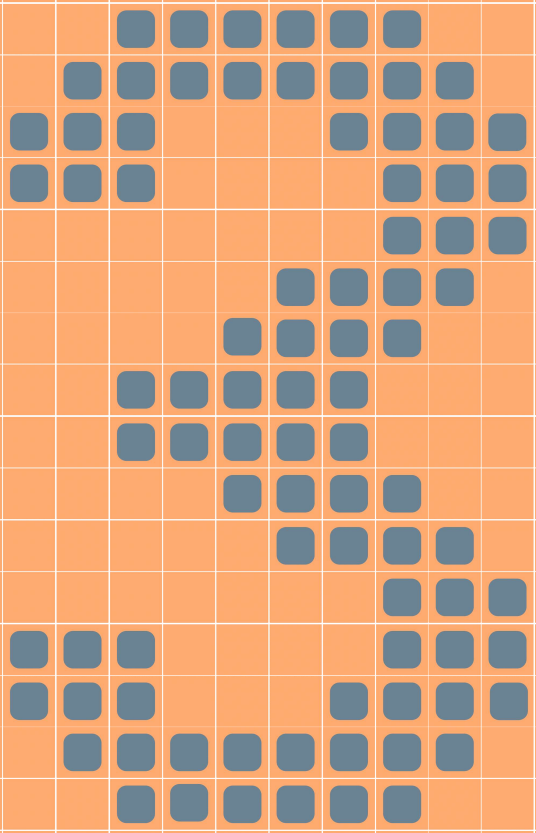
Deep-learning-based image segmentation for uncommon ischemic stroke

From infants to adults

Zoetmulder, R.

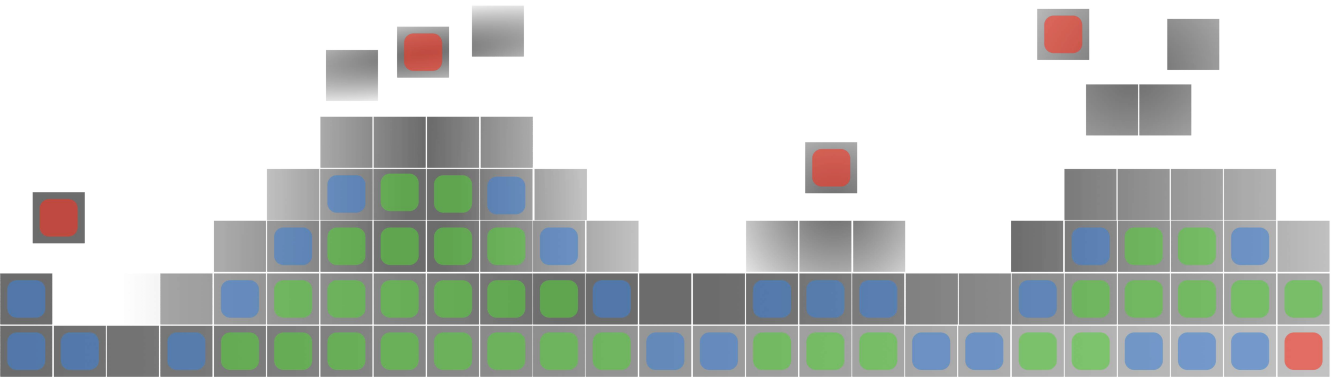
Publication date
2023



Chapter 3

Automated Final Lesion Segmentation in Posterior Circulation Acute Ischemic Stroke using Deep Learning

Based on: R. Zoetmulder, P.R. Konduri, I.V. Obdeijn, E. Gavves, I. Išgum, C.B.L.M. Majoie, D.W.J Dippel, Y.B.W.E.M. Roos, M. Goyal, P.J. Mitchell, B.C.V. Campbell, D.K. Lopes, G. Reimann, T.G. Jovin, J.L. Saver, K.W. Muir, P. White, S. Bracard, B. Chen, S. Brown, W.J. Schonewille, E. van der Hoeven, V.Puetz and H.A. Marquering (2022). "Automated Final Lesion Segmentation in Posterior Circulation Acute Ischemic Stroke using Deep Learning". Diagnostics, 11(9), 1621.



Abstract

Final lesion volume (FLV) is a surrogate outcome measure in anterior circulation stroke (ACS). In posterior circulation stroke (PCS) this relation is plausibly understudied due to a lack of methods that automatically quantify FLV. Applicability of deep learning approaches to PCS is limited due to its lower incidence compared to ACS. We evaluated strategies to develop a CNN for PCS lesion segmentation by using image data from both ACS and PCS patients. We included follow-up NCCT scans of 1018 patients with ACS and 107 patients with PCS. To assess whether an ACS lesion segmentation generalizes to PCS, a CNN was trained on ACS data (ACS-CNN). Second, to evaluate performance of only including PCS patients, a CNN was trained on PCS data. Third, to evaluate performance when combining the datasets, a CNN was trained on both datasets. Finally, to evaluate performance of transfer learning, the ACS-CNN was fine-tuned using PCS patients. The transfer learning strategy outperformed the other strategies in volume agreement with an ICC of 0.88 (95%CI: 0.83-0.92) versus 0.55 to 0.83 and a lesion detection rate of 87% versus 41-77 for the other strategies. Hence, transfer learning improved the FLV quantification and detection rate of PCS lesions compared to the other strategies.

3.1 Introduction

Infarct volume, representing the tissue damage after an Acute Ischemic Stroke (AIS), is commonly considered as a surrogate endpoint for the primary functional outcome (modified Ranking Scale (mRS) after 90 days) [144]. Various trials have shown a strong association of infarct volume with functional outcome in patients suffering from a stroke due to a large vessel occlusion in the anterior circulation [144; 172].

However, in patients with a stroke due to a Posterior Circulation Stroke (PCS), the relation between infarct volume and outcome is understudied [167]. The low number of studies addressing this relation may be due to the combination of two reasons; the relatively low prevalence of PCS compared to Anterior Circulation Stroke (ACS) and the lack of automated analysis of PCS lesion volume assessment.

With the huge effectiveness of endovascular treatment of anterior circulation stroke patients, treatment of posterior stroke has attained renewed interest in various studies and trials. For example, the recently completed BASICS trial [163] could not show a beneficial effect of endovascular treatment with functional outcome used as outcome measure. Alternatively, secondary outcome measures such as infarct volume might show a beneficial effect of certain treatments since functional outcome, as addressed by the mRS, is a rather coarse outcome measure, which is also affected by many other confounders [141]. Developing methods that automatically segment lesions due to a PCS, would help investigate Final Lesion Volume (FLV) as a surrogate outcome for this type of stroke. Solutions for the automatic segmentation of FLV based on convolutional neural networks (CNNs) have been presented in the literature for CT and MR imaging [143; 151]. However, these studies have only considered FLV of patients with an AIS due to an occlusion of the anterior circulation [143; 144]. To achieve good performance, CNNs typically require large amounts of labeled training data. However, PCS constitutes only 26 % of AIS cases [145; 156] and training of CNNs for automatic PCS lesion segmentation is thus hindered by limited availability of data. Furthermore, the applicability of methods developed for ACS FLV segmentation on posterior stroke lesion segmentation is unknown.

