Hence a multi-spectral approach is potentially more powerful than a mono-spectral approach, as it takes information from multiple MR sequences into account. As the lesion swells and shrinks during its evolution, it distorts adjacent tissue and causes anatomical changes such as ventricle enhancement and midline shift. Furthermore, stroke lesions can be topologically unconnected (termed multifocal lesions or embolic shower).

1.3. State of the art of ischemic stroke lesion segmentation

Segmentation of non-chronic ischemic stroke lesions has been summarized and discussed in Rekik et al. (2012). Of the 25 articles reviewed, the majority describe pixel-based (n=13) as opposed to image-based (n=9), atlas-guided (n=1) and deformable model (n=2) methods. Few present fully automatic approaches and none uses supervised training of a classifier.

Segmenting chronic differs from segmenting non-chronic stroke lesions, but similar methods can be employed in both cases. In the following, pixel-based methods for chronic stroke lesion segmentation are described. Agam et al. (2006) apply a mixture-parametric probabilistic model to diffusion tensor (DT), T1w and T2w images, but DT sequences are rarely acquired in routine stroke assessment. Another approach is taken by Seghier et al. (2008), who propose to search for chronic lesion candidates among atypical voxels not corresponding to the tissue type expected at the location, which can be realized with tissue probability maps. They applied fuzzy clustering to T1w images, which are not the most suitable to identify stroke lesions, and they experienced problems

with leakages into the ventricles. A Bayesian multi-sequence hidden Markov model with individual weights for the sequences, accounting for differences in their contributions, is proposed by Forbes et al. (2010). As they provide only results obtained on a single case, it is impossible to judge the general feasibility of their approach. Wilke et al. (2011) propose a semi-automatic as well as an automatic, partially atlas-based method employing 4 class fuzzy-clustering to segment chronic ischemic stroke lesions in T1w sequences. For this purpose, they calculate four prior-maps (tissue abnormality, tissue composition, Jacobian determinant and laterality map) on high quality image data and combine them to segment the lesion. This approach requires a potentially lossy registration between pathological and healthy brains and considerable user interaction to achieve good results.

A recent publication by Mitra et al. (2014) approached the problem of chronic lesion segmentation with a combination of Bayesian-Markov random fields and random decision forests (RDF) for voxel-wise classification in multi-spectral MR volumes. The authors report results obtained over a large evaluation dataset ($n = 36$), but they include secondary and other white matter lesions (WML) in their ground truth by thresholding their MR volumes, which removes the most challenging aspect of ischemic stroke lesion segmentation, namely the differentiation between stroke and other WMLs.

We previously approached the problem with support vector machines for voxel-wise classification (Maier et al., 2014), but these turned out difficult to tune and too sensitive to the selection of the training data.

All of the above articles except Mitra et al. (2014) have evaluated their approach on a small number of cases. Rekik et al. (2012) noted that a detailed description of the image data is required to assess the performance of any stroke lesion segmentation method in detail, as factors such as image quality and artefacts can have a strong impact on the segmentation quality. Even more important is the description of other pathologies such as WMLs and enclosed haemorrhages. These are common occurrences in stroke cases and pose a serious challenge.

1.4. Novel approach to automatic lesion segmentation

In this paper, we propose a novel, fully automatic method to segment subacute ischemic stroke lesions in MR scans based on Extra Tree (ET) forests. ET forest are robust to noise and uncertain training data (Geurts et al., 2006), both of which we are likely to face in our application. Features are extracted voxelwise from the MRI sequences and the trained ET forest is used to classify the voxels of a formerly unseen case into two classes: the non-stroke brain tissue and the stroke lesion. By using a machine learning method, we attempt to capture the complex, non-linear class borders in feature space that allow to separate lesion from non-lesion voxels.

Our method is developed to work on clinical data without special requirements on image quality, resolution or imaging artefacts. Furthermore, it is evaluated using various evaluation measures on 37 well described real cases, some of which contain other WMLs, older stroke lesions and/or haemorrhagic transformation. Thus we follow the recommendations formulated by Rekik et al. (2012) and focus on an application in cognitive neuroscience trials.

We apply ET forests to stroke lesion segmentation and obtain good results with easy to compute features. A thorough investigation of the proposed method is provided, including an analysis of the features. Furthermore, we examine the influence of different MR sequences on the results. Finally, a robust training set sampling strategy is presented and evaluated.

2. Materials

In the previous section, we have noted the importance of a detailed description of the cases used in the method's training and validation. In this section the employed data, its clinical and imaging characteristics and the ground truth are described.

2.1. Data

All 37 datasets have been used in two clinical studies (Machner et al., 2014, 2012; von der Gablentz, 2012) on spatial neglectcarried out at the Department

of Neurology, University Medical Center Schleswig-Holstein (UKSH), Campus Lübeck, Germany. The focus of these studies was on a specific neuropsychological intervention and the MR images were obtained as background information and had been acquired during routine clinical care for these patients. A 3T Achieva 3.0T TX scanner (Phillips, Amsterdam, Netherlands) was employed. The studies were approved by the local ethics committee (#12-064) and all patients gave informed written consent. Of the 37 patients, 17 were women. The mean age was 65.5 ± 15 years (range 26 years to 84 years).

Available sequences are T1w, T2w, FLAIR and DW. Additionally, an apparent diffusion coefficient (ADC) map was computed from the DW images. The FLAIR sequence was recorded for all patients, but only for a subset of $n = 15$ cases all of the five sequences had been acquired.

T1w Turbo Field Echo (TFE) sequences were acquired in axial direction with the following parameters: TR = 8.55 ms, TE = 4 ms, flip-angle = 8° and voxel spacing = $1 \times 1 \times 1$ mm. T2w Turbo Spin Echo (TSE) sequences were acquired in sagittal direction with the following parameters: TR = 3194 ms, TE = 100 ms, flip-angle = 90° and voxel spacing = $0.43 \times 0.43 \times 4.4$ mm for most cases. FLAIR sequences were acquired in sagittal direction with the following parameters: TR = 11 000 ms, TE = 125 ms, TI = 2800 ms, flip-angle = 90° and voxel spacing = $0.43 \times 0.43 \times 5.5$ mm or $0.9 \times 0.9 \times 5.5$ mm for most cases. DW sequences were acquired in sagittal direction with the following parameters: TR = 2365 ms, TE = 52 ms, flip-angle = 90° and voxel spacing = $0.9 \times 0.9 \times 5.5$ mm for most cases. For a detailed listing of the images' resolutions, we refer the supplementary materials.

2.2. Lesion characteristics

Lesion volume ranged from 1.76 to 344.17 ml with an average of 95.48 ± 95.88 ml. The mean lesion age calculated from available clinical data was 8.41 ± 4.32 days (range 1 to 22 days) and thus qualifies the lesions as being sub-acute. Besides the stroke, some of the patients also show other pathologies. Most display either a small or medium load of WMLs and $n = 18$ cases showed

haemorrhagic transformation of part of the stroke lesion area. A detailed listing can be found in the supplementary materials. Most lesions are middle cerebral artery (MCA) strokes, but some occurred in other brain areas. Fig. 1 shows the distribution of lesion voxels normalized to Montreal Neurological Institute (MNI) space and flipped to the same hemisphere. It can be seen that the lesions cover most of the anatomical structure of the brain.

