Automatic Collateral Scoring from 3D CTA Images

Jiahang Su, Lennard Wolff, Adriaan C.G.M van Es, Wim van Zwam, Charles Majoie, Diederik W.J. Dippel, Aad van der Lugt, Wiro J Niessen and Theo van Walsum

Abstract—The collateral score is an important biomarker in decision making for endovascular treatment (EVT) of patients with ischemic stroke. The existing collateral grading systems are based on visual inspection and prone to subjective interpretation and interobserver variation. The purpose of our work is the development of an automatic collateral scoring method. In this work, we present a method that is inspired by human collateral scoring. Firstly, we define an anatomical region by atlas-based registration and extract vessel structures using a deep convolutional neural network. From this, high-level features based on the ratios of vessel length and volume of the occluded and the contralateral side are defined. Multi-class classification models are used to map the feature space to a four-grade collateral score and a quantitative score. The dataset used for training, validation and testing is from a registry of images acquired in clinical routine at multiple medical centers. The model performance is tested on 269 subjects, achieving an accuracy of 0.8. The dichotomized collateral score accuracy is 0.9. The error is comparable to the interobserver variation, the results are comparable to the performance of two radiologists with 10 to 30 years of experience.

Index Terms—collateral score, deep learning, 3D CTA brain image, MR CLEAN

I. INTRODUCTION

A. Clinical Background

Endovascular Treatment (EVT) improves outcome in patients with acute ischemic stroke due to intracranial large vessel occlusion (Goyal et al. [1]). Selection of eligible patients for EVT is important because not all patients benefit and the treatment is not without risk (Venema et al. [2]). Brain tissue at risk can survive longer in the presence of collateral circulation, which is a network of arterial anastomoses that provides blood flow to brain tissue when the principal conduits fail to meet demands (Liebeskind [3]).

The MR CLEAN trial, a multicenter, randomized trial of EVT versus no EVT, showed that baseline computed tomographic angiography (CTA) collateral status modified the treatment effect (Berkhemer et al. [4]): patients with higher collateral score will most likely have better treatment outcome. A clinical decision tool based on multiple baseline clinical and imaging characteristics for individualized predictions of the effect of EVT has been developed and includes grade of collateral circulation as prognostic and predictive marker (Venema et al. [2]). Several collateral status grading systems exist, all based on visual scoring using coarse classification criteria (Seker et al. [5]). Such visual scoring systems suffer from subjective interpretations leading to inter- and also intra-observer variation. An automated scoring system could facilitate an objective and reproducible assessment of the cerebral collateral status. In our work, we use the four-grade score that was proposed by Tan et al. [6] as it has proven correlation with outcome and effect of EVT. The definition of this 4-grade score system is:

- 0: absent collaterals (0% filling in occluded territory)
- 1: poor collaterals (>0% and ≤50% filling in occluded territory)
- 2: moderate collaterals (>50% and <100% in occluded territory)
- 3: good collaterals (100% filling in occluded territory)

Fig. 1 shows example images for four different collateral scores in Maximum Intensity Projection (MIP).

B. Related Work

Collateral status scoring relies on the difference between arterial trees in the middle cerebral artery (MCA) territory of the occluded side and its contralateral side. Therefore, vessel segmentation is an essential step in our application. Kirbas et al. [7] and Lesage et al. [8] provide a review of many vessel segmentation approaches that have been developed over the past decades. For cerebral blood vessel segmentation, Meijls et al. [9] summarized non-convolutional neural network based algorithms with respect to methods, image modality, and cerebral vessel segments. For example, Manniesing et al. [10] utilized a level set approach to detect the circle of Willis (CoW) in 3D CTA images, Schaan et al. [11] utilized a Bayesian tracking approach to segment the internal carotid artery (ICA) in 3D CTA images, Robben et al. [12] utilized graph connectivity in combination with a tracking approach to obtain the label and vessel structure of the CoW in Magnetic Resonance Angiography (MRA) images, and Meijls et al. [9]
utilized a random forest classifier and histogram to extract the complete vessel structure from 4D CTA images.

Nowadays, convolutional neural networks (CNNs) have demonstrated in general superior performance for many medical image segmentation tasks. This trend is also witnessed in vessel segmentation, Moccia et al. [13] summarized vessel segmentation using conventional and CNN based methods. Sanches et al. [14], Livne et al. [15] and Kandil et al. [16] employ 3D U-net based CNN model to extract the cerebral vessel structure from 3D Time-of-flight (TOF) MRA images, Meij and Manniesing [17] use a 3D U-net based CNN model to extract cerebral vessel structure from 4D CTA images and further separate the arterial and vein structure by its spatial features, Tetteh et al. [18] utilize the 3D U-net framework and replace the 3D convolution with a 2D cross-hair filter to segment the cerebral vessel structure and centerline in TOF MRA data. None of these approaches quantifies the collateral status.

