A Topology-Aware Deep Learning Approach for Automated Multi-Class Segmentation of the Circle of Willis in Modeling Applications

Author: E.A. (Emma Anneke) de Bruin

In partial fulfilment of the requirements of the

Master of Science

in **Biomedical Engineering**

Track: **Medical Physics**

at **Delft University of Technology**

To be defended publicly on **Monday July 14, 2025 at 14:00**

Thesis Advisors

Dr. S. Pirola

Dr. T. van Walsum

F.G. te Nijenhuis

Thesis Committee

Dr. S. Pirola (Chair)
Assistant Professor, Department of Biomechanical Engineering, TU Delft

Dr. ir. Theo van Walsum
Associate Professor, Department of Radiology & Nuclear Medicine, Erasmus Medical Center

F.G. te Nijenhuis
Doctoral Candidate, Department of Radiology & Nuclear Medicine, Erasmus Medical Center

Dr. X. Zhang
Assistant Professor, Department of Electrical Engineering, Mathematics and Computer Science, TU Delft

Dr. B. Fereidoonnezhad
Assistant Professor, Department of Biomechanical Engineering, TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl.

TU Delft

A Topology-Aware Deep Learning Approach for Automated Multi-Class Segmentation of the Circle of Willis in Modeling Applications

E. A. de Bruin[1]

[1] *Department of Biomechanical Engineering, Technical University of Delft, Delft, the Netherlands.*

***Abstract*—*Background*: Cerebrovascular diseases, which often involve a disruption in blood flow in the Circle of Willis (CoW) and its branching arteries, pose a major global health risk. Computational fluid dynamics (CFD) analyses present an opportunity to study their pathophysiology but require high-quality vessel segmentations. *Methods*: To generate pseudo-labeled training data, computed tomography angiography (CTA) images were preprocessed and inference was run using two pretrained models: an nnU-Net for multi-class CoW segmentation and a DTUNet for binary cerebral vessel segmentation. These outputs were combined using a region-growing approach; the resulting pseudo-labels were used to train an nnU-Net V2 with a topology-aware loss function. CFD analyses were performed on both a model-generated segmentation and a ground truth segmentation derived from a CTA scan which had been expert-labeled by a neuroradiologist. The resulting velocity, pressure and wall shear stress (WSS) profiles for both segmentations were compared across 10 cross-sections of the middle cerebral artery (MCA). *Results*: After filtering out inaccurate labels, 1,709 of 2,201 pseudo-labeled images were retained for training and testing. Common errors included over-segmentation of small vessels, under-segmentation of large vessels and poor separation of the anterior cerebral arteries when compared to expert-annotated ground truth segmentations. The proposed segmentation model was evaluated on the test set, which used pseudo-labels as a reference standard, and achieved a mean Dice score of 62%, clDice of 40%, IoU of 51%, a HD of 16.9 voxels and an ASD of 3.9 voxels. In terms of centerline-based metrics, the model achieved a mean overlap (OV) of 72% and an average ASCD of 4.26 voxels. In CFD simulations, the predicted segmentation yielded absolute errors of 48.96 ± 30.69 mm/s, 7.47 ± 6.07 Pa and 1.22 ± 0.81 Pa for blood flow velocity, pressure and WSS, respectively, compared to the expert-annotated reference ($p < 0.05$ for all). *Conclusions*: This study demonstrates that a deep learning model, trained using pseudo-labels, can successfully generate anatomically plausible multi-class segmentations of the CoW suitable for downstream CFD analysis. However, discrepancies in key hemodynamic metrics compared to expert-annotated data highlight the need for improved pseudo-label accuracy, especially in regions of complex vascular geometry.**

Keywords: Cerebrovascular disease, Circle of Willis, Computational fluid dynamics, Deep learning, Medical image segmentation.

1. Introduction

Cerebrovascular diseases currently pose a significant global health burden and are one of the main contributors to rising global morbidity and mortality rates [1, 2]. These diseases can lead to significant changes in cerebral blood flow to various regions of the brain [3], which can in turn cause severe neurological damage and, in some cases, death. Stroke, with an estimated prevalence of 1 in 4 adults affected over the course of their lifetime [4], is one of the most common and debilitating cerebrovascular diseases. This condition arises when blood supply to the brain is disrupted by a pathological ischemic or hemorrhagic process which affects one or multiple cerebral arteries. Other cerebrovascular diseases such as aneurysms [5], arteriovenous malformations [6, 7] and arterial stenosis can further contribute to the risk of ischemic or hemorrhagic events [8], which exemplifies the importance of being able to accurately assess cerebral vasculature as a part of the diagnostic and treatment processes.

The Circle of Willis (CoW) is an important cerebral arterial structure which acts as a collateral pathway that maintains adequate blood flow and equalizes blood pressure across both hemispheres of the brain [9, 10]. The physiological importance of this structure is underscored in cases where large cerebral arteries are affected by diseases such as stroke or arterial stenosis. In these situations, the CoW prevents large areas of the brain from suffering neurological damage as a result of oxygen deprivation by providing a collateral blood flow route which bypasses the affected area [11, 12]. Due to its crucial role in cerebral circulation and safeguarding brain tissue against neurological damage, understanding the structure and function of the CoW both in the presence and absence of pathology is imperative for diagnosing and treating cerebrovascular diseases. The ability to visualize and analyze the CoW in detail, therefore, is essential for clinicians and researchers alike, enabling accurate assessments of vascular health, prediction of stroke risk and the planning of interventions.