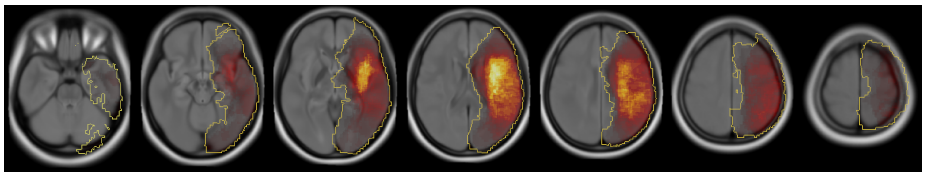


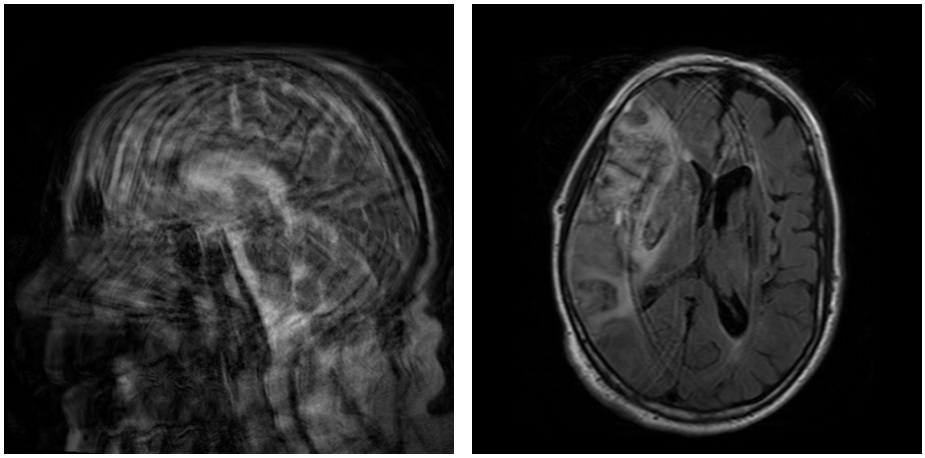
Figure 1: Distribution of lesion voxels over all 37 cases in MNI space in axial slices. The yellow/bright outline denotes the maximum extend. Note that all lesions have been flipped to the same hemisphere for this image.

2.3. Image quality

Acquisition of MR images of actual stroke patients can be difficult, as patients are often irritated or nervous (Rekik et al., 2012). Many of the image are therefore noisy, often with Gibbs and sometimes movement artefacts (see Fig. 2). This reflects that our data had been acquired at a stroke unit in routine settings. See the supplementary materials for complete list of image quality issues. All cases were used in the evaluation, regardless of the image artefacts.

2.4. Ground truth

We train our ET forest classifier with and evaluate our method against a ground truth segmentation delineated by a medical expert on the Flair sequence. Where required and available, other MR sequences were used to resolve ambiguities. The applied rules were to segment only the newest ischemic stroke lesion and to include completely encircled haemorrhages.



(a) Case 01/T2w: Movement artefacts in every second slice

(b) Case 03/Flair: Gibbs artefact

Figure 2: Examples of imaging artefacts in the datasets.

3. Methods

We propose a novel, fully automatic method based on Extra Trees for the segmentation of ischemic stroke lesions in MR images.

As all classification based methods, it can be divided into a training and an application phase. The training phase requires a number of expert-segmented cases to train the forest. For this purposes, the training cases are pre-processed (Sec. 3.1) and voxel-wise local image features are extracted (Sec. 3.3) resulting in a large number of training observations. From these, a training set is sampled randomly (Sec. 3.4) to train a forest of Extra Trees (Sec. 3.2). During the application phase a formerly unseen case is processed. First, it passes through the same pre-processing pipeline as the training cases. Then the same features are extracted and each voxel-wise sample passed to the trained ET forest for classification into either lesion or background. The such obtained preliminary lesion mask is further post-processed (Sec. 3.5), resulting in the final lesion segmentation. For a schematic overview of the complete processing pipeline see Fig. 3.

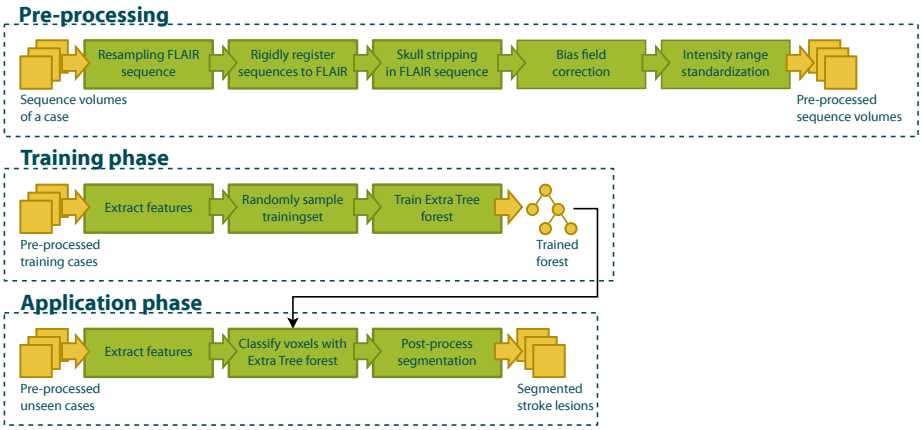


Figure 3: Schematic overview of the method's processing pipeline.

3.1. Pre-processing

A number of pre-processing steps were performed to prepare the training cases as well as new cases before the classification. All methods presented here are fully automatic and applied with a fixed set of parameters described in the experiment section 4.

In the first step, we resampled all FLAIR sequences to a common voxel spacing of $R = x \times y \times z$ mm. Then all remaining sequences were rigidly registered to their associated FLAIR image using the elastix image registration toolbox (Shamonin et al., 2014; Klein et al., 2010) to account for movement and other small errors. In the third step, the intracranial segmentation, we applied the Brain Extraction Tool (BET) proposed by Smith (2002) from the FMRIB Software Library (FSL) (Jenkinson et al., 2012) to the FLAIR sequence to extract the brain. In some cases the skull, eyes or neck are only partially removed, the cases were nevertheless kept. The bias fields of all sequences were corrected with the MR Intensity Bias Field Correction tool (Likar et al., 2001) from the Computational Morphometry Toolkit (CMTK) to increase intensity homogeneity over the tissue types. Finally, an intensity standardization algorithm based on the work of Nyúl et al. (2000) was applied to remove intensity scale differences between scans of the same MR sequence.

3.2. Extra Tree forests

Extra Tree forests are ensemble classifiers similar to RDFs which base on weaker classifiers.

Decision trees (Breiman et al., 1984) are a popular method and show a number of advantages over other classifiers: They are simple to understand and interpret, require practically no data preparation (such as e.g. normalization), use a white box model and are very fast, both in training and application. Furthermore, they are statistically robust (Breiman et al., 1984). On the downside, decision trees do not generalize well, thus are prone to overfitting (Breiman et al., 1984; Criminisi and Shotton, 2013) and show very high variance with respect to their bias (Geurts et al., 2006).

To address these shortcoming, decision forests have been proposed (Ho, 1998; Breiman, 2001; Criminisi and Shotton, 2013). They belong to the class of ensemble methods, which combine multiple weak classifiers, here classification trees, to obtain better predictive performance. A large number T of trees is grown from the same learning set S. Upon application, the new sample is pushed through each tree $t \in T$ and a majority vote decides on the final class membership. Alternatively, class membership probabilities can be obtained from the forests.