To the best of our knowledge, only Boers et al. [19] published an automatic collateral scoring method. The region of interest (ROI) is defined by a probability density map that was generated from an atlas build from lesion segmentation from follow-up CT images. The 3D Frangi filter (Frangi et al. [20]) with visually tuned parameters and a threshold of 200 Hounsfield units (HU) was applied to extract vessel structure in a pre-defined region of interest. The computed feature is the ratio of vessel volume with intensity between the occluded side and the contralateral side. The method was assessed on 59 subjects from which their follow-up non-contrast CT scan was used to construct the probability density map. The method was assessed on CTA images of patients with an occlusion in the M1 segment (for the detailed vessel segments in MCA territory please refer to Fig. 6) of the MCA territory, a maximal slice thickness of 1 mm and full coverage of the intracranial region.

II. Method

The proposed method starts with a pre-processing step to define the anatomical regions of interest: we use an atlas-based approach which was developed by Peter et al. [21] to obtain a 3D CTA brain image $I_b$, an MCA probability density map $M$ and a hemisphere map $H$. This method takes the CTA image and an atlas image as input. This step is followed by a deep learning based segmentation of the brain vasculature (centerlines). In the final step, the output of the previous stages is transformed into a quantitative score, and a collateral class score. The algorithm overview is shown in Fig. 2. Each of the steps is detailed below.

A. Pre-processing: Anatomical Region Definition

In this work, we propose a three-stage algorithm to compute a collateral score. The collateral scoring method was assessed on 269 subjects.

Data preprocessing and vessel centerline segmentation are explained in Section II, followed by feature design and a multi-label classification model for collateral scoring. The data set, collateral score reference standard and annotation strategy are described in Section III, the experiments and results are detailed in Section IV, followed by discussion in Section V, conclusions are drawn in Section VI.

C. Contributions and Organization of Our Work

In this work, we propose a three-stage algorithm to compute a collateral score. The collateral scoring method was assessed on 269 subjects.

Data preprocessing and vessel centerline segmentation are explained in Section II, followed by feature design and a multi-label classification model for collateral scoring. The data set, collateral score reference standard and annotation strategy are described in Section III, the experiments and results are detailed in Section IV, followed by discussion in Section V, conclusions are drawn in Section VI.
of SPM toolbox (Friston [22]). For the MCA region, we use a MCA probability density map (values ranging from 0 to 1) that represent the likelihood of MCA vessels present based on 50 healthy MCA vasculatures from the BraVa (Wright et al. [23]) dataset. This was done by an affine and deformable symmetric diffeomorphic image registration of each subject from the BraVa dataset to the CT-MR template, following by a normalized sum of transformed individual MCA arterial trees. The hemisphere atlas is a three-value map that represents the left hemisphere, the right hemisphere and the background. All anatomical regions are defined in the space of the average CT atlas.

After registration of the CT atlas to the subject CTA image, the binary brain mask, the MCA probability density map and the hemisphere map are transformed to the subject CTA image space. The generated binary brain mask may fail to include vessel structures near the skull base. Therefore, to fine-tune the CT image space binary mask, an iterative morphological dilation with a 3D spherical structuring (radius equals 1 voxel) was applied to this binary mask. In each iteration, we remove dilated voxels if its corresponding 3D CTA voxel value exceed 850 Hounsfield unit (HU). The number of iterations is defined by the maximal gap, 5 mm, divided by the voxel dimension. A brain tissue image is constructed by multiplying the binary mask with the CTA image. After this masking step, the CTA brain image $I_b$ is normalized with min-max normalization.

### B. Deep Learning for Vessel Extraction

In the second step, the vasculature is extracted from the brain-masked and normalized CTA image. We opt for a deep learning approach, which trains a detector for vessel-like structures based on labeled training images; such approaches have been demonstrating excellent performance in the last years for many medical imaging tasks. More specifically, we intend to extract vessel structure with an encoder-decoder architecture (the U-Net model proposed by Ronneberger et al. [24]); this architecture is simple and still very effective (Isensee et al. [25]). Note that for the final goal of obtaining collateral scores, we do not require a very precise segmentation of the vessels, rather we want to highlight the vessel centerlines such that the vasculature can be quantified in subsequent steps. For this 3D vessel extraction task, we use a standard 3D U-Net model as described in Isensee et al. [26] to which we make modifications in the training process to tailor it to our vessel centerline segmentation task. Next, we will detail the network architecture, and the data preparation for the training.