In current clinical practice, non-invasive imaging modalities such as computed tomography angiography (CTA) and magnetic resonance angiography (MRA) are the most common modalities used to visualize the CoW [14–16]. CTA is an imaging technique which makes use of X-ray radiation and an iodine-based contrast agent which is injected intravenously to create detailed images of blood vessels. This fast, high-

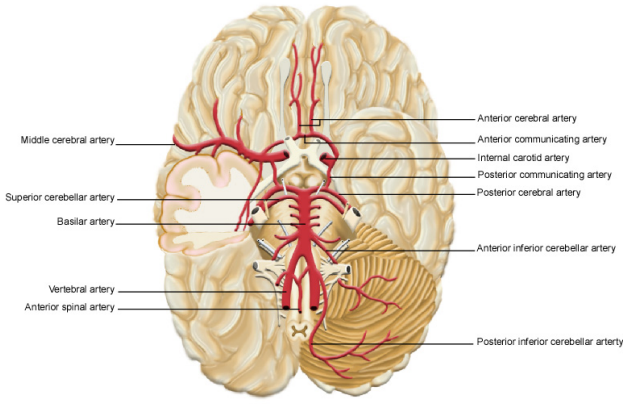

Figure 1. Anatomical schematic of the constituent arteries of the CoW. Adapted from the Royal College of Surgeons of Ireland [13].

resolution imaging modality is commonly used in acute settings for diagnosis and treatment planning for conditions such as stroke or ruptured aneurysms [17]. MRA is a non-invasive imaging technique which uses magnetic fields and radio wave pulses to produce detailed images of blood vessels and soft tissue structures. Unlike CTA, MRA does not require the use of ionizing radiation, however, this modality is associated with longer image acquisition times, higher costs and a lower spatial resolution compared to CTA [18]. For these reasons, the number of MRA scans conducted in clinical practice is currently decreasing, while the number of CTA scans being conducted is increasing at a much faster rate [14]. However, despite the growing reliance on CTA, vessel assessment still largely depends on manual measurements and visual inspection, which reduces efficiency and introduces variability into clinical workflows. To meet the increasing demand for accurate and timely analysis of cerebral vessel structure, there is a pressing need to increase the degree of automation in CTA image analysis for both researchers, who aim to investigate pathological processes in the CoW, and clinicians, who require patient-specific data.

Over the past few years, medical image analysis has experienced a significant revolution in the development of automated tools for medical image analysis. This revolution has been largely powered by the introduction and improvement of machine learning techniques, especially in the field of deep learning [19, 20]. Deep learning is particularly well-suited to medical image analysis as these models possess the capacity to efficiently learn complex patterns and features present in images, which lends itself well for disease detection and diagnosis. Advancements in deep learning architectures such as the U-Net, which was developed specifically for medical image segmentation [21], have improved the performance for segmentation and classification tasks. As a result, deep learning has become a cornerstone technology for both binary and multi-class segmentation tasks.

However, some issues still remain. First and foremost, for almost every type of segmentation there remains the challenge of creating or obtaining a large, accurately labeled dataset which can be used to train segmentation models. Without a large volume of accurate training data, it remains difficult to achieve strong model performance and generalizability. Furthermore, while current methods often achieve high accuracy in terms of traditional evaluation metrics, they fall short in addressing the specific needs of certain downstream applications, particularly those that require precise topological integrity [22].

One such critical application is in the field of computational fluid dynamics (CFD). CFD simulations, an example of which can be seen in Figure 2, can be used to estimate and model physiological parameters such as vessel wall stress, flow velocity and blood pressure distributions within specific vessels under the assumption of specific boundary conditions [23]. These simulations are crucial for understanding the biological and physiological effects caused by various cerebrovascular diseases and could also be used for developing personalized treatment plans for conditions such as aneurysms, stroke and vascular malformations. Accurate CFD analyses depend fundamentally on the availability of high-quality, 3D segmented models of vascular structures, as any errors or inconsistencies in segmentation can propagate through simulations and significantly affect the reliability of predicted hemodynamic parameters [24]. This means that for this application, the segmentation process of vascular networks must aim to preserve their exact anatomical shape and topology. Additionally, the segmentations need to be continuous, avoiding any breaks that would split vessels into disconnected parts. Lastly, the segmentation should capture the full extent of the vascular structures, including small and branching vessels, in order to ensure the accuracy of flow dynamics and pressure distribution predictions.

It has been shown that CFD simulations can be used to gain an insight into general cerebral hemodynamics in the cases of intracranial arterial stenosis [25, 26], and aneurysms [27]. However, producing models that are tailored to specific patients remains challenging. Currently, segmentation is also most often done using semi-automated approaches involving region-growing, active contours, centerline-based methods or simple intensity thresholding [28]. Some have already tried to use deep learning to create an automatic segmentation tool for CFD analysis for other blood vessels which showed promising results [29–31]. It should be noted that these models focus mainly on larger arteries, which calls for the design of an equally accurate model tailored specifically to the CoW.