Since decision trees are grown deterministically, a notion of randomness has to be introduced, otherwise all trained trees would be equal. Breiman (1996) proposed to train each tree with a bootstrap replica of the training set. Shortly afterwards, Ho (1998) proposed to search for the optimal split not over all, but only K randomly selected features at each node. Finally, Breiman (2001) combined both of these methods, terming the approach random decision forests. For RDFs, two parameters are important: The number of trees T trained and the number of features K considered at each node. These popular classifiers have been shown to be very robust against noisy data as well as overfitting and Geurts et al. (2006) found them to strongly reduce the variance of the decision trees while only slightly increasing the bias. A number of empirical studies (see, among others, Criminisi and Shotton (2013)) found the decision forests to

outperform their predecessors and many other classifiers. In Mitra et al. (2014), RDFs were the classifier of choice.

Extra Trees or extremely randomized trees were introduced by Geurts et al. (2006) and add another layer of randomness to decision forests. The additional randomization step is introduced at the node during the tree training: Instead of searching for the optimal cut-point i.e. threshold τ, a random threshold value is selected for each feature. Subsequently, the search space is reduced, leading to faster training. On the down side, the forest size/depth is increased due to the introduction of sub-optimal cuts. From a theoretical point of view, the introduction of this additional randomness allows for a non-intuitive approach to the classification problem by increasing the chance to jump out of local minima at the cost of sub-optimal cuts. Geurts et al. (2006) have compared ET forests with RDFs and other tree-based classifiers and shown them to perform largely equal to or better than their competitors, including RDFs. Furthermore, they have established that ET forests lead to a further decrease in overall variance.

ET forests are suitable classifiers for our problem, as they are very robust to noise and uncertain training data. Our ground truth can contain errors, noise and image artefacts, which may lead to uncertainties.

3.3. Local image features

For each voxel a number of features are extracted. The information encoded in these features is the knowledge made available to the classifier during training and application. It is therefore important to select meaningful and discriminative features.

We decided on a number of mainly intensity based features to put the focus on the method and to show that a good segmentation can already be obtained with simple features. The default parameters for each feature are given in Sec. 4. All features are extracted from all available MR sequences.

Intensity feature: Depending on the MR sequence and the lesion age, stroke lesions often show up as hyper- or hypointensities. Therefore, we take the intensity value of each voxel after the pre-processing as our first feature f_{int}.

Weighted local mean feature: Many sequences are corrupted by noise or other image artefacts. The intensity value of a single voxel alone can thus be misleading and many false-positives and false-negatives are bound to occur. Therefore, we chose the weighted mean intensity value over a small area around each voxel, computed by a convolution with a Gaussian kernel of size σ mm for our second feature $f_{lmg,\sigma}$.

2D centre distance feature: There are some spatially locatable parts of the brain, such as the eyes and the ventricle, where no ischemic stroke lesions can occur. To improve robustness and to provide spatial information without running the risk of reducing generality, we select as third feature $f_{cd,\{axial,coronal,sagittal\}}$ each voxels Euclidean distance to the presumed brain centre in each of the three 2D views (axial, coronal and sagittal). Since the true brain centre is not known, the centre of the image is used as rough approximation.

Local histogram feature: Ischemic stroke lesions are largely inhomogeneous in appearance and incorporate a wide range of intensities that often overlap with healthy tissue types. Assuming that the local distribution of intensity values differs between lesioned and healthy brain tissue, we model the intensity distribution in a cubical area of r^3 mm^3 through a normalized histogram with b bins. The histogram range is set to the whole image's intensity range. The number of bins should be selected large enough to be able to model a large number of possible distributions while being small enough to not inflate the feature vector; and the neighbourhood patch size r^3 has to be selected to fit smoothly inside the smallest lesions. Thus, the feature is defined by two parameters as $f_{lh,b,r}$.

3.4. Sampling the training set

A careful selection of the training samples is a crucial step for all learning based methods. Depending on the image resolution, around $128^3 = 2,097,152$ samples can be extracted per training case. This abundance of training samples lead us to sample only a subset of all available training samples into the training set. We keep the inherent lesion to background ratio of each case intact to introduce the prevalence of background voxels to the classification process (which

differs from the widespread approach to use samples equally distributed over the classes). Hence, we employ stratified random sampling, which corresponds to uniform random sampling while keeping the class ratio intact, to extract a number of N_S samples from each case. For the exact value chosen for this parameter, please refer to the experiment section 4.

3.5. Post-processing

The only post-processing step applied to the classification result is the closing of holes and the removal of small binary objects which are unlikely to represent ischemic stroke lesions but rather constitute false positives. These have been noticed to appear in the skull and neck areas when the intracranial segmentation failed partially. The size of the threshold t_s has to be chosen carefully to avoid deleting true positives.

4. Experiments

In this section, the evaluation scheme, the parameter values and the experiments are described in detail.

4.1. Evaluation scheme

The evaluation of our method was performed using leave-one-out cross-validation. The resulting segmentations were compared to the expert ground truth (which has been resampled to the working resolution R beforehand).

We employed a number of evaluation measures to highlight different aspects of our results: Dice's coefficient (DC) rates the overlap between two volumes. A value of 1 denotes a perfect fit, while a value of 0 signifies no overlap. The DC is known to be affected by the size of the rated volumes, where larger objects receive higher DC values than smaller objects for a comparable segmentation quality. Therefore, we furthermore computed the average symmetric surface distance (ASSD, mm), which gives a measure of how well the volumes' surfaces fit. Outliers, which are an important aspect of a segmentation's quality, are neither reflected by DC nor the ASSD. We thus introduced the Hausdorff distance

(HD, mm) as a measure of the maximum error. Finally, precision and recall (both in range $[0, 1]$) are provided. The higher the precision compared to the recall, the stronger the under-segmentation and vice-versa.

4.2. Parameter configuration

Our proposed method allows for a number of settings and depends on some parameters. Three different configurations were used during the experiments: By default, a mono-spectral approach was chosen, that employs only the FLAIR sequence and was evaluated over all 37 cases. All experiments involving more MR sequences (multi-spectral) used the same set of default parameters, but were trained and evaluated over the subset of 15 cases, for which all five sequences are available. For the third configuration, the method's parameters and the utilised sequences were tuned for maximum accuracy to assess the proposed methods optimal performance.

The parameters values, chosen heuristically and not tuned, were as follows: For resampling (see Sec.3.1), an isotropic spacing of $R = 3$ mm^3 was selected. A total of $N_S = 250,000$ samples were drawn over all training cases (see Sec. 3.4) and the following features extracted for each sampled voxel: 1. intensity value f_{int}; $2-4$. weighted local mean with $\sigma \in 3, 5, 7$ mm $f_{lmg,\{3,5,7\}}$; $5-7$. the distance in mm to the image centre in axial, coronal and saggital views $f_{cd,\{axial,coronal,sagittal\}}$; $8-40$. the local histogram (11 elements feature) of an area of size 5^3, 10^3 and 15^3 mm around each voxel $f_{lh,11,\{5,10,15\}}$. That makes to a total of 40 features. The once sampled training set was used in all experiments to allow for a sound comparisons independent of any variance introduced by the random sampling.