1) **Network Architecture:** Our network architecture is shown in Fig. 3. The network utilizes the classic encoder and decoder architecture with a depth of 5 layers. The input data of the proposed model is a cube of $64 \times 64 \times 64$ voxels with 28 convolutional filters in the first stage. The number of filters was determined according to a set of experiments we describe in Section IV-B. We apply instance normalization to each convolution layer. In the encoder path, activations are calculated using residual blocks. A residual block is a type-1 block in Fig. 3, including two cascaded 3x3x3 3D convolutional layers and the identity short connection (dashed lines in Fig. 3). This combination is similar to the context module described in Isensee et al. [26], however, in our cases, we didn’t use dropout layer in between. Two cascaded 3x3x3 3D convolutional layers with a stride of 2 (type-2 block in Fig. 3) are added in front of residual blocks in order to obtain more abstract feature maps as the encoder goes deeper. Leaky rectified linear units (ReLU) (Maas et al. [27]) are used as the activation function. In the concatenated decoder path, deconvolutional layers are constructed with an extra deep supervision (Kayalibay et al. [28]) path. This can avoid information loss and vanishing gradients in each convolutional layer. The proposed model uses a sigmoid activation layer in the final step to output a 3D voxel-wise vessel probability map.

2) **Data Preparation:** For each brain, we construct ground truth voxel trees. The vessel tree is a 3D binary mask resulting from the vessel centerline annotation process described in Section III-C. Then we split the whole brain and vessel tree into 3D cubes of 128x128x128 voxels for both training data and validation data. During the training process, we extract 64-voxel 3D cubes out of 128x128x128 voxels. In the validation process during training, we use 128x128x128 voxels as input data size for convenience. Data augmentation is applied to obtain different training images in each iteration. In the data preparation stage, we did not resample the image into common space as the training data is representative for the whole dataset and exhibits little variation in slice spacing and pixel size.

3) **Deep Learning Post Processing:** The proposed CNN model outputs a vessel probability map with values between 0 and 1. We threshold the probability map to obtain a binary vessel map $B_0$ (Fig. 4b). The threshold value is found by optimizing a Dice cost function on the deep learning validation dataset. There are some small isolated parts (mostly false positive parts) in the predicted vessel map $B_0$ and connectivity based noise removal (Risser et al. [29]) is applied to remove these small isolated parts. This results in a binary vessel tree map denoted by $B$ (Fig. 4c).

### C. Quantification

The purpose of the quantification step is to compute a collateral score from the results of the previous processing steps. Human collateral scoring is based on comparing the amount of vessels visible in the affected and non-affected side, and we follow a similar strategy: we compare the affected and non-affected hemisphere of subjects using a combination of the binary vessel structure from deep learning model output $B$, the corresponding MCA probability density map $M$ and hemisphere map $H$ as shown in Fig.2. We assume that it is known a priori (from clinical symptoms) which hemisphere (left or right) is affected. Based on this information and the hemisphere map, we generate an affected side binary map $H_A$ and a non-affected side binary map $H_N$. With this information, we compute four different ratios, each representing a different aspect of the vasculature, as detailed below.

1) **Volume:** An obvious quantification is difference in the number of vessels between the affected and the non-affected hemisphere. Assuming there are more vessels visible in the
non-affected hemisphere, the ratio of vessel volumes (affected divided by non-affected) should give a number between 0 (no vessels visible in the affected side) to 1 (same amount of vessels visible in both sides). We propose two variants of this comparison: one where only the volume (number of voxels) is taken into account, and one where each voxel is weighted with its intensity \( r \) and \( r_{vi} \) respectively:

\[
\begin{align*}
    r_u &= \frac{\sum_{p \in B} M(p) \cdot H_A(p)}{\sum_{p \in B} M(p) \cdot H_N(p)}, \\
    r_{vu} &= \frac{\sum_{p \in B} I(p) \cdot M(p) \cdot H_A(p)}{\sum_{p \in B} I(p) \cdot M(p) \cdot H_N(p)},
\end{align*}
\]

where \( B \) is the set of voxels that are 1 in isolated part removed binary vessel tree \( B \), \( p \) is a voxel and \( I \) is the image intensity.

2) Length: In addition to volume, we also consider vessel length. For the vessel length computation, a medial axis skeletonization approach that was developed by Lee et al. [30] is applied first. This step yields a pixel-wise skeleton structure \( S_0 \), a 26-connected structure (Fig. 5 left) which is a basic representation of the vessel network. Next, the vessel segments are determined from the vessel skeleton using a tree topology approach employed in the work of Risser et al. [29].

Subsequently those vessel segments were fitted by 3D spline curves and further smoothed (Garcia [31], Garcia [32]) and interpolated (Fig. 5 right), yielding a set of world coordinates \( S \). For all points \( p \) from \( S \) we obtain the corresponding MCA probability value \( M(p) \) from the 3D MCA probability density map. In the same way, we obtain intensity values \( H_A(p) \), \( H_N(p) \), and \( I(p) \). The ratios \( r_1 \) and \( r_{li} \) are accumulated values over all points (weighted with the mean distance \( w \) between the point and its neighbours) in the affected hemisphere and the ones in the non-affected hemisphere:

\[
\begin{align*}
    r_1 &= \frac{\sum_{p \in S} M(p) \cdot H_A(p) \cdot w(p)}{\sum_{p \in S} M(p) \cdot H_N(p) \cdot w(p)}, \\
    r_{li} &= \frac{\sum_{p \in S} I(p) \cdot M(p) \cdot H_A(p) \cdot w(p)}{\sum_{p \in S} I(p) \cdot M(p) \cdot H_N(p) \cdot w(p)}.
\end{align*}
\]