Several methods exist for dealing with a lack of data to train a CNN. One method that reduces the data needed to train CNNs by reusing knowledge is transfer learning [158]. To perform transfer learning, a CNN is pre-trained on a task for which large amounts of image data are available and fine-tuned on a different task for which little image data is available. Transfer learning has been successfully applied to solve various medical image analysis problems [153].

We evaluate strategies to create automated PCS lesion segmentation by using image data from patients with ACS and patients with PCS. We hypothesize that transfer learning utilizing data of ACS lesions improves automatic PCS lesion segmentation performance compared to alternative strategies: training a CNN on only ACS lesions, only on PCS lesions, or on the combination of ACS and PCS lesions.

3.2 Materials & Methods

3.2.1 Patient and Image Data

All involved patients in this retrospective study or their legal representatives provided written informed consent. The medical ethics committee of each participating hospital approved the use of the data after anonymization. The Hermes dataset consists of 1665 patients who suffered from an ACS and was obtained from the HERMES collaboration [149], which investigated the effectiveness of endovascular therapy for treating ACS. This collaboration combined data from seven clinical randomized trials and collected data between December 2010 and December 2014. The inclusion criteria are shown in Figure 3.4a, in the appendix. Patients were excluded if no Follow-Up Non-Contrast Computed Tomography (FU-NCCT) was made in the time window of 12 hours and 2 weeks after stroke onset or if the preprocessing steps were unsuccessful. In total 1018 patients out of the 1665 patients were included. Baseline characteristics of the included patients are shown in Table 3.1.

The BASICS dataset consists of 168 patients who suffered from a PCS and was obtained from the BASICS trial [152; 166], which investigated the effectiveness of endovascular therapy for treating patients with a PCS. This trial

Parameter	Posterior Stroke	Anterior Stroke
Clinical		
Age, years, Mean (Standard Deviation)	65.65 (12.2)	66.1 (13.3)
Sex, F, No. [%]	34/107 [31.8]	458/1018 [45]
NIHSS at baseline, mean[median](N)	21.4 [19] (107)	17 [17] (1015)
Prior Conditions		
Diabetes mellitus, No. [%]	28/107 [26.2]	169/1018 [16.6]
Hypertension, No. [%]	64 /107 [59.8]	564/1018 [55.4]
Stroke, No. [%]	21/107 [19.6]	121/1018 [11.9]
Posterior Circulation Stroke, No. [%]	7/107 [6.5]	NAV
TIA, No. [%]	10/106 [9.4]	NAV
Posterior Circulation TIA, No. [%]	2/106 [1.9]	NAV
Atrial Fibrillation, No. [%]	13/107 [12.1]	314/1018 [30.8]
Atrial Fibrillation (history or 12 lead ECG), No. [%]	23/107 [21.5]	NAV
Pre-Stroke mRS		
0, No. [%]	80/107 [74.8]	836/1017 [82.1]
1, No. [%]	11/107[10.3]	129/1017 [12.7]
2, No. [%]	13/107 [12.1]	29/1017 [2.9]
3, No. [%]	3/107[2.8]	23/1017 [2.3]
Treatment		
IV Thrombolysis, No. [%]	92/107 [86]	872/1018 [85.7]
Time		
Stroke onset to IVT, minutes, Mean [Standard Deviation](N)	176.9 [176.102] (90)	112.2 [57.2] (871)

Table 3.1: Baseline characteristics, treatment and time data for patients with posterior circulation stroke and anterior circulation stroke. Prior Posterior stroke, Transient Ischemic Attack, posterior circulation TIA and Atrial Fibrillation (history or 12 lead ECG) were not available (NAV) for the HERMES dataset.

included patients from 23 centers, collected between 2011 and 2019. Inclusion criteria for our study are shown in Figure 3.4b, in the appendix. The latest FU-NCCT scan was used if multiple scans were available for the same patient. Patients were excluded if no FU-NCCT was made or if the follow up image was of insufficient quality. In total, 107 patients out of the 168 available patients were included. Baseline characteristics of the included BASICS patients are shown in Table 3.1. The infarcted regions are shown in Table 3.2, in the appendix.

3.2.2 Reference Segmentations

For patients with an ACS, reference segmentations were obtained by manual annotation by one of two experienced observers on the most recent FU-NCCT. The annotation procedure is outlined in [144]. In summary, a window width of 30 Hounsfield Units (HU) and a center level of 35 HU was set in ITK-Snap [171]. All hypodense regions on the ipsilateral hemisphere including

edema were included in the segmentations. Infarcted tissue in the ipsilateral hemisphere with signs of an old infarct were excluded from the reference segmentations. Parenchymal hemorrhages adjacent to or within the affected area were included in the reference segmentations. Finally, reference segmentations were checked and, if necessary, corrected by one of three radiologists, each of whom had more than 5 years of experience.

Reference segmentations of lesions caused by a PCS on FU-NCCT scans were manually segmented by a single trained observer (IVO) and were checked by an experienced radiologist (CBLMM) who had more than 5 years of experience. Lesions were segmented by using the aforementioned window width and center level on ITK-Snap software [171]. Posterior Circulation Alberta Stroke Program Early Computed Tomography Score (PC-ASPECTS) [159] were used when available to identify the infarcted territory.

3.2.3 Preprocessing

The intracranial region as a volume of interest was obtained automatically using a combination of preprocessing steps [143]. The bone was segmented using a threshold-based segmentation by selecting all voxels with an intensity of 170 HU or higher. Subsequently, the foramina, except the foramen magnum, were closed using morphological filters, and a region growing algorithm was applied to select the intracranial volume. To obtain the final volume of interest, the region caudal to the foramen magnum was excluded. To ensure the same size, orientation and voxel sizes, all scans were aligned by automatically registering the images to a common space using rigid and affine transformations. Images were registered using the Mattes Mutual Information [155] with a gradient descent optimizer. In addition to registration, the scans were downsampled to allow the entire scan to be passed into the CNNs. After the preprocessing each scan had a size of 256 x 256 x 32, with a slice thickness of 5 mm. The voxel intensities were clipped between -20 and 120 HU, and subsequently normalized between minus one and one. The preprocessing was done using SimpleITK [154; 168] and Python 2.7.