Furthermore, despite significant advances in medical image segmentation, most binary and multi-class models have not been validated in application-specific contexts like CFD, where preserving vascular topology is essential. Although several studies report high Dice scores for CoW segmentation

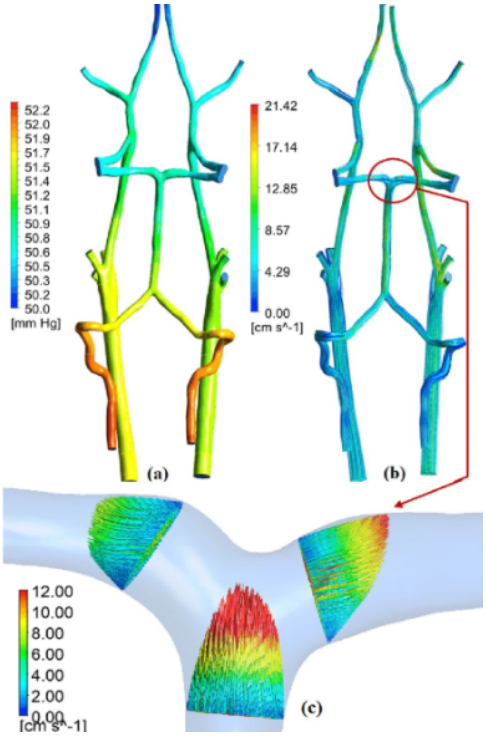

Figure 2. Example of CFD analysis of hemodynamic parameters in the CoW. Simulations can be performed to estimate various parameters, such as blood pressure in mm Hg (a) and velocity in cm/s (b), to analyze their magnitude and distribution throughout different vessel regions. In some cases, specific sections of the CoW may need to be examined individually, and flow dynamics can be assessed over time for parameters such as velocity (c). Adapted from Yankova et al. [32].

[22, 33–35], only the models from Yang et al.'s TopCoW 2023 challenge explicitly focus on topological accuracy [22]. Additionally, systematic evaluation of segmentation outputs for their suitability in CFD simulations, particularly at the CoW or individual vessel segment level, is rare. This gap is exacerbated by the limited availability of large, labeled CTA or MRA datasets needed for effective model training. As a result, the potential advantages of applying deep learning to multi-class CoW segmentation for CFD remain largely unexplored. Therefore, the main purpose of this study is to investigate whether deep learning models trained using pseudo-labels can produce anatomically accurate, multi-class segmentations of the CoW that are suitable for downstream CFD analysis. The main contributions of the study are therefore threefold:

- This article presents a novel preprocessing pipeline which generates a large volume of pseudo-labeled images that can be used for training segmentation models, thereby aiming to reduce the need for extensive manual annotations.
- A topology-aware deep learning framework is trained and evaluated, with an emphasis on topological and anatomical accuracy and consistency, to provide more precise and reliable results for downstream analyses at both the whole-network and individual vessel levels.
- The segmentation model is validated for CFD applications by comparing blood flow velocity, pressure and wall shear stress (WSS) in the middle cerebral artery (MCA) against expert-annotated ground truth, demonstrating its potential for clinical and research use.

By combining pseudo-labeling with topology-aware learning, this work aims to streamline CFD workflows, reduce manual annotation and vessel analysis burdens and enhance reproducibility in vascular analysis.

2. Methods and Materials

A. Dataset

The CTA dataset used in this study was compiled retrospectively using four different sources, including both public and private sources, as can be seen in Table 1. The dataset includes 130 CTA images which were made publicly available by the organizers of the Topology-Aware Anatomical Segmentation of the Circle of Willis for CTA and MRA grand challenge in 2024, also known as the TopCoW grand challenge [22]. The subjects in this dataset are patients which were admitted to the Stroke Center of the University Hospital Zurich between 2018 and 2019 for a stroke-related neurological disorder. These CTA images had a voxel size of 0.45 mm in the x- and y-dimensions and 0.7 mm in the z-dimension.

A further 143 CTA images were sourced from the publicly available Clinical, Morphological, Hemodynamic Data for Aneurysms (CMHA) dataset [36]. This dataset includes scans of patients who received a CTA scan between 2012 and 2018 at the Second Affiliated Hospital of Anhui Medical University in China. The dataset contains 44 CTA images of healthy adults and 99 CTA images of patients diagnosed with an intracranial aneurysm. These images have varying voxel sizes which are all comparable to those of the TopCoW CTA images.

A third publicly available dataset, the Large Intracranial Aneurysm Segmentation Dataset (LIASD), was used as an additional source of imaging data [37]. This dataset includes 1,476 CTA images of patients diagnosed with either a ruptured or non-ruptured intracranial aneurysm, gathered from eight separate institutions across China. These images also have varying voxel sizes similar to those of the TopCoW CTA images.

Lastly, 458 CTA images were obtained from the Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands (MR CLEAN) NO-IV trial [38], which is stored and curated by the Collaboration for New Treatments for Acute Stroke (CONTRAST) consortium. This is a private dataset which includes CTA images from patients that underwent endovascular stroke treatment across multiple hospitals in the Netherlands, Belgium and France between March 2014 and December 2018. These images have varying voxel sizes, with mean voxel dimensions of 0.45x0.45x1.35 mm in the x-, y- and z-dimensions respectively.

Table 1. An overview of the different sources and patient populations included in the dataset.

Dataset	Source Type	CTA Images	Patient Population	Acquisition Details
TopCoW	Public	130	Stroke patients	University Hospital Zurich, 2018–2019.
CMHA	Public	143	99 aneurysm patients 44 healthy adults	Anhui Medical University, China, 2012–2018.
LIASD	Public	1476	Aneurysm patients	8 hospitals across China, unknown timeframe.
MR CLEAN NO-IV	Private	458	Stroke patients	20 hospitals across the Netherlands, Belgium and France, 2014–2018.