3) Multi-label Classification: In the last step, multi-class classification is used to predict collateral score \((0,1,2,3)\) from an input feature vector \( r = [r_u, r_{vu}, r_1, r_{li}] \). We start from the baseline model, in which we take the median of feature vector \( r \) as input and define the threshold value by utilizing the clinical definition of collateral score. We then define our second model by using a support vector classifier (SVC) with linear kernel to find the optimal threshold value for the median of feature vector \( r \). For the third method, we use the complete feature vector \( r = [r_u, r_{vu}, r_1, r_{li}] \) as input to a random forest classifier. Finally, we use ordinal regression with the complete feature vector for ordered categorical prediction.
III. DATA AND ANNOTATION

A. Data Overview

The images used for training and assessing the methods were obtained from the MR CLEAN Registry (Jansen et al. [33]) and MR CLEAN trial (Berkhemer et al. [34]). The MR CLEAN Registry is an on-going registry that contains patients who underwent endovascular treatment at a stroke intervention center in the Netherlands since March, 2014. The CTA images are acquired in clinical routine at several different sites, and thus there is large variation in image quality, as well as in imaging equipment and acquisition protocols (contrast phase, brain coverage).

Hence, in order to get a representative set of images, the following selection criteria have been used to select images in this study:

- **Spatial resolution**: The average diameter of the M1 segment of the MCA region is 3.1 ± 0.4 mm and of the M2 segment 2.4 ± 0.4 mm according to Rai et al. [35]. The spatial resolution should be sufficient to visualize the major arterial tree. Therefore, we only select images of which the slice thickness is smaller than 1.5 mm; additionally we require the slice spacing to be smaller than or equal to the slice thickness.

- **Contrast phase**: Five different contrast phases have been previously defined by Rodriguez-Luna et al. [36]: early arterial, peak arterial, equilibrium, peak venous, and late venous, based on the image intensities in the contralateral ICA and the transverse sinus. In the early arterial phase, collateral vessels are likely not enhanced yet. In the late venous phase, the venous structures are more pronounced than arterial structures. Therefore, in this study we only select images with peak arterial, equilibrium or the early venous phase.

- **Image quality**: Image quality in MR CLEAN Registry is rated by the core lab into the following categories: good image quality, moderate image quality and bad image quality. A good quality image implies that the image is sufficiently informative for radiologist to rate. We similarly included good quality images in our analyses.

- **Brain coverage**: Brain coverage, the image should cover at least half of the vertical distance between skull base and vertex.

At the start of the current study, baseline CTA data of 1594 subjects had been collected in the MR CLEAN Registry. These data was acquired from 16 March 2014 till 15 June 2016. Of these 1594 subjects, the images of 1058 subjects had good image quality (based on MR CLEAN Registry core lab readings). 736 subjects fulfill both image quality and contrast phase criteria. At the end, 585 subjects fulfill all selection criteria. From this set of 585 subjects, 49 subjects were manually selected for annotation and training of the CNN. From the remaining 536 subjects, we randomly selected around half of these subjects (270 images) for our study. This number was assumed to be sufficient for our evaluation, and also reduced the amount of work for obtaining a consensus score compared to using the full set. Those 270 subjects were originally from 14 intervention centers with different vendors (mainly from Philips, Siemens, Toshiba and GE). For all cases, we obtained the occlusion side and occlusion position and initial collateral score from the registry information. From those 270 subjects, one subject was additionally excluded which was considered to have insufficient image quality by the expert readers. We did not have additional selection criteria on occlusion location. However, only 6 subjects with A1 or A2 occlusions were found among the 1574 MR CLEAN registry subjects, and our final set of images did not contain any A1 or A2 occlusions.

In addition, for training we added five collateral score 0 cases from the MR CLEAN Trial. Collateral score 0 is rare, and due to the random selection, only two images with collateral score 0 were in our initial selection. The additional five subjects were only used in the classifier training process. The accuracy of the algorithm was thus evaluated on the randomly selected 269 cases.

B. Collateral Score Reference Standard

To get a consistent and reliable collateral score reference standard for this study, collateral scoring was performed by three radiologists from different medical centers with 10 to 30 years of experience. Two radiologists rated the collateral score independently and the third radiologist independently rated the cases in which there was disagreement by the first two radiologists. The radiologists were asked to rate the collateral status according to the criteria of Tan et al. [6].

In this 269 subjects, the two independent raters had an interobserver agreement of 0.64, and their scores compared to the consensus score were 0.81 and 0.82. The details of the 269 test subjects are listed in the Table I. Fig. 6 shows the location of occlusion in the vessel segments that are listed in Table I.