3.2.4 CNN for Automatic Posterior Circulation Lesion Segmentation

The preprocessed images were input to a CNN which consisted of three-dimensional convolutional kernels. The architecture of the CNN (shown in Figure 3.5a in the appendix), was inspired by U-Net [161] and ResNet [150]. The CNN consisted of a downsampling path and an upsampling path. The downsampling path started with an input block, consisting of a convolution with a kernel size of three and a stride of one, followed by a max pooling layer with a pooling size of two and a stride of two. The features generated by the input block consisted of eight channels.

Subsequent to the input block, three downsampling blocks were added, consisting of three 3D ResNet layers, shown in Figure 3.5b, in the appendix. The first two blocks were followed by average pooling with a stride and pooling size of two.

The upsampling path started with a transposed convolution using a stride of two. Next, two upsampling blocks followed by an output layer were added. Each upsampling block consisted of two ResNet layers, followed by a transposed convolution with a stride of two and a kernel size of three. Each upsampling block took the features from the previous block and the corresponding downsampling block and concatenated them. The output block consisted of two ResNet blocks followed by a convolutional layer. The CNNs were implemented using Tensorflow 1.5.

3.2.5 Experimental Setup

Four different training strategies for CNNs were evaluated: A CNN was randomly initialized and trained on images of patients in the HERMES dataset (ACS-CNN), BASICS dataset (PCS-CNN) and the HERMES and BASICS datasets combined (CD-CNN). The ACS-CNN was used to establish the generalization ability of a CNN trained on ACS to PCS lesion segmentation. The PCS-CNN served as a baseline for training with limited but representative data. The CD-CNN was used as a benchmark if both ACS and PCS data were available, but no transfer learning was used. Transfer learning reused the weights from the trained ACS-CNN to initialize all but the last block of

the CNN, and fine-tuned by updating all the weights, using images from the BASICS dataset (TL-CNN).

The CNNs used group normalization with four groups, the Leaky Relu activation a batch size of two and the Adam optimizer. The loss function used was the weighted binary cross entropy. The initial learning rate for the ACS-CNN and CD-CNN was 10^{-3} and was decayed stepwise after 5, 10, 15 and 20 epochs to respectively, $5 \cdot 10^{-4}$, $2 \cdot 10^{-4}$, 10^{-4} , and 10^{-5}. These networks were trained for 25 epochs. The initial learning for the PCS-CNN and TL-CNN was 10^{-5} and was decayed after 25, 50 and 75 epochs to respectively, $5 \cdot 10^{-6}$, $2 \cdot 10^{-6}$, and 10^{-6}. These networks were trained and fine-tuned for 100 epochs. The weight decay was set to 10^{-5}.

Data augmentation was applied at training time. The images were rotated at a randomly chosen angle between zero and ten degrees along the axial plane in either direction or were randomly flipped along the sagittal plane.

We evaluated the performance of the ACS-CNN to check whether the model converges during pre-training. The ACS dataset was split randomly into scans for training (85 %), validation (5 %) and testing (10 %). For this approach, we used the entire PCS dataset for evaluation. Thus, we used stratified five-fold cross-validation. Given the number of available PCS patients, the first four testing splits consisted of 20 % and the fifth of 22 % of the data. The training splits were of equal size and consisted of 78 % of the PCS data.

3.2.6 Evaluation

The reliability between the automatically and manually segmented volumes was evaluated with the Intraclass Correlation Coefficient (ICC) including the 95% Confidence Interval (95% CI). The ICC was interpreted in accordance to the American Psychological Association [147]. Following their guidelines, an ICC < 0.4 is defined as poor, an ICC between 0.4 and 0.6 as fair, an ICC between 0.6 and 0.75 as good and an ICC greater than 0.75 as excellent. In addition, a Bland-Altman analysis was performed to assess the bias and spread in volume measurements. Statistical significance between ICCs was evaluated by using Fisher's r-to-z transformation.

To determine whether our model accurately detected lesions independent of size, the number of detected lesions was determined. For this measure,

we calculated the ratio of the total number of correctly detected lesions and the total number of lesions as determined by the ground truth in the dataset. A lesion was defined as detected if the percentage of overlapping voxels between the automatic and reference segmentations was larger than a predefined threshold. In case of small thresholds, a non-zero overlap of automated and reference lesion segmentation could be caused by chance. To account for this issue the required percentage of overlapping voxels was set to either greater than zero percent or a more conservative 20%. Next, the effect of lesion volume on the lesion detection rate was studied by excluding lesions of a progressively larger volume. This latter cutoff was set between zero and 4 mL, with increments of 0.5 mL.

The segmentation performance of the automatic methods was evaluated by calculating the Dice coefficient as an overlap measure between the reference and the automatic segmentation. Normality of the distribution of the Dice coefficients was assessed before pairwise statistical testing by using Shapiro-Wilk test. If the Dice coefficients were normally distributed, a paired t-test was used, otherwise a Wilcoxon rank sum test was used. P-values were corrected for the family wise error rate using the Bonferroni correction. All statistical testing was done using the python library Pingouin, version 0.3.1 [165].

3.3 Results

Baseline characteristics of the patients in BASICS and HERMES datasets were compared. Patients in the HERMES dataset had a similar age to patients in the BASICS dataset. Diabetes (26.2% vs 16.6%, $p<0.05$) and prior stroke (19.6% vs 11.9%, $p<0.05$) occurred more frequently in patients in the BASICS dataset. However, atrial fibrillation (12.1% vs 30.8%, $p<0.01$) occurred more frequently in patients in the HERMES dataset.

The median FLV in patients with PCS was 11 (IQR: 3.4 - 36) mL. The ICCs for volume assessments for the TL-CNN, PCS-CNN, CD-CNN, and ACS-CNN were 0.88 (95% CI: 0.83-0.92), 0.80 (95% CI: 0.72-0.86), 0.83 (95% CI: 0.76-0.88) and 0.55 (95% CI: 0.4-0.67), respectively. The ICC of the TL-CNN, was significantly larger than the ICCs of the ACS-CNN ($p<0.01$) and PCS-CNN ($p = 0.02$). The ICC of the ACS-CNN was significantly smaller than the ICCs

of the PCS-CNN ($p<0.01$) and CD-CNN ($p<0.01$). The bias and spread of the volume measurements are shown in Figure 3.1 and Table 3.3.

The lesion detection rate of the TL-CNN was higher than for the other learning strategies (Figure 3.2). Figure 3.2 shows that the lesion detection rate increases with increasing lesion volume and that the lesion detection rate decreases with increasing thresholds of overlapping voxels.