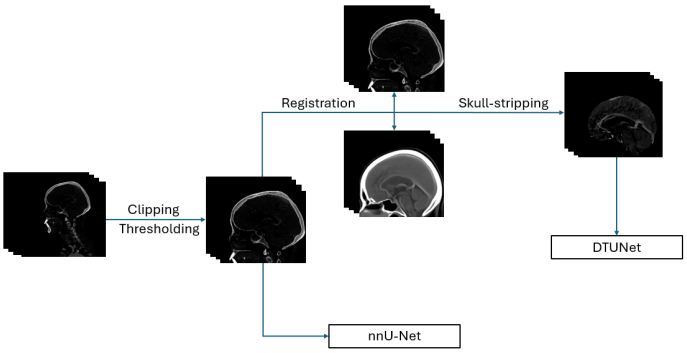

Figure 3. A schematic representation of the preprocessing steps required to produce input images suitable for both the nnU-Net and the DTUNet.

B. Preprocessing

To prepare the CTA images for the semi-automatic generation of pseudo-labels and subsequent model training, the images underwent a standardized preprocessing pipeline to ensure anonymization and compatibility with pretrained segmentation models. All CTA images in the LIASD and MR CLEAN datasets were first defaced using TotalSegmentator in order to anonymize the images [39]. CTA images in the TopCoW and and CMHA datasets had already been defaced and were therefore not subjected to this preprocessing step. The anonymized dataset was further preprocessed using two similar but separate approaches to generate suitable input data for two pretrained segmentation models, a DTUNet and an nnU-Net, as can be seen in Figure 3.

For the nnU-Net, input images were clipped and their intensity values thresholded in order to address memory constraints and ensure consistency across all CTA images. The images were clipped starting from the cranial region of the skull, extending 20 cm in the caudal direction. Subsequently, the intensity values were thresholded to fall within the range of -1024 to 1600.

For the DTUNet, clipping and thresholding were performed as described above. Following these steps, atlas-based registration was conducted using the ANTs toolbox in order to facilitate skull-stripping of the images [40]. The input images were registered to a template CTA image in atlas-space using an affine transformation. The decision was made to avoid using deformable transformations since these can stretch, warp, or bend structures. This could possibly introduce artificial connections, alter the shape of vessels or remove small vessels, which could lead to topological errors in the resulting images. Following the atlas-based registration, skull-stripping of the CTA images was conducted. However, this process resulted in the internal carotid arteries (ICAs) and the basilar artery (BA) being truncated prematurely, limiting the ability to follow these vessels along their full course. For this reason, prior to skull-stripping a partial dilation of the brain mask was performed in this region for all images in order to preserve the ICAs and BA, at the expense of retaining a small amount of skull tissue in this region.

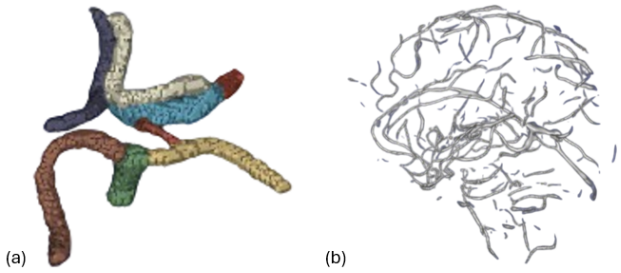

Figure 4. An example of the segmentation output produced by the pretrained models. (a) The nnU-Net produces a multi-class segmentation of the CoW which segments up to 12 distinct arteries. (b) The DTUNet produces a binary segmentation of all cerebral vessels, including both arteries and veins.

C. Pretrained segmentation models

Once the CTA images had been preprocessed, inference was run using two separate deep learning models. The first model was developed by the Charité Lab for Artificial Intelligence in Medicine (CLAIM) research group [22, 41, 42]. This makes use of the Residual Encoder Medium (ResEnc M) architecture, which is a variation of the nnU-Net V2 [43]. The nnU-Net is a special variant of the U-Net that is able to automatically determine the best configuration, including preprocessing steps, network architecture and postprocessing steps in order to optimize the model for a specific task. This model uses skeleton recall and cross-entropy in its loss function to produce a multi-class segmentation of the CoW using raw CTA images as input. The network classifies each voxel into one of 14 different classes, with 0 being the background and 1-13 being different arterial segments associated with the CoW. The segments which the model classifies can be found in Table 2. An example of the output produced by this model can be seen in Figure 4a.

The second deep learning network which was utilized was a Dual-branch Topology-aware U-Net (DTUNet) model [44]. This model is based on the 3D U-Net and contains two main branches, as can be seen in Figure A1 (Appendix A). One

Table 2. The labels and classes segmented by the pretrained nnU-Net.

Label	Class
0	Background
1	Basilar artery (BA)
2	Right posterior cerebral artery (R-PCA)
3	Left posterior cerebral artery (L-PCA)
4	Right internal carotid artery (R-ICA)
5	Right middle cerebral artery (R-MCA)
6	Left internal carotid artery (L-ICA)
7	Left middle cerebral artery (L-MCA)
8	Right posterior communicating artery (R-Pcom)
9	Left posterior communicating artery (L-Pcom)
10	Anterior communicating artery (Acom)
11	Right anterior cerebral artery (R-ACA)
12	Left anterior cerebral artery (L-ACA)
13	3rd A2 segment (3rd-A2)

branch is responsible for segmentation of the lumen of a blood vessel, while the other branch predicts the vessel centerline. The model also contains a fusion path which integrates the features identified in the two main branches and passes these through spatial and channel attention modules. This model makes use of a topology-aware loss function which employs the Dice loss for lumen segmentation and a combination of both the Dice loss and the centerline-Dice loss for the vessel centerline prediction. This model produces a binary segmentation of cerebral blood vessels, including both cerebral arteries and veins. An example of the output produced by this model can be seen in Figure 4b.