![Fig. 6: An example of vessel segments in right hemisphere. The vessel segments include the ICA segment, M1, M2 and the more distal part of the MCA territory, for simplicity, we have combined M2 and the more distal part into M2.](image-url)
TABLE I: Data distribution of 269 test subjects.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slice thickness</td>
<td>[0.5-0.75mm]</td>
</tr>
<tr>
<td></td>
<td>[0.75-1.5mm]</td>
</tr>
<tr>
<td></td>
<td>[1-1.5mm]</td>
</tr>
<tr>
<td>Acquisition phase</td>
<td>peak arterial phase</td>
</tr>
<tr>
<td></td>
<td>equilibrium phase</td>
</tr>
<tr>
<td></td>
<td>peak venous phase</td>
</tr>
<tr>
<td>Collateral score</td>
<td>score 0</td>
</tr>
<tr>
<td></td>
<td>score 1</td>
</tr>
<tr>
<td></td>
<td>score 2</td>
</tr>
<tr>
<td></td>
<td>score 3</td>
</tr>
<tr>
<td>Occlusion location</td>
<td>ICA</td>
</tr>
<tr>
<td></td>
<td>M1</td>
</tr>
<tr>
<td></td>
<td>M2 and above</td>
</tr>
<tr>
<td></td>
<td>No occlusion found</td>
</tr>
</tbody>
</table>

C. Data Annotation

Training and assessment of the deep learning based vessel extraction requires (manually) annotated images. In total 9 whole brains and 40 cubes of 128x128x128 voxels were annotated. The 9 subjects were selected by a radiologist as being cases representative for daily clinical practice. The 40 subjects for cube annotation were selected to cover large variation of image quality and acquisition parameters, as well as collateral scores. The 40 cubes were randomly selected from the intracranial region of 40 subjects. Manual annotation was performed using an in-house developed tool build with MevisLab. The annotation task was performed by 1 pre-med student, 3 experienced physicians and the first author of this paper. The purpose of the annotation task was to label centerlines of all intracranial vessels. Annotation points needed to be placed in the center of candidate vessel structures, after which semi-automated processing (shortest path connectivity, followed by a graph cut segmentation, yielding contours orthogonal to the centerline (Schaap et al. [37]), after which the real centerline defined as the centers of the segmentation result) was applied. The labelled region also include vessels running through the skull base and isolated vessel structures. Fig. 7 shows the result of a whole brain annotation.

IV. EXPERIMENTS AND RESULTS

A. Implementation

The method consists of three parts: the pre-processing (atlas-based registration), deep learning based vessel centerline extraction, and classification. The deep learning model is written using the Keras and Tensorflow frameworks. Model training and validation are implemented on a local PC equipped with one NVIDIA Titan Xp GPU and 64 Gb of RAM. The pre-processing, quantification and classification of proposed method were implemented in MATLAB 2018a.

Registration of the CT atlas to the CTA image was performed with ANTs (Avants et al. [38], Klein et al. [39]), following a previously described CT atlas based registration [21] that consists of a two-step approach: an initial rigid registration followed by a diffeomorphic non-rigid registration.

In the deep learning model training process, the network was trained with a batch size of 8 and 50 steps per epoch. In each epoch, we iterate twice over all 200 training images of 128x128x128 voxels: once with random shift and flipping along each axis and once with elastic deformation. We also introduced additive Gaussian noise at the input layer for regularization. The standard Dice score served as the loss function. We chose the root mean square propagation (RMSProp) (Tieleman and Hinton [40]) optimizer with an initial learning rate of 0.1, and halved it every 10 epochs. We stopped training after 300 epochs, as the learning rate was approaching zero.

In the multi-label classification part, the baseline model has fixed threshold values. The three threshold values($\theta_1$, $\theta_2$, $\theta_3$) were determined by the clinical definition of collateral score (Tan et al. [6]). We introduced a small margin to the collateral 0 case, since, in practice, there were always some vessels in the occluded side. Therefore, we define the collateral 0 case as less than 10%. The other three threshold values follow the clinical definition, i.e. $\theta_1 = 10\%$; $\theta_2 = 50\%$; $\theta_3 = 100\%$. For the random forest classifier, we used five fold cross validation to optimize the maximal depth of the trees, the number of trees in forest and the number of features used.

B. Vessel Extraction Model Training and Hyperparameter Optimization

In the first experiment, we trained the convolutional neural network for vessel extraction. In this experiment, we first investigated the performance of 3D U-Net model with different hyperparameters and configurations. Compared to a standard 3D U-Net (Isensee et al. [25]) (our baseline model), we first assessed the model performance enhanced by deep supervision and context modules, and subsequently also assessed the added value of varying the number of filters at convolutional layers.

The annotation dataset was randomly divided into a training and validation dataset. The training dataset consisted of 7 whole annotation brains and 20 cubes. The validation dataset consisted of the other 2 brains and 20 cubes. In order to guarantee the continuity and completeness of the centerline structure in the training process, we dilated every single-pixel centerline from manual annotation with a 3x3x3 square structuring element. The resulting ground truth image are shown in Figure. 7.