The TL-CNN, PCS-CNN, CD-CNN and ACS-CNN achieved a Dice coefficient of 0.25±0.08, 0.21±0.06, 0.16±0.06 and 0.07±0.03, on the overall PCS test set, respectively (Figure 3.3). The Dice coefficients were not normally distributed. Hence, a Wilcoxon rank sum test was used. Results of the Wilcoxon rank sum test are shown in Table 3.4. For the anterior circulation stroke lesions, the ACS-CNN achieved an average Dice coefficient of 0.60±0.07.

3.4 Discussion

Our study found that transfer learning results in a high level of agreement between manually delineated and automatically quantified lesion volumes on follow-up NCCTs of patients with a PCS. Furthermore, we found that transfer learning resulted in higher spatial accuracy and larger volume agreement of automatic PCS lesion segmentation compared to the other strategies. In addition, the TL-CNN models also detected a larger number of PCS lesions in comparison to the other strategies. Moreover, our results indicate that the ACS-CNN models, which were trained on only patients who suffered from an ACS, do not generalize to PCS lesion segmentation.

Our study is one of the few that addressed automated segmentation of posterior circulation stroke lesions on follow-up NCCT. Previous work include anterior stroke lesion segmentation in a variety of imaging modalities, such as baseline CTP [148] baseline CTA [157], follow-up NCCT [143] and baseline and follow-up DWI [170] using multiple approaches and addressing various types of stroke.

A previous study focused on developing a CNN-based method for automatic ACS lesion segmentation on FU-NCCT [143] with a higher spatial overlap accuracy than found in our study. This could be explained by the larger lesion volumes in their population (median FLV of 48 vs 11 mL) and

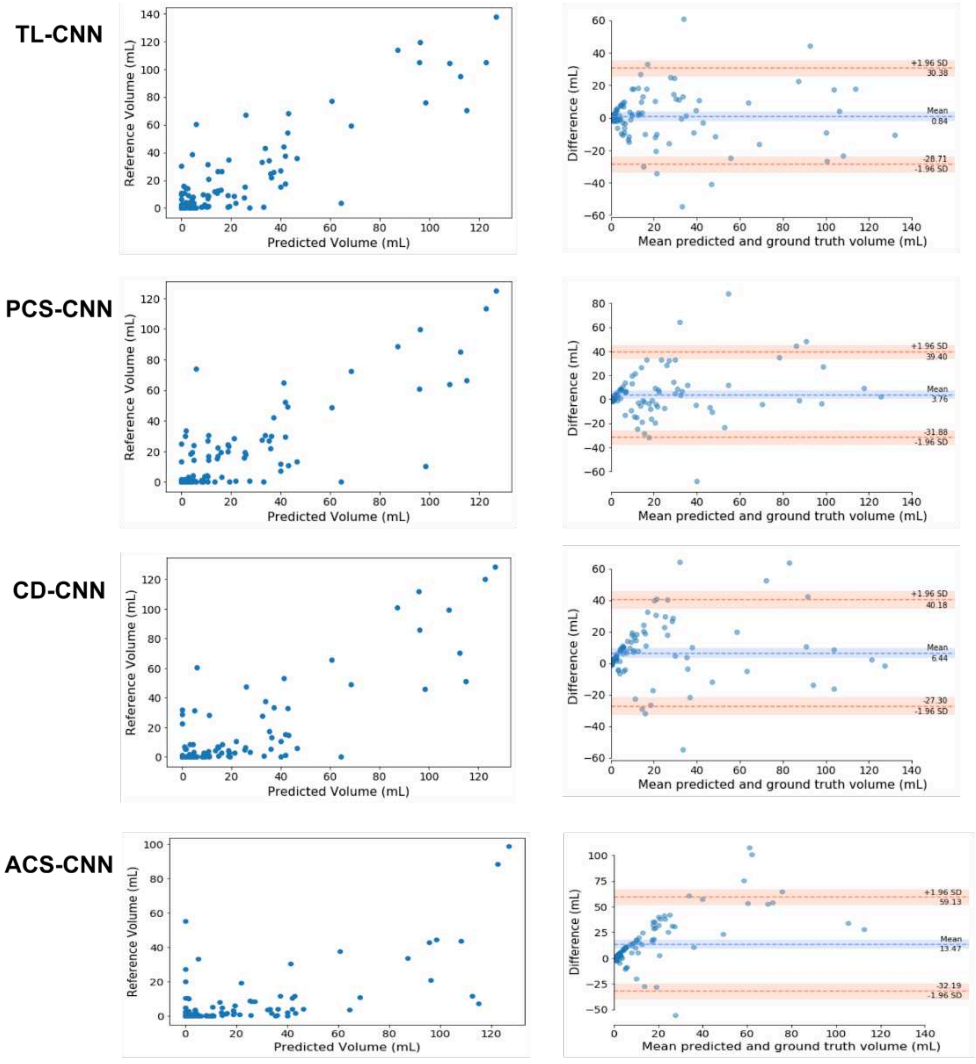


Figure 3.1: Comparison of the automated and reference segmentation volume for the Transfer Learned CNN (TL-CNN), Posterior Circulation Stroke CNN (PCS-CNN), Combined Datasets CNN (CD-CNN) and Anterior Circulation Stroke CNN (ACS-CNN). Left column: Scatter plots comparing lesion volumes derived from the reference segmentations (y-axis) and from the automatic segmentations determined by the CNN (x-axis). Right column: Bland Altman plots of the lesion volumes. The volumes corresponding to the reference and automatic segmentations are shown on the x-axis and the volume difference y-axis.