For ET forest training, the training set was resampled using bootstrapping to train a total of $T = 200$ trees. At each node, $K = 6$ ($\approx \sqrt{40}$) features were considered in the search for the best split, which was rated by its information gain. The trees were grown until a depth of 100 (which, for all practical purposes, corresponds to an unlimited growth) or until each leaf contained only samples of a single class, whichever occurred first. This implies that we grew our trees to

their maximum extent, neglecting all available measures to counter overfitting. The class probability maps produced by the forest were thresholded at $t = 0.5$.

After segmentation, the post-processing steps described in Sec. 3.5 were employed, including the removal of all binary structures of size <= 1.5 ml to reduce false positives.

The configuration of tuned parameters is described in the respective section 4.8 of the experiments.

4.3. Segmentation results

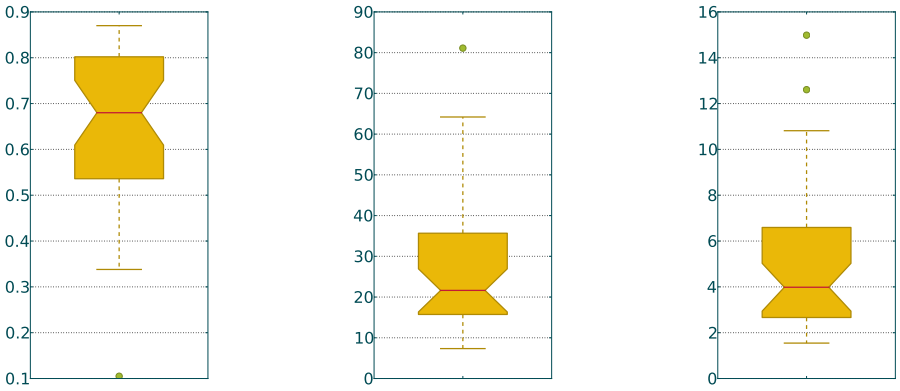
Employing the default mono-spectral configuration as laid out above, our proposed ischemic lesion segmentation method led to the segmentation results presented in Tab. 1. The boxplots for DC, HD and ASSD are shown in Fig. 4

	DC[0,1]	HD(mm)	ASSD(mm)	prec.	recall
average	0.65	28.61	5.02	0.83	0.58
std	0.18	17.82	3.37	0.17	0.21
median	0.68	21.63	3.71	0.86	0.59

Table 1: Average results obtained using the default mono-spectral configuration. Note that the failed cases 30 and 36 have been exempt from the computation.

and a detailed listing of the results for each case can be found in the supplementary materials. [*For online version insert Supplementary Table 1 here.*] Our approach failed in two cases to segment the stroke lesion. These two cases 30 and 36 were excluded from calculation of the evaluation measures for all further experiments (but remained in the training set).

As noted above, the evaluation was conducted in the resampled space R of the FLAIR sequence, which involves the potentially lossy resampling of the ground truth. Reversing the process and resampling the resulting segmentation to the original resolution of each case instead led to comparable results of $DM = 0.61$, $HD = 29$ mm, $ASSD = 4.83$ mm, $prec. = 0.80$ and $recall = 0.54$ on average.



(a) DC(0,1) (b) HD(mm) (c) ASSD(mm)

Figure 4: Boxplots of the evaluation measures with median line in red and outlier points in green. Cases 30 and 36 are exempt.

4.4. Multi-spectral results

To assess the superiority of multi- over mono-spectral approaches for stroke lesion segmentation in the sub-acute phase, we trained ET forest classifiers with all possible sequence combination of FLAIR, DW, ADC, T1 & T2 and evaluated them over the 15 cases of the second configuration. From each sequence the complete range of features described in Sec. 4.2 was extracted. The results are shown in Fig 5.

4.5. Feature combinations

We employed four different types of cardinal features to train our classifier, some of which might be redundant or irrelevant. Although forest ensemble methods are not known to suffer from overfitting due to the presence of redundant and irrelevant features, it is useful to perform a feature analysis to improve the model interpretability and possibly reduce the size of the feature vector. Fig. 6 displays the DC results obtained with ET forests trained with all possible combinations of our feature types using the first configuration.

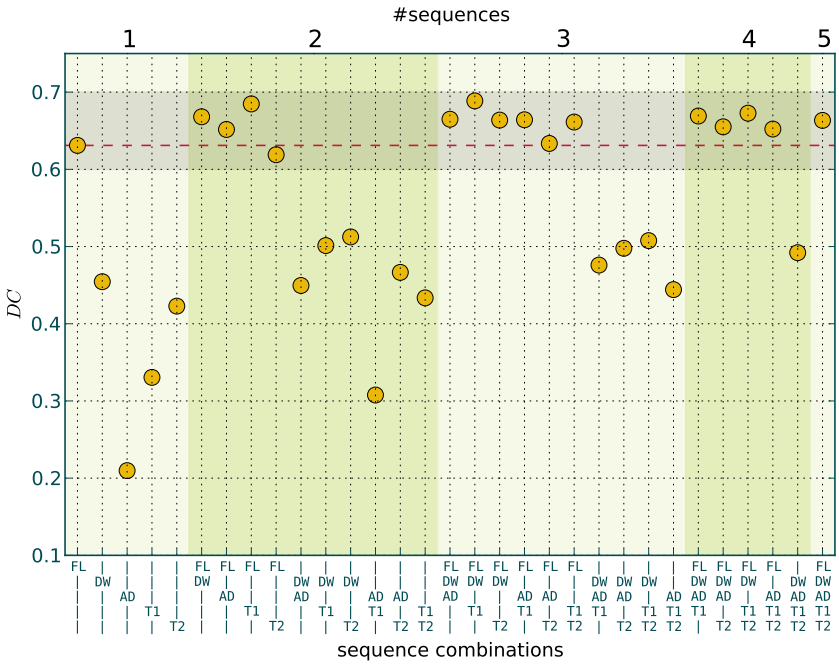


Figure 5: Average DC values obtained for all possible sequence combinations. The lower x-axis denotes the combination of sequences used in the corresponding results, where `FL`=FLAIR, `DW`=DW, `AD`=ADC, `T1`=T1w and `T2`=T2w, while the upper x-axis as well as the alternating background colours denote the changes in the number of sequences used. With the default settings we obtained a DC of ≈ 0.63, marked by the horizontal dotted line in the graph. The shaded horizontal band in the upper end of the graph is a visual aid to highlight the group of best combination, which all contain the FLAIR sequence.

4.6. Influence of the post-processing

During post-processing, we eliminated small outliers under the assumption that most ischemic stroke lesions are larger than 1.5 ml. Omitting this step, we obtained the mean results presented in Tab. 2.