The Dice score on the validation dataset was used to measure model performance. In the first experiment, we assessed the added value of the various configurations. For this
experiment, 28 input filters were used, as this was the maximal number that fitted into our GPU memory. The performance of the baseline model is shown in the first row of Table II. Improvement in Dice score is observed when adding deep supervision and the context module to the baseline model. A combination of all components (last row of Table II) shows the highest validation score. We also performed a paired T-test for two average values $\mu_i$ and $\mu_i$ ($i = 2, 3, 4$), where $\mu_i$ is the average of Dice score on all validation subjects with 3D U-net configuration and $\mu_i$ the average dice score corresponding to the network configuration in $i$-th entry of Table II. Whereas the improvement of the final configuration is not statistical significant at the common 0.05 level, the trend in the Table II is clear and we attribute the lack of statistical significance to the limited validation set size. We therefore choose the 3D U-net with deep supervision and context module for the subsequent experiments.

<table>
<thead>
<tr>
<th>Network Configuration</th>
<th>Dice mean</th>
<th>Dice std</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D U-net + deep supervision</td>
<td>0.53</td>
<td>0.19</td>
<td>-</td>
</tr>
<tr>
<td>3D U-net + context module</td>
<td>0.54</td>
<td>0.19</td>
<td>0.69</td>
</tr>
<tr>
<td>3D U-net + deep supervision + context module</td>
<td>0.56</td>
<td>0.16</td>
<td>0.07</td>
</tr>
</tbody>
</table>

TABLE II: Deep learning test result for different configurations.

Whereas the number of input filters increases model capacity, it also greatly affects the number of parameters of the model, and thus may impact the training process. In the next experiment, we therefore vary the number of input filters from 4 to 28 for the configuration in the last row of Table II. The learning curves are shown in Fig. 8. As the number increases, both training and validation scores increase. This trend ends when the filter number goes beyond 24, where the improvement to the test score is marginal. Going from 24 filters to 28 filters, the test score only improves with less than 1%, compared to an improved of 4% when going from 4 filters to 8 filters.

Next, we evaluated the proposed collateral scoring method. First, we assessed the accuracy of the three proposed models. Then, we assessed the accuracy of proposed methods applied to a dichotomized decision based on collateral score. In the accuracy test for collateral scoring we evaluated the baseline model, a linear support vector classifier (SVC) with single feature, ordinal regression with four features, and a random forest model as describe in Section II-C3. We used the consensus score as ground truth label.

In the baseline model, we applied threshold values $\theta_1$, $\theta_2$, $\theta_3$ to the median of $r = [r_v, r_{vi}, r_1, r_{i1}]$ and derived the collateral scores 0, 1, 2, and 3. Similarly, we use the median of $r = [v, r_{vi}, r_1, r_{i1}]$ with linear SVC model to find another set of threshold values that maximize the collateral score test accuracy on 269 subjects. The average threshold values for the SVC model were: $\theta_1 = 7\%$; $\theta_2 = 55\%$; $\theta_3 = 99\%$. We use ordinal regression with $r = [r_v, r_{vi}, r_1, r_{i1}]$. The averaged odds ratios of four features are [1.17, 0.9, 0.92, 1.15]. We further evaluated the added value of random forest in terms of accuracy. We explored the feature vector $r$ starting from a single feature towards combined features and tested all 16 possible combinations. We performed a nested cross-validation with 20 splits at the outer level and 5-fold cross-validation for parameter tuning. More specifically, we first randomly splitted the 274 subjects into 20 subsets with a stratified sampler. Then we trained a model with data from 19 subsets and tested on 1 left-over subset. During training, we use 5-fold cross validation for hyper-parameter optimization of feature vector $r$. In the end, we have 20 models with similar performance but with different parameter settings. In this way, we could fully utilize 269 subjects for testing and reduce the possible bias caused by a smaller test dataset.

The average Dice is 0.56 with a minimum of 0.42 and a maximum of 0.67. There are two outliers (0.27 and 0.28). Fig. 9, shows two example vessel trees extracted by the proposed deep learning model.
TABLE III: The accuracy, dichotomized accuracy of three proposed methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>Dichotomized accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>median + fix (baseline)</td>
<td>0.75</td>
<td>0.87</td>
</tr>
<tr>
<td>median + SVC</td>
<td>0.76</td>
<td>0.87</td>
</tr>
<tr>
<td>ordinal regression</td>
<td>0.75</td>
<td>NA</td>
</tr>
<tr>
<td>logistic regression</td>
<td>NA</td>
<td>0.89</td>
</tr>
<tr>
<td>random forest</td>
<td>0.80</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Furthermore, in the MR CLEAN trial, substantial treatment effects were reached in patients with collateral scores of 2 or 3, whereas in patients with scores of 0 and 1, outcomes were poor and treatment effects small. Therefore, we focus on the accuracy of dichotomized prediction, wherein scores 2 and 3 are merged into one class, and score 0 and score 1 are merged into another class. We follow the same experimental setup as in the previous accuracy test. We use logistic regression with \( r = [r_v, r_{vi}, r_i, r_{li}] \) for this dichotomized binary classification problem. The performance of all four models is similar. The random forest classifier with features \( r_v \) and \( r_{li} \) performs slightly better. The overall dichotomized accuracy is listed in the dichotomized score column of Table III. For comparison, clinician 1 and clinician 2 have an accuracy of 0.91 and 0.90 respectively in this case. We further assessed the performance of this binary class classifier by plotting receiver operating characteristic (ROC) curve of our proposed methods. Fig. 10 shows the ROC of the baseline model, and the performance of the clinicians, linear SVC and random forest classifier.