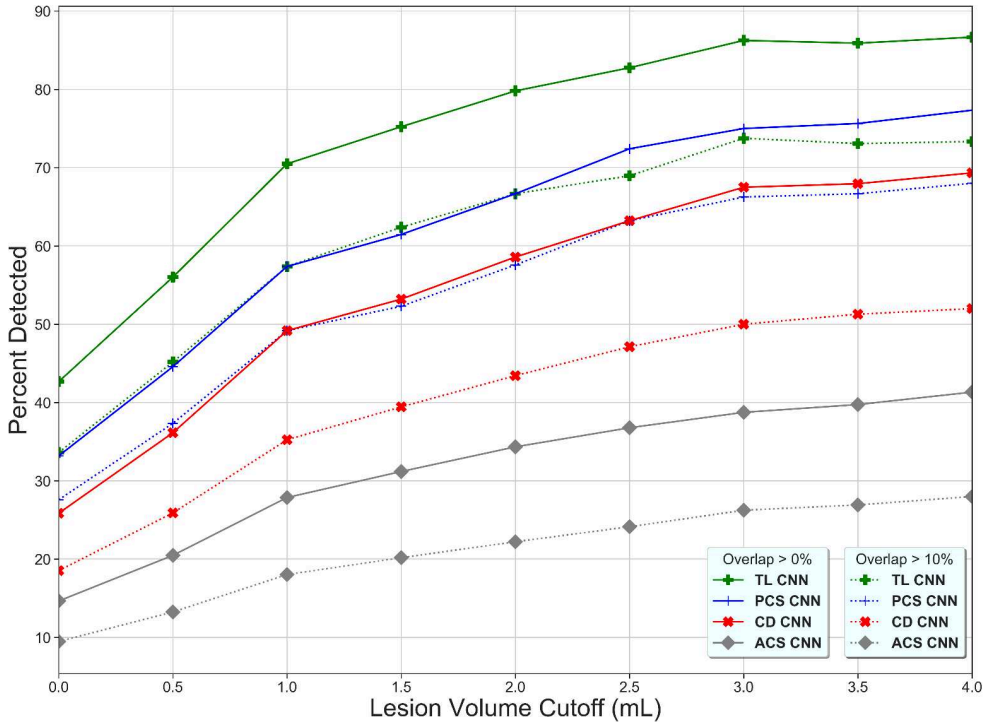


Figure 3.2: Percentage of detected lesions (y-axis) as a function of the minimum volume requirement (x-axis) and the minimum percentage of overlapping voxels for the Transfer Learned CNN (TL-CNN), Posterior Circulation Stroke CNN (PCS-CNN), Combined Datasets CNN (CD-CNN) and Anterior Circulation Stroke CNN (ACS-CNN) (green, blue, red and gray lines). For all methods a higher lesion volume cutoff results in a higher percentage of detected lesions. The lower overlapping voxel requirement, the higher the percentage of detected lesions (dotted versus solid lines).

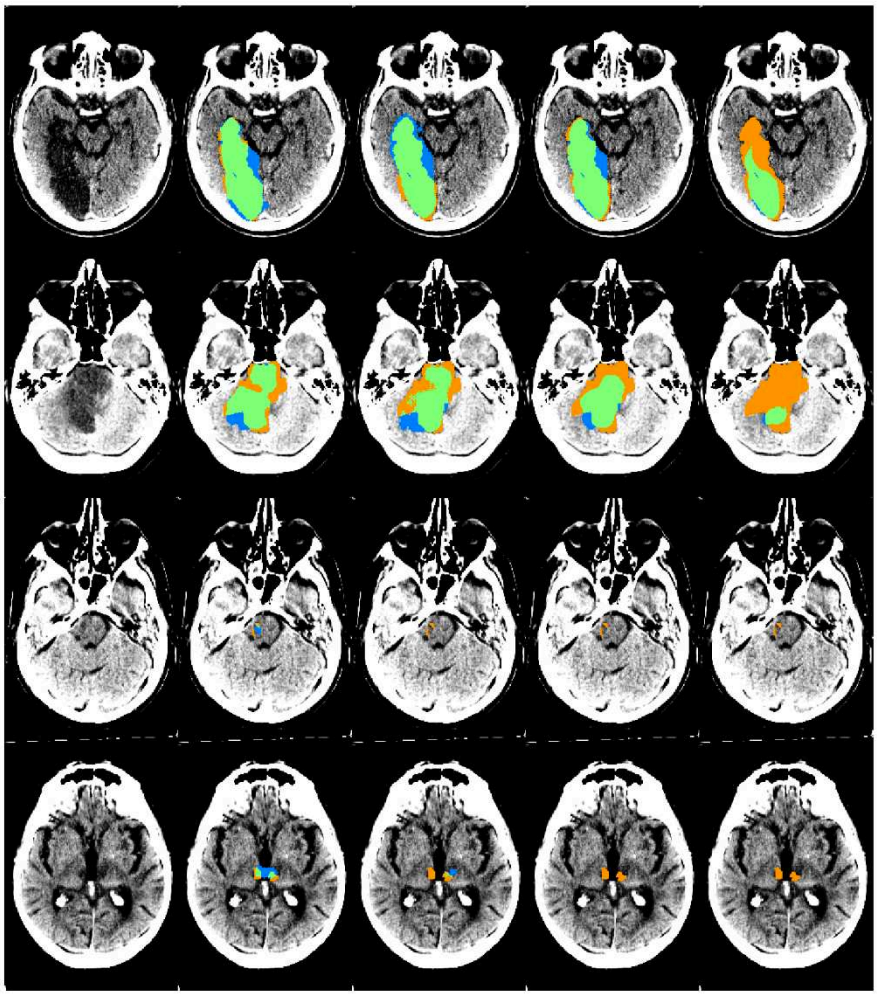


Figure 3.3: An example of automatic segmentation results obtained by the 4 CNNs on the PCS test set. From the left to the right column: the original scan, the automatic segmentation results from the TL-CNN, PCS-CNN, CD-CNN and ACS-CNN are shown. The segmentation maps show true positives (green), false positives (blue) and false negatives (orange). The scans were plotted using a window center around 35, with a window width of 30.

the larger dataset available for training. In addition, the Dice coefficient is a global overlap metric which works well for large lesions. If lesions are smaller the Dice coefficient may be too sensitive to small errors and detection metrics are more important.

Other methods for automatic ACS lesion segmentation used information from the contralateral hemisphere to improve segmentation accuracy [142; 157]. PCS lesions can affect both hemispheres, hence comparing information between the ipsilateral and its contralateral hemisphere is unlikely to improve the accuracy.

Automated stroke lesion segmentation has also been developed for chronic stroke lesions on T1 MRI. Chronic stroke lesions have been segmented by using a random forest classifier to segment lesions in the left hemisphere by using hand-crafted image features [160]. Another study used a deep residual network to segment lesions on images of the ATLAS dataset, which contains manually traced lesions on 304 T1-weighted MRI images [164]. Both studies achieved higher similarity scores than our method, which could be explained by the larger FLVs and higher sensitivity provided by T1-weighted MRI images.

In previous research, transfer learning has also been successfully applied to improve accuracy of various other medical image segmentation tasks. One study used CNNs pre-trained on eight different medical image segmentation tasks on various imaging modalities to improve automatic lung, liver and liver tumor segmentation [146]. Unlike the aforementioned study, our study pre-trained on a single imaging modality and task. Another study, pre-trained CNNs using self-supervised tasks to improve lung nodule, liver and brain tumor segmentation [173]. In agreement with our study, the results of prior work indicate that for medical image segmentation transfer learning can be beneficial.

Other work using transfer learning for medical image tasks included CNNs pre-trained on ImageNet [162] as a benchmark. These studies used transfer learning to improve performance on medical image analysis tasks on 2D images. However, using ImageNet for transfer learning was less likely to be suitable for our study because prior work has shown that ImageNet pre-training improves performance on medical image analysis tasks less than

using a pre-trained 3D model [173].