D. Pseudo-labels

1) Pseudo-label generation: Pseudo-labels were created by first conducting inference using the pretrained nnU-Net and DTUNet on the preprocessed CTA images. Two separate segmentation outputs were therefore obtained for each input image: the nnU-Net model produced a multi-class segmentation of the CoW, while the DTUNet produced a binary segmentation of the intracranial arteries and veins present in the image. To isolate and generate pseudo-labels for the individual CoW segments along with their branching arteries, a series of operations was performed, as outlined in Figure 5.

First, the inverse of the transformation matrix used to register the DTUNet input images to the atlas space, $\mathbf{A}^{-1}$, was applied to the DTUNet output using linear interpolation. This ensured that the outputs of both the nnU-Net and DTUNet models were spatially aligned in the original image space.

Following this, a binary mask of each of the segments present in the output of the multi-class nnU-Net model was created. Subsequently, multiple binary masks were generated from the full multi-class nnU-Net output, with each mask excluding one specific class. Each of these masks was subtracted from a new, binary mask of the DTUNet output, creating multiple binary images of the DTUNet output which were each missing the whole CoW except for a single, unique segment. This volume was then used for region-growing, where the remaining segment of the CoW was used as the seed. Region-growing was performed only for the major arteries present in the CoW: the basilar artery, the right and left posterior cerebral arteries, the right and left middle cerebral arteries and the right and left anterior cerebral arteries.

The resulting images after region-growing for each individual segment were binary images with a background value of 0 and a foreground value corresponding to the segment number as seen in Table 2. Segment 13 was excluded due to its rare occurrence and limited clinical relevance. Finally, the binary output images from the region-growing process were combined by taking the maximum voxel value between the combined image and each individual segment image. The remaining segments that were not used for region-growing were then added into the combined image to complete the pseudo-labeled image.

2) Analysis of the accuracy of the pseudo-labels: To assess the accuracy of the assigned labels, the total number of segments and volume of individual segments were analyzed for each pseudo-labeled image. Following this, the mean number of segments per image and the mean volume per segment class were calculated, along with their corresponding standard deviations (STD). Pseudo-labeled images containing outliers, either in terms of segment count or segment volume, were excluded from the training and test sets under the assumption that these cases likely reflected segmentation or classification errors. Outliers were defined as images containing four or fewer segments, or those in which one or more segment volumes deviated more than 1.96 times the SD from the class-specific mean volume.

To quantitatively estimate the accuracy of the pseudo-labels, the Dice score was computed between the binarized pseudo-labeled output and the corresponding original CTA images. For this, aggressive skull-stripping was performed on both the pseudo-labeled and the original images in order to ensure that no skull tissue remained in any images. This was achieved by eroding the original brain mask and applying this to all images. Following this, the images were binarized by applying a threshold. For the pseudo-labeled images, a fixed threshold of 1 was applied. For the CTA images, a percentile-based thresholding approach was applied in which only values above the 99.8^{th} percentile were retained. Dice scores were calculated for each image to evaluate how well the pseudo-label generation pipeline segmented and labeled the arteries present in the original CTA images. A qualitative assessment was also performed to compare the pseudo-labels to the thresholded images.

To provide stronger validation, the first author manually created a voxel-wise multi-class segmentation of the CoW and its branching arteries for a single, randomly selected CTA image. This segmentation was reviewed and corrected by an experienced neuroradiologist. The corresponding pseudo-labeled image was then quantitatively and qualitatively compared to this ground truth segmentation.

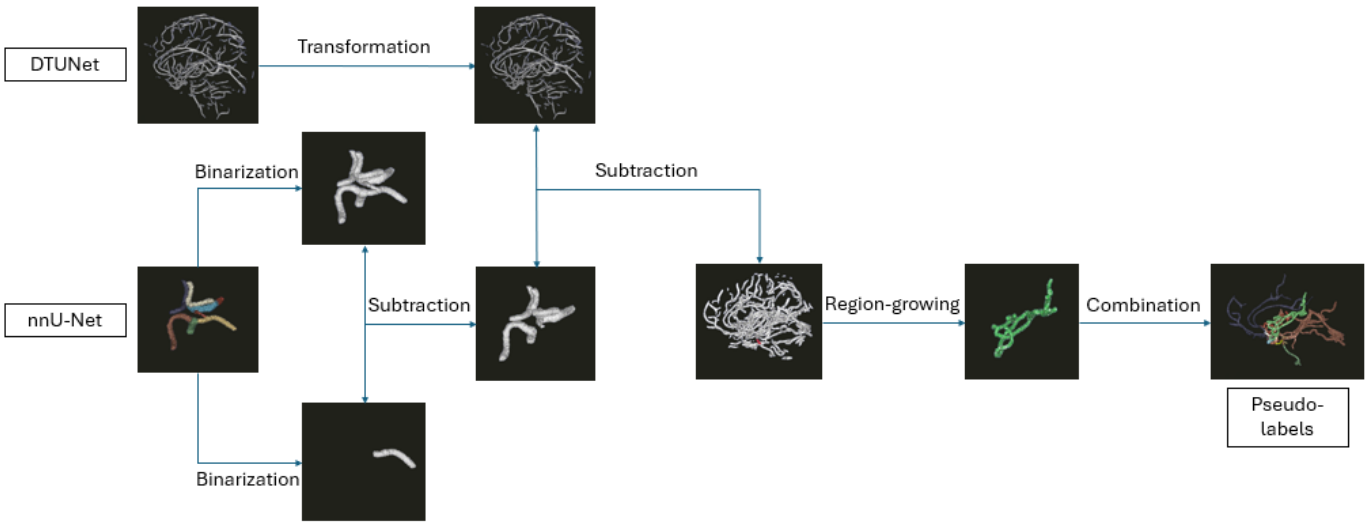

Figure 5. A schematic representation of the creation of the pseudo-labeled images. This outlines the step-by-step process for generating pseudo-labeled images using operations such as transformations, binarization, subtraction and region-growing.