![ROC curve](image)

Next, we investigated the misclassified subjects. Fig. 11 shows the confusion matrix test results on 269 subjects, for the two radiologists and the random forest classifier versus the consensus score. In total, the random forest classifier predicts an incorrect collateral score for 53 subjects; 26 out of these 53 subjects have predictions that are consistent with one of the two clinicians. The baseline model predicts an incorrect collateral score for 66 subjects; 30 out of these 66 subjects have agreement with one clinician. Finally, in Table IV we present a subgroup analysis.

The computation time for one brain with 0.5 mm slice thickness is about 15 min. Of this, the atlas to CTA space registration, MCA map, hemisphere map transformation with binary skull dilation takes 12 min, of which most time is spent on the Ants registration. It is likely that this can be optimized for clinical applicability. The vessel extraction is around 0.5 min. The quantification part in total is around 2.5 min. Those times were obtained using an implementation that was not optimized for computation time.

![Confusion matrix](image)

**Fig. 11:** Confusion matrix of clinician 1 / clinician 2 / random forest versus consensus score.

### V. DISCUSSION

In this work, we proposed an automatic collateral scoring method that is inspired by human visual collateral scoring. In the method, we compute the collateral score by comparing the difference of vessel structures in the occluded hemisphere and contralateral hemisphere. Vessel structure extraction is important for collateral status quantification. Therefore, we first investigated the performance of CNN and found a positive effect of context module and deep supervision on the performance of the 3D U-net vessel centerline segmentation model. We further evaluated network performance with an increased number of filters at convolutional layers. On average, we achieve a Dice score of 0.56 on the validation dataset. Whereas this might seem low, a value of 0.56 is reasonable for thin linear structures in 3D, as they contain a large proportion of boundary voxels. For such long thin structures, a single pixel shift in a direction orthogonal to the structure have a major impact on the Dice score.

In the collateral scoring assessment, the baseline model with using the median of feature vector \( r = [r_v, r_{vi}, r_i, r_{li}] \) and fixed threshold values achieved an accuracy of 0.75 and a dichotomized accuracy of 0.87. This demonstrates that the features are relevant. We observed that the misclassified subjects are mostly at the border of decision boundary. The average error distance is 0.25. The error is computed from the floating point score, and represents the distance to the closest value of interval of the correct collateral score. For example, \([0,1]\) corresponding to collateral score 0. On this scale, the error made in classification is the distance to the closest border of the correct class. The random forest model on average achieves an accuracy of 0.8, which outperforms the baseline model on the border cases. With two principle features \( r' = [r_v, r_{li}] \), the random forest model achieves a dichotomized accuracy of 0.9, which is comparable to two clinicians (0.91 and 0.9).
In Table IV, a subgroup analysis is presented for the acquisition phase, collateral score, slice thickness and occlusion position. The accuracy does not seem to depend on acquisition phase or collateral score. It also shows that collateral scoring is less accurate for more distal occlusions. This can be explained by the fact that a smaller region is affected in more distal occlusion. This trend can also be observed in the human scoring. In terms of slice thickness, both error rates increase along the slice thickness. When the slice thickness increases, the vessel structure is less pronounced in the 3D CTA image.

The anatomical regions are defined by a conventional atlas-based registration method due to the fact that the MCA probability density map is an essential element of the feature computation in the quantification step. In order to align the MCA probability mask, we need to first register the CT template to the CTA space. The registration result then can similarly be used to bring the hemisphere and brain segmentation to the patient space.

Collateral scoring involves only the arterial tree. In our approach, we do not discriminate between arteries and veins in the centerline segmentation. Application of the MCA territory mask, which was build from arterial trees of 50 subjects, removes some of the venous structures from the segmentation. Any remaining venous structures may affect the final quantification. As the amount of veins in the remaining region generally is small in the region of interest especially in peak arterial, equilibrium and early venous phase, and as their presence is expected to be symmetric, the remaining veins may have a minor effect on the subsequent quantification, which is demonstrated by our current results. Still, including an artery-vein separation may be an interesting direction for future research.