This study has several limitations. First, most of the FU-NCCT scans of PCS patients included in this study were mostly obtained 24 hours after onset of AIS. Creating manual reference segmentations of the lesions on these early FU-NCCT scans is more challenging owing to the subtle differences in HU values after 24 hours. In addition, FLV segmentations for patients with PCS were performed by only one trained observer, hence inter-observer agreement could not be assessed. However, the reference segmentations were verified by an experienced radiologist. Second, this study suffered from a low number of available PCS patients. Therefore, an even lower number of patients would be included if the data was divided into training, validation and test sets. This would have lowered the generalizability of the results to the PCS patient population. To overcome this, five-fold cross validation was used to allow all the data to be used as testing data in the analysis and to assess the stability of the presented results. Third, the CNNs were not accurate at detecting lesions with a volume smaller than 2mL. If a patient is suspected of having lesions with a small volume, results from the presented algorithm should be verified by an expert evaluation. Transfer learning allowed the CNNs to reuse information learned from ACS lesion segmentation to segment lesions caused by PCS. The resulting improvement in PCS lesion segmentation is likely due to the similarity between the pre-training and the fine-tuning tasks [169]. However, in our approach the detection and segmentation of small lesions and the segmentation of lesions that are connected to cerebrospinal fluid filled areas is still suboptimal.

Deep learning is potentially valuable for automating demanding tasks in the quantification of radiological imaging. It is well-known that deep learning requires large amounts of data to train algorithms, which may suggest a limited applicability of deep learning in less common diseases. This study also shows that deep learning models that are trained on a more general, less specific disease may not be sufficient. Here we presented an alternative approach based on transfer learning and showed that deep learning models can be pre-trained on similar diseases and fine-tuned on the specific rarer disease.

To conclude, the presented transfer learning approach improves automatic

detection and segmentation of posterior circulation stroke lesions compared to the evaluated commonly used training strategies. The presented automated posterior stroke lesion segmentation method allows inclusion of lesion volume as an image outcome measure and as a metric to predict outcome in large-scale clinical trials and potentially as a first step towards clinical application.

3.5 Supplementary materials

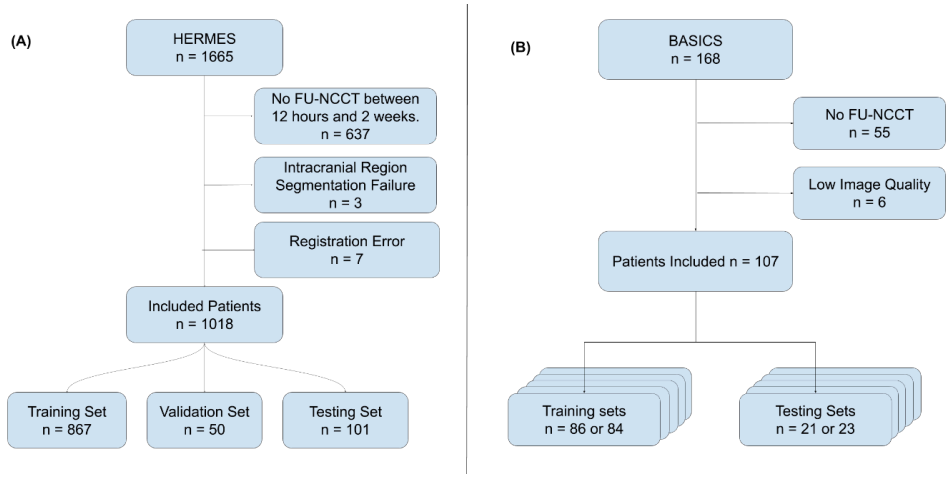


Figure 3.4: (A) Flowchart showing the exclusion criteria used for the HERMES dataset. Exclusion criteria were: No FU-NCCT available which was acquired between 12 hours and 2 weeks (n = 637), an intracranial region segmentation error (n = 3) and a registration error (n = 7). In total, 1018 patients were included. These were split into a training set (n = 876), a val-idation set (n = 50) and a testing set (n = 101). (B) Flowchart showing the exclusion criteria used for the BASICS dataset. Exclusion criteria were: No FU-NCCT available (n = 55) or an image quality that was too low (n = 6). In total, 107 patients were included. These were split into 5 training and testing sets for the five-fold cross-validation. Patients could only belong to one test set.

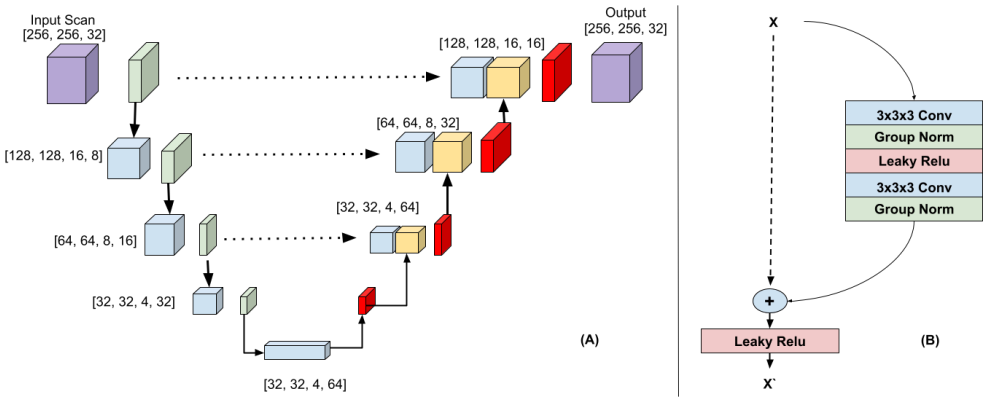


Figure 3.5: A) 3D-Unet Architecture. The downsampling path (left) consisted of 3DResNet blocks with max pooling (green). The upsampling path (right) consisted of ResNet blocks followed by transposed convolutions (red). The features created in the downsampling path are colored blue and the features created in the upsampling path are colored yellow. The dotted arrows indicate the skip connections. The feature maps from the downsampling path were concatenated to the feature maps in the upsampling path. The input image and output probability map are colored purple. (B) 3D ResNet block.

Lesion Location	**Count/Total**
No lesion	15/107
Left thalamus	33/107
Left cerebellum	40/107
Left PCA territory	19/107
Right thalamus	26/107
Right cerebellum	40/107
Right PCA territory	19/107
Midbrain	34/107
Pons	46/107
Other	4/107

Table 3.2: Lesion location in the posterior fossa, scored manually by using the PC-ASPECTS.