3) Dataset split: The pseudo-labeled images were split into training and test sets using a stratified split. Stratification was done by ensuring that pseudo-labeled images with nine segments or less, and pseudo-labeled images with 10 segments or more, were evenly distributed between the training and test sets. This cut-off point was chosen as prior research suggests that the expected mean number of CoW segments present in the pseudo-labeled images should be between 10-11 [22, 45], which places images with nine segments or less at risk of being poor quality due to missing segments. This cut-off was also supported by the distribution of segment counts across all images. Since the CoW is notorious for its anatomical variation, images with segment counts below nine could not be discarded but were divided evenly in case errors were present. In this way, it was ensured that an equal distribution of high and poor quality images was achieved across both sets. Since there is no standardized dataset split for this dataset, a standard split of 80/20 for the training and test sets was applied. An analysis of the demographic statistics of both sets was conducted to ensure that there were no significant differences between the training and test sets. The organizers of the TopCoW grand challenge did not provide demographic statistics for their data which meant that these images were left out of this analysis. Five-fold cross-validation was applied.

E. Segmentation model

The decision was made to use an nnU-Net V2 due to the high performance of the nnU-Net in medical image segmentation applications. In order to ensure that the network produced a segmentation output that was suitable for CFD simulations, a custom loss function was created. The first component of the loss function is the multi-class Dice loss, which was computed using the following formula:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{v\in V}\left(S^{(c)}_{\text{pred},v}\cap S^{(c)}_{\text{gt},v}\right)}{\sum_{v\in V}S^{(c)}_{\text{pred},v}+\sum_{v\in V}S^{(c)}_{\text{gt},v}} \tag{1}$$

where C is the total number of classes, V is the total number of voxels present in the image, S_{pred} is the segmentation prediction and S_{gt} is the ground truth segmentation, which in this case is the pseudo-labeled image. This component is a measure of similarity that prioritizes overlap between predicted and ground truth masks for each class.

The second component of the loss function is cross-entropy (CE), which is calculated as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{v\in V}\sum_{c=1}^{C} S^{(c)}_{\text{gt},v}\log S^{(c)}_{\text{pred},v} \tag{2}$$

This component ensures that the model prioritizes voxel-wise prediction accuracy and can combat class-imbalance present in the images.

The final component of the loss function is skeleton recall. This is a recent development designed to promote the preservation of connectivity, which is a crucial factor for accurately segmenting thin, tubular structures such as blood vessels [41]. It offers a computationally efficient alternative to centerline-Dice (clDice) and is calculated as follows:

$$\mathcal{L}_{\text{SRec}} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{|\text{SK}_{\text{pred},c}\cap \text{SK}_{\text{gt},c}|}{|\text{SK}_{\text{gt},c}|} \tag{3}$$

Where SK_{pred} is the skeleton of the predicted segmentation and SK_{gt} is the skeleton of the ground truth segmentation. The final and complete loss function can therefore be given as:

$$\mathcal{L}_{\text{total}} = \omega_{\text{Dice}} \cdot \mathcal{L}_{\text{Dice}} + \omega_{\text{CE}} \cdot \mathcal{L}_{\text{CE}} + \omega_{\text{SRec}} \cdot \mathcal{L}_{\text{SRec}} \tag{4}$$

Where ω_{Dice}, ω_{CE} and ω_{SRec} are the weights of the respective loss components. Each loss component was given an equal weight of one. A soft skeletonization algorithm was used to extract vessel centerlines from the segmented volumes. These were then used to compute a memory-efficient soft skeleton recall loss, which scales well to large 3D image inputs. The initial learning rate was set to 1×10^{-4} and was determined in later epochs by a polynomial learning rate scheduler. Optimization was performed using AdamW. A foreground oversampling percentage of 75% was chosen in order to ensure that the model was able to successfully learn all classes, including the smaller and underrepresented classes. The model was trained until convergence.

F. CFD experiment

In order to be able to validate the output of the model for CFD, the ground truth segmentation created by the neuroradiologist was used to provide a comparison independent of the pseudo-labeling step which could be used for further analysis. Inference of the proposed segmentation model was run on the corresponding original CTA image to produce a second segmentation for comparison. To minimize the risk of setup or execution errors, postprocessing of both segmentations and the simulation itself were carried out by an experienced researcher specialized in CFD. First, the MCA and its three branching arteries were isolated. The surfaces of the segmentations were smoothed, small defects were removed and any holes present were filled in using Autodesk Meshmixer [46]. To generate a finite element model, both smoothed segmentations were remeshed, extremities were cut and extensions of 5 mm were added. Five boundary layers of a maximum thickness of 0.2 mm were added. The rest of the fluid domain was meshed with 0.3 mm elements. All surfaces were identified for boundary condition definition. These steps were performed using MatLab [47]. Following this, a CFD simulation was run using three specific boundary conditions: a no-slip boundary condition on the external surface, a fluid normal velocity with parabolic profile and a fixed fluid pressure of 0.00144 MPa at the outlets. A dynamic viscosity of 3.5×10^{-9} MPa was applied. The simulation was run for 3000 time steps with a step size of 0.001 s using MatLab and FEBio solver [48]. The blood flow velocity and blood pressure data were extracted directly from the simulation output. The WSS was calculated manually using a standard traction decomposition formula, following the same method implemented in an in-house postprocessing tool developed by researchers at the same affiliated institution. Specifically, the WSS magnitude was computed as:

$$\|\boldsymbol{\tau}\| = \sqrt{\|\mathbf{T_n}\|^2 - \sigma_n^2} \tag{5}$$

Where $\boldsymbol{\tau}$ is the WSS vector, $\mathbf{T}_n$ is the traction tensor and σ_n is the normal stress. A derivation of this formula can be found in Appendix C. A rough, rigid registration of both meshes was performed using the iterative closest point algorithm to improve spatial alignment of the two models prior to evaluation of the results. This was necessary as the models can become misaligned in the postprocessing steps when models are transformed and remeshed in order to identify the in- and outlets.

G. Metrics

The complete pipeline was evaluated at several distinct locations. First of all, the accuracy of the pseudo-labels was estimated using the Dice score calculated between the binarized pseudo-labeled images and the skull-stripped original CTA images using:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{6}$$

where TP is the amount of true positives, FP is the number of false positives and FN is the number of false negatives.

After evaluating the accuracy of the pseudo-labels, the segmentation accuracy of the proposed model was evaluated. For this, the Dice score was calculated for each segment class. The clDice was also calculated for each class by performing skeletonization of the lumen segmentation and applying the Dice formula to the centerline points. Besides this, the Intersection over Union (IoU) was calculated for each class using:

$$IoU = \frac{TP}{TP + FP + FN} \tag{7}$$

The Hausdorff Distance (HD) was calculated for each class using:

$$\text{HD}(\text{pred}, \text{gt}) = \max \left\{ \sup_{p \in C_{\text{pred}}} \inf_{g \in C_{\text{gt}}} \|p - g\|, \quad \sup_{g \in C_{\text{gt}}} \inf_{p \in C_{\text{pred}}} \|g - p\| \right\} \tag{8}$$

The Average Surface Distance (ASD) was also calculated for each segment using:

$$\text{ASD} = \frac{1}{2} \left(\frac{1}{|S_{\text{pred}}|} \sum_{p \in S_{\text{pred}}} \min_{g \in S_{\text{gt}}} \|p - g\| + \frac{1}{|S_{\text{gt}}|} \sum_{g \in S_{\text{gt}}} \min_{p \in S_{\text{pred}}} \|g - p\| \right) \tag{9}$$

The precision and recall were calculated for each class using:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

The overlap (OV) was calculated using:

$$\mathrm{OV} = \frac{\mathrm{TP}_{\mathrm{gt}} + \mathrm{TP}_{\mathrm{pred}}}{\mathrm{TP}_{\mathrm{gt}} + \mathrm{TP}_{\mathrm{pred}} + \mathrm{FN}_{\mathrm{gt}} + \mathrm{FP}_{\mathrm{pred}}} \tag{12}$$

The Average Symmetric Centerline Distance (ASCD) was calculated using:

$$\mathrm{ASCD} = \frac{1}{2}\left(\frac{1}{|C_{\mathrm{pred}}|}\sum_{p\in C_{\mathrm{pred}}}\min_{g\in C_{\mathrm{gt}}}\|p-g\| \; + \; \frac{1}{|C_{\mathrm{gt}}|}\sum_{g\in C_{\mathrm{gt}}}\min_{p\in C_{\mathrm{pred}}}\|g-p\|\right) \tag{13}$$

In order to evaluate the proposed model output following the CFD experiments, 10 cross-sections perpendicular to the centerline of the MCA were selected, an example of which can be seen in Figure 6. Corresponding cross-sections from both the predicted and ground truth models were compared. The mean absolute error was then calculated for blood velocity, blood pressure and WSS for each cross-section. The mean absolute error and its corresponding standard deviation across all 10 cross-sections was then calculated. Paired T-tests were conducted to ascertain whether there were statistical differences present between the ground truth and the predicted values. Bland-Altman plots were created in order to analyze the error distribution and determine whether there was a systematic bias present. In addition, Pearson's correlation coefficient (r) and the coefficient of determination (R^2) were calculated to assess the strength and linearity of the relationship between predicted and ground truth values.

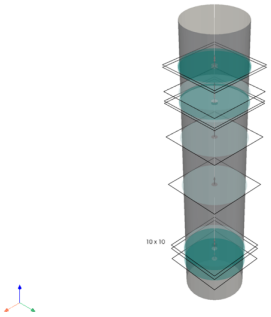

Figure 6. Example workflow for cross-sectional analysis of the MCA. A number of centerline points were randomly selected, at which point a cross-section was made perpendicular to the centerline at that location. Velocity, pressure and WSS were then compared across these slices for corresponding points from the ground truth and predicted models.

H. Implementation details

Training, testing and inference of models were conducted using PyTorch v2.4.1 with CUDA v12.4. All processes were executed on an NVIDIA Tesla A100 80GB GPU offered by the Delft High Performance Computing Centre (DHPC) [49].