4.7. Training set sampling

Our trainings set was randomly sampled from all available training cases as described in Sec. 3.4, which might have introduced a variance. To investigate the influence of this variance on the final segmentation quality and to examine the

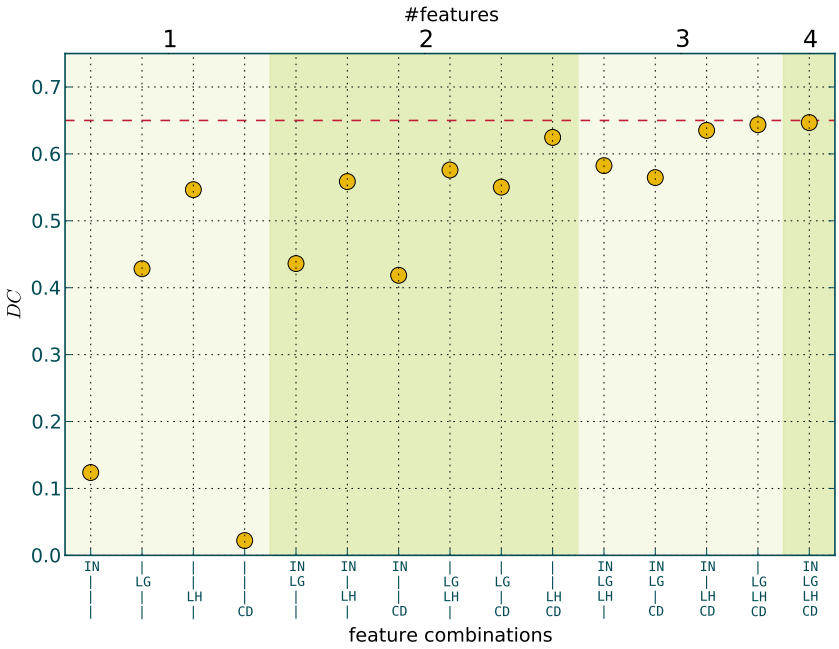


Figure 6: Average DC values obtained for all possible feature combinations. The lower x-axis denotes the combination of feature types used in the corresponding results, where `IN`=Intensity, `LG`=Weighted local mean, `LH`=Local histogram and `CD`=2D centre distance, while the upper x-axis as well as the alternating background colours denote the changes in the number of features used. With the default settings we obtained a DC of ≈ 0.65, marked by the horizontal dotted line in the graph.

robustness of the chosen sampling method, we repeated the process of sampling, training and evaluation 10 times. The average results and standard deviations obtained are $DM = 0.65 \pm 0.002$, $HD = 28 \pm 0.9$ mm, $ASSD = 4.96 \pm 0.08$ mm, $prec. = 0.82 \pm 0.004$ and $recall = 0.59 \pm 0.001$.

We chose to only use $250,000$ of the total $10,899,312$ training samples. It is possible, that a larger training set would lead to better results. Fig. 7 displays the resulting mean DC scores obtained with ET forests trained on sampling sets of different sizes.

Configuration	DC[0,1]	HD(mm)	ASSD(mm)
post-processing: on	0.65	29	5.02
post-processing: off	0.65	39	5.18

Table 2: Mean results obtained with and without post-processing.

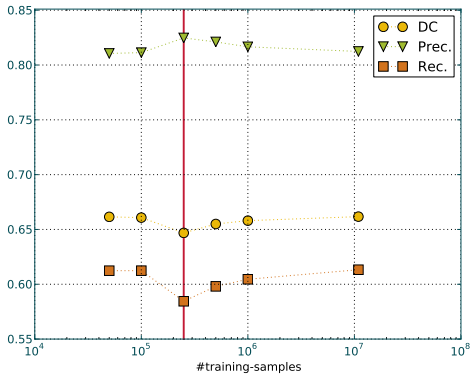


Figure 7: Influence of the training set size on the segmentation quality. The vertical line denotes the default value.

4.8. Parameter tuning

We also investigated the influence of the forest parameters on the segmentation quality (see supplementary materials). From these results and the findings on suitable feature and sequence combinations made above, we derived a set of ideal, tuned parameters. They differed from the default configuration used in the experiments in four aspects: (1) The forests were trained on the FLAIR, DW and T1w sequence; (2) all features were considered at each node i.e. K is set to the feature vector length; (3) *gini* rather than information gain was used to evaluate the quality of a split; and (4) all ten million available training samples were used. The results obtained with the tuned as well as with the default parameters on the 15 cases for which all MR sequences were available are presented in Tab. 3.

parameter set		DC[0,1]	HD(mm)	ASSD(mm)	prec.	recall
default	mean	0.63	32	5.60	0.75	0.63
	std	0.18	22	3.80	0.25	0.21
tuned	mean	0.70	25	3.99	0.78	0.69
	std	0.13	13	1.90	0.20	0.16

Table 3: Comparison of the default and the tuned set of parameters.

5. Discussion

Exempting the two failed cases, we obtained an average DC of 0.65 ± 0.18 for our method, which compares favourably with previously published results: Wilke et al. (2011) reported a DC of 0.6 for their semi-automatic and 0.49 for their automatic approach. Hevia-Montiel et al. (2007) reached 0.54 ± 0.18 and Seghier et al. (2008) even 0.64 ± 0.10, although only on eight real cases. In a recent publication Mitra et al. (2014) proposed a method with which they have achieved an average DC of 0.60 ± 0.13 and ASSD of 3.06 ± 3.17 mm. These comparisons have to be treated with care, as they have not been obtained on the same dataset, with the same amount of cases or even for lesions in the same stage of development. Regrettably, no publicly available dataset exists to compare ischemic stroke lesion segmentation methods.

A similar evaluation approach can be found in Mitra et al. (2014), who likewise employed a voxel-wise classification scheme with an ensemble of decision trees and evaluated their approach on a dataset of similar size. However, where Mitra et al. (2014) used a RDF with a preceding Bayesian-Markov random field, we applied an ET forest. Also, where they aimed to segment chronic, we concentrated on sub-acute lesions. Most importantly, whereas they included old lesions and other WMLs as well, we tried to segment the most recent stroke lesion only. Furthermore, they obtained their WML ground truth by simple thresholding of the images. This last item means that they did not address the difficult problem of stroke lesion to WML similarity.

Despite facing the challenge of excluding WMLs, we obtained a higher DC

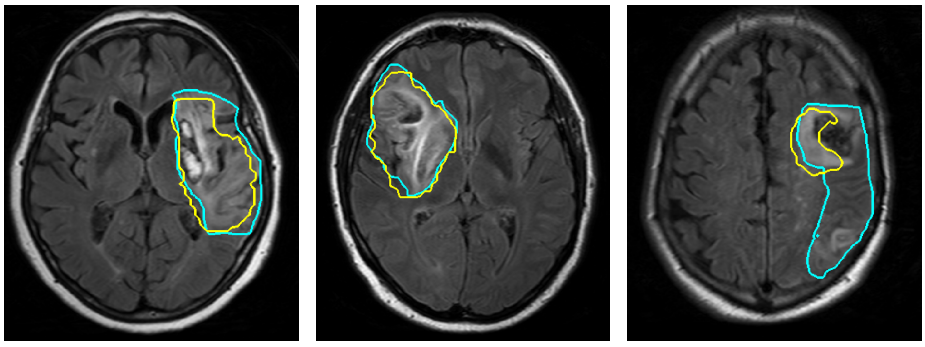
score when compared with Mitra et al. (2014) and the HD values show that outliers occurred only in three cases (02, 25 & 34, see Fig.10). Our ASSD is higher than theirs with a similar standard deviation. For some cases we reached ASSD values as low as 1.55 mm, which lies in the area of sub-pixel accuracy. The precision is mostly higher than the recall, revealing a tendency to under-segmentate the lesions, a behaviour which is not optimal for our envisioned application, although it might be desirable for other cases.

5.1. Qualitative Evaluation of segmentation performance

In this section the results of a visual examination are presented to elaborate the potential causes for the observed failures as well as successes.

In two cases, our method failed to detect the stroke lesions present. For patient 30, this can attributed to a very small lesion (1.8 ml) coupled with a heavy WML load. For patient 36 the reason is less obvious, but it seems that the intensity standardization failed for this case.