Method	ICC	Dice	Bias	Limits of Agreement
TL-CNN	0.88 (95% CI: 0.83-0.92)	0.25±0.07	0.84 mL	-28.7 to 30.4 mL
PCS-CNN	0.80 (95% CI: 0.72-0.86)	0.21 ±0.06	3.8 mL	-31.9 to 39.4 mL
CD-CNN	0.83 (95% CI: 0.76-0.88)	0.16±0.06	6.4 mL	-27.3 to 40.2 mL
ACS-CN	0.55 (95% CI: 0.4-0.67)	0.07±0.03	13.5 mL	-32.2 to 59.1 mL

Table 3.3: The ICC, Dice coefficients, bias and limits of agreement between the automatically quantified and manually segmented volumes, respectively for the transfer learned (TL-CNN), anterior circulation stroke (ACS-CNN), combined dataset (CD-CNN) and posterior circulation stroke (PCS-CNN) convolutional neural networks, tested on the PCS test set.

		Dice Coefficient		Bias	
Method 1	**Method 2**	**W**	**P-Value**	**W**	**P-Value**
TL-CNN	**PCS-CNN**	766	<.05	2205	.28
TL-CNN	**CD-CNN**	216	<.01	1018	<.01
TL-CNN	**ACS-CNN**	62	<.01	938	<.01
PCS-CNN	**CD-CNN**	535	<.01	1958	<.05
PCS-CNN	**ACS-CNN**	114	<.01	1350	<.01
CD-CNN	**ACS-CNN**	88	<.01	1443	<.01

Table 3.4: Wilcoxon rank sum test on pairwise differences between the Dice coefficient and the bias of the volume differences. The W-statistic and P-value are shown in this table. The TL-CNN produced a significantly greater Dice coefficient than the other methods. The PCS-CNN produced a significantly greater Dice coefficient than the CD-CNN and ACS-CNN and the CD-CNN produced a significantly greater dice coefficient than the ACS-CNN. The TL-CNN produced a significantly smaller bias than the CD-CNN and ACS-CNN, the PCS-CNN produced a significantly smaller bias than the CD-CNN and ACS-CNN and finally, the CD-CNN produced a significantly smaller bias than the ACS-CNN.

Bibliography

[141] J. L. Banks and C. A. Marotta. Outcomes validity and reliability of the modified rankin scale: Implications for stroke clinical trials - A literature review and synthesis, 3 2007.

[142] A. Barman, M. E. Inam, S. Lee, S. Savitz, S. Sheth, and L. Giancardo. Determining ischemic stroke from ct-angiography imaging using symmetry-sensitive convolutional networks. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2019-April, pages 1873–1877. IEEE Computer Society, 4 2019.

[143] R. S. Barros, M. L. Tolhuisen, A. M. Boers, I. Jansen, E. Ponomareva, D. W. Dippel, A. Van Der Lugt, R. J. Van Oostenbrugge, W. H. Van Zwam, O. A. Berkhemer, M. Goyal, A. M. Demchuk, B. K. Menon, P. Mitchell, M. D. Hill, T. G. Jovin, A. Davalos, B. C. Campbell, J. L. Saver, Y. B. Roos, K. W. Muir, P. White, S. Bracard, F. Guillemin, S. D. Olabarriaga, C. B. Majoie, and H. A. Marquering. Automatic segmentation of cerebral infarcts in follow-up computed tomography images with convolutional neural networks. *Journal of NeuroInterventional Surgery*, 12(9):848–852, 9 2020.

[144] A. M. Boers, I. Jansen, l. F. M Beenen, T. Devlin, L. san roman, J. hoe heo, M. Ribó, S. Brown, M. Almekhlafi, D. Liebeskind, J. Teitelbaum, h. F. Lingsma, W. van Zwam, P. Cuadras, M. Beaumont, M. M. Brown, a. J. Yoo, r. J. van Oostenbrugge, B. K. Menon, G. Donnan, J. louis Mas, Y. B. W M roos, C. Oppenheim, A. van der lugt, r. J. Dowling, M. D. Hill, A. Davalos, T. Moulin, N. Agrinier, a. M. Demchuk, D. K. Lopes, L. aja rodríguez, D. W. J Dippel, B. V. Campbell, P. J. Mitchell, F. Al-ajlan, T. Jovin, J. Madigan, g. W. Albers, S. Soize, F. Guillemin, V. K. Reddy, S. Bracard, J. Blasco, K. W. Muir, R. Nogueira, P. M. White, M. Goyal, s. M. Davis, H. Marquering, and c. B. M Majoie. Association of follow-up infarct volume with functional outcome in acute ischemic stroke: a pooled analysis of seven randomized trials. *J NeuroIntervent Surg*, 10:1137–1142, 2018.

[145] J. Bogousslavsky, G. V. Melle, and F. Regli. The lausanne stroke registry: Analysis of 1,000 consecutive patients with first stroke. *Stroke*, 19(9):1083–1092, 1988.

[146] S. Chen, K. Ma, and Y. Zheng. MED3D: Transfer learning for 3D medical image analysis, 2019.

[147] D. V. Cicchetti. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, 6(4):284–290, 1994.

[148] A. Clèrigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, and X. Lladó. Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. *Computers in Biology and Medicine*, 115, 12 2019.

[149] M. Goyal, B. K. Menon, W. H. Van Zwam, D. W. Dippel, P. J. Mitchell, A. M. Demchuk, A. Dávalos, C. B. Majoie, A. Van Der Lugt, M. A. De Miquel, G. A. Donnan, Y. B. Roos, A. Bonafe, R. Jahan, H. C. Diener, L. A. Van Den Berg, E. I. Levy, O. A. Berkhemer, V. M. Pereira, J. Rempel, M. Millán, S. M. Davis, D. Roy, J. Thornton, L. S. Román, M. Ribó, D. Beumer, B. Stouch, S. Brown, B. C. Campbell, R. J. Van Oostenbrugge, J. L. Saver, M. D. Hill, and T. G. Jovin. Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. *The Lancet*, 387(10029):1723–1731, 4 2016.

[150] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 12 2016.

[151] Y. C. Kim, J. E. Lee, I. Yu, H. N. Song, I. Y. Baek, J. K. Seong, H. G. Jeong, B. J. Kim, H. S. Nam, J. W. Chung, O. Y. Bang, G. M. Kim, and W. K. Seo. Evaluation of Diffusion Lesion Volume Measurements in Acute Ischemic Stroke Using Encoder-Decoder Convolutional Network. *Stroke*, 50(6):1444–1451, 2019.

[152] L. C. Langezaal, E. J. van der Hoeven, F. J. MontAlverne, J. J. de Carvalho, F. O. Lima, D. W. Dippel, A. van der Lugt, R. T. Lo, J. Boiten, G. J.