End-to-end training might have been an alternative approach to computing collateral scores. We chose for a slightly more conventional approach for four reasons. First, the data and corresponding ground truth required for training may be need to be larger than the set we are using now. This would require additional expert radiology screening. Second, such approach generally require a network that is trained with full-size CTA images whereas our current vessel segmentation is trained with 3D patches. Such a training would be challenging to commonly available GPUs. Thirdly, the result of such end-to-end training is difficult to interpret, whereas our approach also gives insight in the vessel segmentation and quantification on which the scoring is based. Finally, we aim for a more quantitative analysis of collaterals with a tool that provides a continuous output instead of a semiquantitative scale with 4 items.

In comparison with Boers et al. [19], the dataset and the level of ground truth are different: Boers et al. [19] used the data from MR CLEAN Trial (Berkhemer et al. [34]) and we used dataset from MR CLEAN Registry (Jansen et al. [33]). Their collateral score was derived by direct use of feature $r_{ct}$, and there was no direct assessment over collateral score; instead they perform a correlation test between their single feature and the manually obtained collateral score with a Spearman correlation test. The result on 59 subjects was a Spearman $\rho = 0.68$, $p < 0.001$. For our method, a Spearman correlation $\rho = 0.80$ was obtained on a test set of 269 subjects.

During this study, a collateral scoring product (e-Stroke Suite) became available from Brainomix. This software was evaluated by Grunwald et al. [42] recently. Ninety-eight subjects were used in their work. Their selection criteria were more restrictive than ours (1 mm slice thickness), and no information is provided on the occlusion location or number of excluded scans. Also, their consensus score may be biased towards the software performance, as the consensus score was determined after knowing the software score, which makes it hard to directly compare their result with ours. Reported accuracy of software compared to their reference standard is 90%, and they also demonstrate, similar to our work, that the errors of the software are within the interobserver variation.

All patients from the MR CLEAN registry that were included in the study had an acute large vessel occlusion, which was assessed on CTA. In all patients a large vessel occlusion was detected in one hemisphere only. Although we cannot rule out that small peripheral emboli were present in the contralateral hemisphere (which were not visible on CTA) in patients with a cardioembolic etiology, in general, the symptoms caused by a large vessel occlusion are more prominent than the symptoms caused by potential small peripheral emboli. Based on that we assume that the symptoms

<table>
<thead>
<tr>
<th>Property</th>
<th>Category</th>
<th>Number of subjects</th>
<th>Accuracy Baseline</th>
<th>RF</th>
<th>Clinician 1</th>
<th>Clinician 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slice thickness</td>
<td>[0.5-0.75mm)</td>
<td>66</td>
<td>0.79</td>
<td>0.80</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>[0.75-1mm)</td>
<td>94</td>
<td>0.79</td>
<td>0.87</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>[1-1.5mm)</td>
<td>109</td>
<td>0.71</td>
<td>0.74</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td>Acquisition phase</td>
<td>peak arterial phase</td>
<td>75</td>
<td>0.76</td>
<td>0.79</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>equilibrium phase</td>
<td>115</td>
<td>0.77</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>peak venous phase</td>
<td>79</td>
<td>0.73</td>
<td>0.81</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>Collateral score</td>
<td>score 0</td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>score 1</td>
<td>83</td>
<td>0.69</td>
<td>0.75</td>
<td>0.77</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>score 2</td>
<td>124</td>
<td>0.89</td>
<td>0.90</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>score 3</td>
<td>60</td>
<td>0.57</td>
<td>0.68</td>
<td>0.85</td>
<td>0.75</td>
</tr>
<tr>
<td>Occlusion position</td>
<td>ICA</td>
<td>64</td>
<td>0.81</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>154</td>
<td>0.75</td>
<td>0.81</td>
<td>0.86</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>M2 and above</td>
<td>45</td>
<td>0.67</td>
<td>0.71</td>
<td>0.69</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>no occlusion found</td>
<td>6</td>
<td>0.83</td>
<td>0.83</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

TABLE IV: Subgroup analysis.
indicate the hemisphere in which the large vessel occlusion is present. However, outside the MR CLEAN Registry, indeed cardioembolic stroke accounts for 14-30% of ischemic stroke population (Arboix and Alió [43]), and for those cases, the assumption of occluded side as prior might not be valid. Our method is not intended for those cases.

VI. CONCLUSIONS

We proposed a robust and automated collateral scoring method and evaluated it on a large set of images acquired in clinical routine as demonstrated in the MR CLEAN Registry. The proposed method achieves 80% accuracy on the 4 score prediction. In a dichotomized test, the proposed method achieves 90% accuracy. The result are comparable to two radiologists with 10 to 30 years of experience.

ACKNOWLEDGMENT

The authors would like to thank NVIDIA Corporation for the donation of a TITAN XP GPU used in these experiments. Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The authors would thank for the generous GPU resources provided by Dutch national e-infrastructure. The authors would also like to show our gratitude to the excellent brain vessel annotation work of Yvonne Veltman, Sven Luijten and Nikki Boodt.

REFERENCES


[33] I. G. Jansen, M. J. Mulder, and R.-J. B. Goldhoorn, “En-