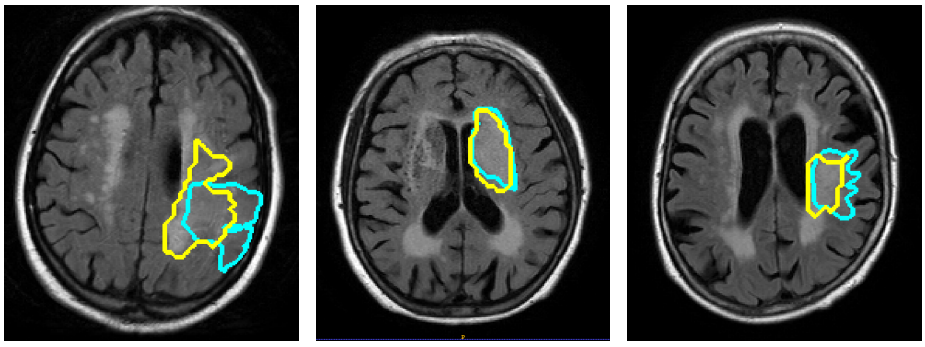
Eighteen of our cases were afflicted with hypointense haemorrhagic transformation of parts of the ischemic stroke lesion. As can be seen in Fig. 8, the proposed method was largely able to cope with these difficult cases, although Fig. 8(c) shows that large haemorrhages that are not completely surrounded by infarcted tissue can be missed.



(a) Successfull case 04 (b) Successfull case 13 (c) Failed case 08

Figure 8: Examples of the methods performance in the presence of haemorrhages. The blue/dark outline denotes the ground truth, the yellow/bright outline the segmentation result.

Although WMLs appear similar to stroke in FLAIR sequences, our method has been proven successful in excluding them as can be seen in Fig. 9. This dis-



(a) Partial leakage in case 25 (b) Successfull case 27 with many stroke-simmilar WMLs (c) Successfull case 31 with no leakage in adjacent WMLs

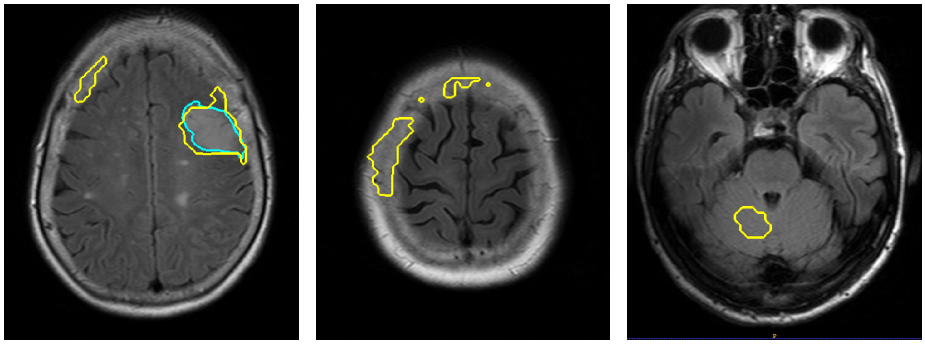
Figure 9: Examples of the methods performance in the presence of lesion-similar WMLs. The blue/dark outline denotes the ground truth, the yellow/bright outline the segmentation result.

tinguishes our method from others such as Mitra et al. (2014), who circumvented the problem by including WMLs into their ground truth.

Failed differentiation between WMLs and stroke lesions would result in outliers. These occurred only in three cases (see Fig.10) and can primarily be attributed to errors in the skull-stripping.

Our method copes implicitly with the problem of multi-focal lesions (Rekik et al., 2012), as no assumption about connectedness is made.

Some cases displayed an under-segmentation towards the skull as can be seen in Fig. 11(a). This is not a side-effect of an inaccurate skull-stripping, as the bone and background had been removed successfully at the affected location. Furthermore, the presented method shows a tendency to keep a distance from the ventricles (see Fig. 11(b)). Both of these behaviours are likely to be connected. Presumably, the strong bias towards features that average the intensity values over a larger region coupled with the downsampling of the images causes the classifier to keep such a security margin. On the downside, this



(a) Outlier in only partially removed skull of case 02

(b) Outlier in only partially removed skull of case 34

(c) Outlier in crebellum of case 18

Figure 10: The three occurences of outliers observed. The blue/dark outline denotes the ground truth, the yellow/bright outline the segmentation result.

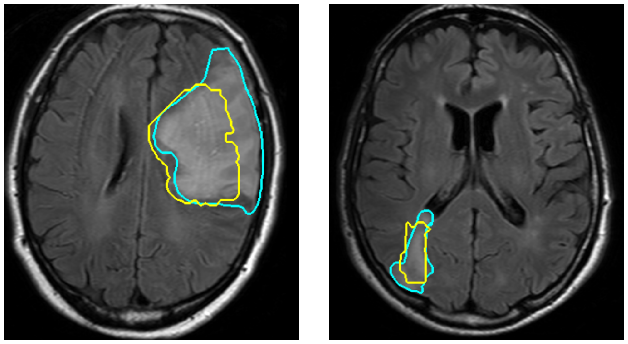
causes under-segmentation in cases where the lesion reaches up to the skull or ventricle. On the upside, it is exactly this tendency that enables our method to successfully differentiate between stroke lesions and other WMLs.

We face cases with lesions at different stages of development. The results displayed in Fig.12 show that the proposed method successfully segmented young (i.e. only slightly hyperintense) as well as old (i.e. with hypointense patches of cell decomposition) lesions.

5.2. Multi-spectral results

Different MR sequences potentially contain complementary information about the stroke lesions, e.g. hyperintensities in the DW sequence can help to differentiate between acute stroke and other WMLs. Multi-spectral approaches have been reported to be superior to mono-spectral methods (Agam et al., 2006; Forbes et al., 2010; Mitra et al., 2014). To assess if this holds true for the segmentation of sub-acute ischemic stroke lesions, we evaluated our method's performance using any possible combination of the five sequences available for our data (Fig. 5).

Of the mono-spectral approaches, the FLAIR image clearly outperformed the other sequences. This is consistent with the observation that sub-acute



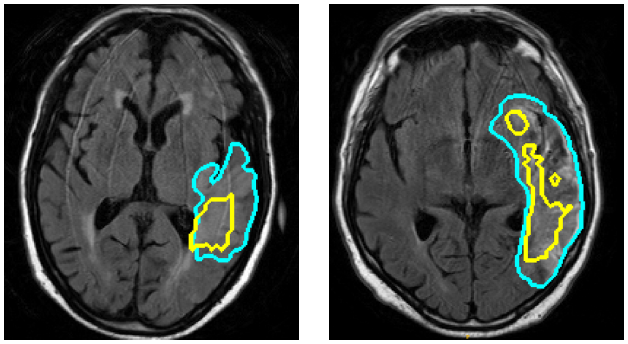
(a) Margin kept to skull in case 06, despite strong lesion evidence

(b) Margin kept to ventricle in case 18, despite strong lesion evidence

Figure 11: Examples of the method tendency to keep a margin to the skull and ventricles. The blue/dark outline denotes the ground truth, the yellow/bright outline the segmentation result.

stroke lesion areas are distinctly visible in the FLAIR sequence (Gonzalez et al., 2006). Also the DW and T2w sequence provided information suitable for the segmentation task, albeit less. When combining the FLAIR image with another sequence, the addition of T1w and DW led to the largest gain in score. Adding the T2w on the other hand, did not lead to a greater DC as could have been expected from its mono-spectral performance. This is not surprising, as the FLAIR sequence can be considered a special case of T2w. Overall, the sequence combinations including the FLAIR image markedly surpassed all others in terms of segmentation quality.