Lycklama à Nijeholt, J. Staals, W. H. van Zwam, P. J. Nederkoorn, C. B. Majoie, J. C. Gerber, M. Mazighi, M. Piotin, A. Zini, S. Vallone, J. Hofmeijer, S. O. Martins, C. H. Nolte, K. Szabo, F. A. Dias, D. G. Abud, M. J. Wermer, M. J. Remmers, H. Schneider, C. M. Rueckert, K. F. de Laat, A. J. Yoo, P.-J. van Doormaal, A. C. van Es, B. J. Emmer, P. Michel, V. Puetz, H. J. Audebert, O. M. Pontes-Neto, J.-A. Vos, L. J. Kappelle, A. Algra, and W. J. Schonewille. Endovascular Therapy for Stroke Due to Basilar-Artery Occlusion. *New England Journal of Medicine*, 384(20):1910–1920, 5 2021.

[153] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis, 12 2017.

[154] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek. The design of simpleITK. *Frontiers in Neuroinformatics*, 7(DEC), 12 2013.

[155] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellyn, and W. Eubank. Nonrigid multimodality image registration. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 1609–1620. SPIE, 7 2001.

[156] T. Moulin, L. Tatu, F. Vuillier, E. Berger, D. Chavot, and L. Rumbach. Role of a stroke data bank in evaluating cerebral infarction subtypes: Patterns and outcome of 1,776 consecutive patients from the Besancon Stroke Registry. *Cerebrovascular Diseases*, 10(4):261–271, 2000.

[157] O. Öman, T. Mäkelä, E. Salli, S. Savolainen, and M. Kangasniemi. 3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke. *European radiology experimental*, 3(1):8, 2019.

[158] S. J. Pan and Q. Yang. A survey on transfer learning, 2010.

[159] V. Puetz, P. N. Sylaja, S. B. Coutts, M. D. Hill, I. Dzialowski, P. Mueller, U. Becker, G. Urban, C. O'Reilly, P. A. Barber, P. Sharma, M. Goyal, G. Gahn, R. Von Kummer, and A. M. Demchuk. Extent of hypoattenuation on CT angiography source images predicts functional outcome in patients with basilar artery occlusion. *Stroke*, 39(9):2485–2490, 9 2008.

[160] D. Pustina, H. Branch Coslett, P. E. Turkeltaub, N. Tustison, M. F. Schwartz, and B. Avants. Automated Segmentation of Chronic Stroke Lesions Using LINDA: Lesion Identification With Neighborhood Data Analysis. 2016.

[161] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241, Munich, Germany, 2015. Springer Verlag.

[162] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 9 2014.

[163] W. J. Schonewille, C. A. Wijman, P. Michel, C. M. Rueckert, C. Weimar, H. P. Mattle, S. T. Engelter, D. Tanne, K. W. Muir, C. A. Molina, V. Thijs, H. Audebert, T. Pfefferkorn, K. Szabo, P. J. Lindsberg, G. de Freitas, L. Jaap Kappelle, A. Algra, A. M. Weber, G. A. Donnan, V. Thijs, A. Peeters, G. de Freitas, A. B. Conforto, M. Miranda-Alves, A. Massaro, P. Ijäs, T. Bogoslovsky, P. J. Lindsberg, C. Weimar, J. Benemann, K. Kraywinkel, C. Haverkamp, D. Michalski, K. Weissenborn, M. Goertler, A. Kloth, A. Bitsch, T. Mieck, J. Machetanz, P. Möller, R. Huber, S. Kaendler, C. Rueckert, H. Audebert, R. Müller, B. Vatankhah, T. Pfefferkorn, T. E. Mayer, K. Szabo, C. Disque, O. Busse, C. Berger, W. Hacke, Y. Schwammenthal, D. Orion, D. Tanne, M. Bergui, E. Pozzati, W. J. Schonewille, A. Algra, L. J. Kappelle, G. J. Luijckx, P. Vroomen, M. D. Vergouwen, Y. Roos, J. Stam, P. Bienfait, F. E. de Leeuw, P. de Kort, D. Dippel, J. Pagola, M. Ribo, C. Molina, A. Gonzales, A. Gil-Peralta, B. Norrving, M. Arnold, U. Fischer, J. Gralla, H. Mattle, G. Schroth, P. Michel, S. T. Engelter, S. Wetzel, P. Lyrer, J. Gandjour, N. Michael, R. Baumgartner, B. Tettenborn, H. Hungerbuehler, T. Baird, K. Muir, C. A. Wijman, A. Finley Caulfield, M. Lansberg, N. Schwartz, C. Venkatasubramanian, Z. Garami, S. Bogaard, F. Yatzu, and J. Grotta. Treatment and outcomes of acute basilar artery occlusion in the Basilar Artery International Cooper-

ation Study (BASICS): a prospective registry study. *The Lancet Neurology*, 8(8):724–730, 8 2009.

[164] N. Tomita, S. Jiang, M. E. Maeder, and S. Hassanpour. Automatic post-stroke lesion segmentation on MR images using 3D residual convolutional neural network. *NeuroImage: Clinical*, 27, 1 2020.

[165] R. Vallat. Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31):1026, 11 2018.

[166] E. J. Van Der Hoeven, W. J. Schonewille, J. A. Vos, A. Algra, H. J. Audebert, E. Berge, A. Ciccone, M. Mazighi, P. Michel, K. W. Muir, V. Obach, and V. Puetz. The Basilar Artery International Cooperation Study (BASICS): study protocol for a randomised controlled trial. Technical report, 2013.

[167] K. Villringer, M. Florczak-Rzepka, U. Grittner, P. Brunecker, H. Tepe, C. H. Nolte, and J. B. Fiebach. Characteristics associated with outcome in patients with first-ever posterior fossa stroke. *European Journal of Neurology*, 25(6):818–824, 6 2018.

[168] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *Journal of Digital Imaging*, 31(3):290–303, 6 2018.

[169] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 4, pages 3320–3328, 11 2014.

[170] Y. Yu, Y. Xie, T. Thamm, E. Gong, J. Ouyang, C. Huang, S. Christensen, M. P. Marks, M. G. Lansberg, G. W. Albers, and G. Zaharchuk. Use of Deep Learning to Predict Final Ischemic Stroke Lesions From Initial Magnetic Resonance Imaging. *JAMA network open*, 3(3):e200772, 3 2020.

[171] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128, 7 2006.

[172] S. F. Zaidi, A. Aghaebrahim, X. Urra, M. A. Jumaa, B. Jankowitz, M. Hammer, R. Nogueira, M. Horowitz, V. Reddy, and T. G. Jovin. Final infarct volume is a stronger predictor of outcome than recanalization in patients with proximal middle cerebral artery occlusion treated with endovascular therapy. *Stroke*, 43(12):3238–3244, 12 2012.

[173] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang. Models genesis: generic autodidactic models for 3d medical image analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11767 LNCS, pages 384–393, 2019.