We can conclude that the FLAIR sequence is necessary and mostly sufficient for a successful segmentation of lesions in the sub-acute state. Employing a multi-spectral approach led to only small improvements. Thus, a best-effort approach can be used: Multiple classifiers are trained and always the one is selected that provides the best estimated segmentation result for the sequences available for a new case. Finally, it should be noted that the expert ground truth had been drawn in the FLAIR image, which is likely to have introduced a bias in favour of this sequence. Nevertheless, it is remarkable, that the FLAIR



(a) One day old and only slightly hyperintense lesion of case 25

(b) Old (22 days) lesion of case 23 with hypointense signs of cell decomposition

Figure 12: Segmentation results on youngest (a) and oldest (b) lesion. The blue/dark outline denotes the ground truth, the yellow/bright outline the segmentation result.

sequence alone has proven sufficient to distinguish between stroke and other WMLs, which share many similarities in intensity profile and appearance.

5.3. Feature importances

Using all possible combinations of our feature types, we obtained the results presented in Fig. 6. From there we can derive the different features' importances for the classification task and examine their interaction in detail.

With only a single feature type, the best results were obtained using the local histogram. The weighted local mean performed reasonably well, while the low DC obtained for the intensity feature shows that multi-thresholding, which it approximates, is not a suitable approach for stroke lesion segmentation. Used alone, the 2D centre distance reacheed a DC near zero, but in conjunction with the local histogram, it led to results better than the sum of their respective single performances. Adding the weighted local mean or the intensity to the local histogram did not lead to equally high results, which tells us that these three features are correlated. Combining all available feature types led to the best results.

Summarizing we can say that using more features is unlikely to worsen the

segmentation quality, as the ET forests can cope elegantly with redundant or irrelevant features.

5.4. Optimal performance of the proposed method

Using a tuned set of parameters, all employed quality measures could be improved (see Tab. 3), the DC even statistically significant (paired student's t-test, $p < 0.05$). Even more important is the reduction of the standard deviation, which signifies that especially the difficult cases were improved. The number of outliers was reduced and the distribution of results condensed to a smaller range, especially at the lower end.

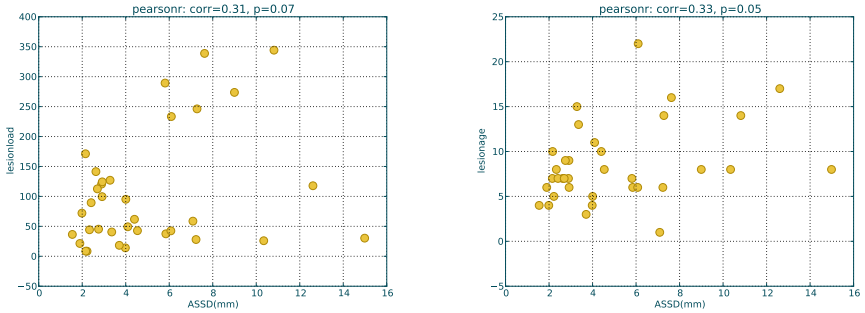
With this configuration, our method appears to outperform all previously published approaches. A head to head comparison of our method with other methods as well as a validation of this optimal configuration on a new set of cases are important tasks for the future.

5.5. Influence of clinical and imaging factors

Clinical parameters, such as the size or age of a stroke lesion, can have a severe effect on the segmentation quality. Wilke et al. (2011), for example, noted that their segmentation approach for chronic lesions performs significantly better for larger lesions according to Kendall's rank correlation. But they used the DC in the rank comparison, which is known to yield higher values for larger volumes.

We employed Pearson's r to assess the linear correlation between the lesion size and the ASSD score under the assumption of a normal distribution of the variables, leading to the results presented in Tab. 4 and the scatter plot in Fig. 13(a). Besides the correlation coefficient r, we also supply the 2-tailed p-value p as a measure of the test's significance at a 95% confidence level (i.e. $p < 0.05$). No statistically significant correlation between the lesion size and the methods performance measured by the ASSD has been found.

The current set of cases ranges from the earliest to the later sub-acute stages of stroke and hence shows a marked differences in MR-appearance. A correlation



(a) lesion size/load (ml) vs. ASSD (mm)

(b) lesion age (days) vs. ASSD (mm)

Figure 13: Scatter plots to assess correlations between clinical factors and our methods performance.

	ASSD	HD	DC
lesion size (ml)	$r = +0.31, p = 0.07$	$r = +0.01, p = 0.95$	
lesion age (days)	$r = +0.33, p = 0.05$	$r = +0.05, p = 0.79$	$\mathbf{r = -0.40, p = 0.02}$
FLAIR image quality ($\{0, 1, 2\}$)	$\rho = \mathbf{+0.55, p = 0.00}$	$\rho = \mathbf{+0.45, p = 0.01}$	$\rho = \mathbf{-0.40, p = 0.02}$
WML load ($\{0, 1, 2, 3\}$)	$\rho = -0.02, p = 0.90$	$\rho = -0.01, p = 0.97$	$\rho = -0.04, p = 0.81$
Haemorrhage ($\{no, yes\}$)	$\rho = +0.06, p = 0.73$	$\rho = +0.02, p = 0.93$	$\rho = -0.01, p = 0.97$

Table 4: Person's r and Spearman's ρ correlation coefficients as well as their respective two-tailed p-values relating a number of clinical and imaging parameters to the method's scores. Results with a confidence level higher than 95% are highlighted.

between the lesion age and the methods performance would thus imply that our approach is biased towards lesions of a specific age. Again the Pearson's r test was employed to look for linear correlations between the lesion age and ASSD as well as the HD score. The obtained results displayed in Tab. 4 show no tendency of our method to perform better (ASSD) on lesions of a specific age or to produce more outliers (HD). For the associated scatter plots, see Fig. 13(b).

Another important question is whether the image quality (rated at a scale of $1 = good$ to $3 = bad$) has an impact on the method's performance. We observed such a correlation for all quality measures (see Tab. 4) at a high confidence level of 98%, suggesting that our proposed method is slightly dependent on the

quality of the input images.

Some of our cases displayed no WMLs, whereas they were abundant in others. We rated the WML load on a scale of 0 to 3 and searched for an influence on the method's performance (ASSD) and the presence of outliers (DM). The results obtained for a Spearman's ρ test and the two-sided p-value of a hypothesis test are displayed in Tab. 4. No relation could be observed, i.e. our method seems to perform independent of the presence of WMLs. The same test has been applied to investigate how the presence of haemorrhages influence the accuracy. The values in Tab. 4 show our approach to also be robust against haemorrhages.

5.6. Other influences

A number of other factors in the pre- as well as post-processing have a potential influence on the resulting segmentation quality.

Our post-processing included the removal of small binary objects to reduce outliers. Inspecting the results presented in Tab. 2, it becomes clear that the only observable influence of this measure is a lowering of the HD values i.e. it fulfils it designated purpose of eliminating small outliers. Visual examination revealed that in no case a real ischemic stroke lesion has been falsely removed.

Only a subset of all available samples was used for training, which might have introduced variance. The results obtained by repeated sampling presented in Sec. 4.7 show that our sampling method is very robust: The standard deviations for all evaluation methods ranged in the third decimal.

Increasing the amount of training samples would lead to only a marginal gain in segmentation quality at a high cost of training time as Fig. 7 shows.

6. Conclusion

In this work, we presented a new segmentation method for sub-acute ischemic stroke lesion segmentation based on ET forests trained with simple image features. Evaluation on a large and well-described dataset has shown the proposed approach to perform well even on challenging cases and to compare favourably with other propositions from literature.

The ET forests ability to generalize well from noisy training data and their robustness against overfitting make them an adequate choice for the challenging classification problem at hand. Very good performance could be reached with only a few simple feature.

Of these, the local histogram has been established as highly discriminative and an investigation of the features' importances illustrated the methods's robustness against redundant information.

Among the MR sequences, FLAIR has been found to be required and largely sufficient for a successful segmentation of sub-acute stroke lesions, although the method benefits from an addition of the T1 and DW sequences.

Furthermore, a simple sampling strategy was proposed, allowing for a considerable reduction of the training set size without any loss in segmentation accuracy.

Statistical analyses indicate that our method is largely independent of clinical and anatomical stroke parameters. Only the image quality appears to exert some influence on the outcome.

The fully automatic nature of the method, its proficiency on routine images of varying stroke appearance and its customisability to different MR sequence input combinations renders it an ideal candidate for applications in neuroscientific context as, for example, the screening of large databases.

In two cases, our approach failed to detect the stroke lesion and in some others the segmentation quality was poor. This can largely be attributed to a low FLAIR image quality. Further improvements are necessary to reach inter-observer quality.

In future, we would like to test more elaborated features to support the discrimination process and to employ a subsequent narrow-band segmentation to counter the under-segmentation. Furthermore, an multicenter evaluation with ground truth data prepared by different experts is planned.

References

Agam G, Weiss D, Soman M, Arfanakis K. Probabilistic brain lesion segmentation in DT-MRI. In: IEEE ICIP 2006. 2006. p. 89–92. doi:10.1109/ICIP.2006.312369.

Breiman L. Bagging predictors. Machine Learning 1996;24(2):123–40. doi:10.1007/BF00058655.

Breiman L. Random forests. Machine Learning 2001;45(1):5–32. doi:10.1023/A:1010933404324.

Breiman L, Friedman J, Stone CJ, Olshen R. Classification and Regression Trees (Wadsworth Statistics/Probability). Chapman and Hall/CRC, 1984.

Criminisi A, Shotton J, editors. Decision Forests for Computer Vision and Medical Image Analysis. Springer, 2013.

Fiez JA, Damasio H, Grabowski TJ. Lesion segmentation and manual warping to a reference brain: Intra-and interobserver reliability. Human Brain Mapping 2000;9(4):192–211. doi:10.1002/(SICI)1097-0193(200004)9:4<192::AID-HBM2>3.0.CO;2-Y.

Forbes F, Doyle S, Garcia-Lorenzo D, Barillot C, Dojat M. Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation. In: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on. 2010. p. 69–72. doi:10.1109/ISBI.2010.5490413.

von der Gablentz J. Visuelle Exploration von dynamischen Szenen bei Schlaganfallpatienten mit Neglect-Syndrom. Ph.D. thesis; Klink für Neurologie, Universität zu Lübeck, Lübeck; 2012.

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning 2006;63(1):3–42. doi:10.1007/s10994-006-6226-1.

Gonzalez RG, Hirsch JA, Koroshetz W, Lev MH, Schaefer PW, editors. Acute Ischemic Stroke - Imaging and Intervention. Springer, 2006.

Hevia-Montiel N, Jimenez-Alaniz J, Medina-Banuelos V, Yanez-Suarez O, Rosso C, Samson Y, Baillet S. Robust nonparametric segmentation of infarct lesion from diffusion-weighted MR images. In: IEEE EMBS 2007. 2007. p. 2102–5. doi:10.1109/IEMBS.2007.4352736.

Ho TK. The random subspace method for constructing decision forests. Pattern Anal Mach Intell, IEEE Trans on 1998;20(8):832–44. doi:10.1109/34.709601.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. NeuroImage 2012;62(2):782–90. doi:10.1016/j.neuroimage.2011.09.015.

Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. elastix: A toolbox for intensity-based medical image registration. Med Imag, IEEE Trans 2010;29(1):196–205. doi:10.1109/TMI.2009.2035616.

Krämer UM, Solbakk AK, Funderud I, Løvstad M, Endestad T, Knight RT. The role of the lateral prefrontal cortex in inhibitory motor control. Cortex 2013;49(3):837–49.

Likar B, Viergever MA, Pernus F. Retrospective correction of MR intensity inhomogeneity by information minimization. Med Imag, IEEE Trans 2001;20(12):1398–410. doi:10.1109/42.974934.

Machner B, Dorr M, Sprenger A, von der Gablentz J, Heide W, Barth E, Helmchen C. Impact of dynamic bottom-up features and top-down control on the visual exploration of moving real-world scenes in hemispatial neglect. Neuropsychologia 2012;50(10):2415–25. doi:10.1016/j.neuropsychologia.2012.06.012.

Machner B, Könemund I, Sprenger A, von der Gablentz J, Helmchen C. Randomized controlled trial on hemifield eye patching and optokinetic stimulation in acute spatial neglect. Stroke 2014;doi:10.1161/STROKEAHA.114.006059; preprint.

Maier O, Wilms M, Gablentz J, Krämer U, Handels H. Ischemic stroke lesion segmentation in Multi-Spectral MR images with support vector machine classifiers. In: SPIE Medical Imaging 2014, Computer-Aided Diagnosis. San Diego; volume 9035; 2014. doi:10.1117/12.2043494.

Mathers CD, Boerma T, Ma Fat D. Global and regional causes of death. British Medical Bulletin 2009;92(1):7–32. doi:10.1093/bmb/ldp028.

Mitra J, Bourgeat P, Fripp J, Ghose S, Rose S, Salvado O, Connelly A, Campbell B, Palmer S, Sharma G, et al. Lesion segmentation from multimodal MRI using random forest following ischemic stroke. NeuroImage 2014;doi:10.1016/j.neuroimage.2014.04.056.

Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. Med Imag, IEEE Trans 2000;19(2):143–50. doi:10.1109/42.836373.

Rekik I, Allassonniere S, Carpenter TK, Wardlaw JM. 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 2012;1(1):164–78. doi:10.1016/j.nicl.2012.10.003.

Seghier ML, Ramlackhansingh A, Crinion J, Leff AP, Price CJ. Lesion identification using unified segmentation-normalisation models and fuzzy clustering. NeuroImage 2008;41(4–3):1253–66. doi:10.1016/j.neuroimage.2008.03.028.

Shamonin DP, Bron EE, Lelieveldt BP, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer's disease. Frontiers in Neuroinformatics 2014;7(50). doi:10.3389/fninf.2013.00050; preprint.

Smith SM. Fast robust automated brain extraction. Human Brain Mapping 2002;17(3):143–55. doi:10.1002/hbm.10062.

Wilke M, de Haan B, Juenger H, Karnath HO. Manual, semi-automated, and automated delineation of chronic brain lesions: a comparison of methods. NeuroImage 2011;56(4):2038–46. doi:10.1016/j.neuroimage.2011.04.014.