3. Results

A. Pseudo-label generation

A total of 2,207 CTA images were processed for pseudo-label generation. The complete pseudo-labeling pipeline required approximately five minutes per image, which amounted to a total processing time of around 184 hours for the full dataset. The pipeline failed in six instances: three during preprocessing due to corrupted image data, and three during nnU-Net inference. This left 2,201 pseudo-labeled images for further analysis.

To filter out low-quality pseudo-labels, images were evaluated for outliers in terms of segment count and segment volume. The majority of images contained nine or more segments, as shown in Figure B1 in Appendix B. A total of 165 images were identified as outliers for having four segments or less. Segment volume analysis showed expected mean values across all classes, with the exception of segments 11 and 12, corresponding to the right and left anterior cerebral artery (R-ACA and L-ACA), respectively. In overlapping cases, the R-ACA was consistently under-segmented, while the L-ACA was consistently over-segmented. As this effect was systematic, it was not considered grounds for exclusion. In total, 351 scans contained at least one segment with an outlier volume. Lastly, 24 images were marked as outliers due to both segment count and segment volume. After the filtering process, a total of 1,709 pseudo-labeled images were deemed to be of acceptable quality for use in training, validation and testing of the proposed model.

B. Pseudo-label evaluation

To assess the reliability of the filtered pseudo-labels, both quantitative and qualitative evaluations were conducted. The mean Dice score between the filtered, binarized, pseudo-labeled images and the original, thresholded CTA images was found to be 48%. To further evaluate pseudo-label quality, a visual assessment was performed on 20 images selected to represent a range of Dice scores when compared to their thresholded counterparts. All selected images passed the initial filtering criteria for segment count and volume. The subset intentionally included 14 failure cases with Dice scores below 10% and six successful cases with scores above 50%. In most failure cases, low Dice scores were attributed to poor original CTA scan quality (n=5), registration or transformation errors (n=5), incomplete skull-stripping (n=2), or failure in the region-growing process (n=2). It was noted that segmentations with nine classes or less often contained more labeling inaccuracies, which resulted in lower Dice scores. In all successful cases, the pseudo-labeled images effectively captured the vast majority of the vessels segmented in the original image. However, it should be noted that the pseudo-labeled images tended to slightly under-segment larger vessels and over-segment smaller ones compared to the thresholded images, which serve only as a rough approximation and may not reflect anatomically accurate segmentation. An example of one of the successful cases can be seen in Figure 7.

To provide further insight into pseudo-label quality, an additional comparison was performed for a single image between the expert-annotated ground truth segmentation and the corresponding pseudo-labeled image, as shown in Figures 8 and 9. The class-averaged Dice score between the two binarized

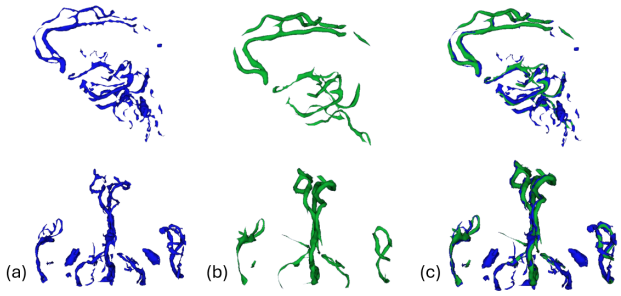

Figure 7. A comparison of the eroded and thresholded CTA image (a) and the eroded pseudo-labeled output (b). In (c), a substantial overlap is observed, however, the pseudo-labeled output exhibits a small degree of over-segmentation, particularly in smaller vessels.

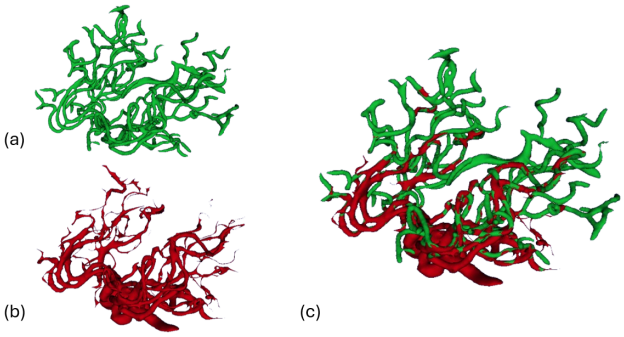

Figure 8. A comparison of the binarized pseudo-labeled image (a) and the expert-labeled ground truth image (b). In (c), it is evident that there is over-segmentation of the smaller vessels and under-segmentation of the larger vessels in the pseudo-labeled image compared to the expert-labeled image.

images seen in Figure 8 is approximately 40%. Visual analysis of the multi-class segmentations in Figure 9 yielded three main observations. First, as highlighted in red, the ICAs appear prematurely truncated in the pseudo-labeled image, despite the use of a partially dilated mask during skull-stripping intended to prevent this. Second, as indicated by the yellow circles, the pseudo-labeled image fails to clearly distinguish the ACAs. In nearly all pseudo-labeled images, there is overlap between the region-grown volumes of the R-ACA and L-ACA (segments 11 and 12, respectively). The pseudo-labeling pipeline retains the maximum segment value in overlapping voxels, resulting in consistent under-segmentation of segment 11 and over-segmentation of segment 12. This overlap also posed challenges during expert annotation, where the close proximity of the ACAs often made voxel-level classification difficult. Close proximity of different vascular structures also led, in this pseudo-labeled image as well as in a large number of other pseudo-labeled images, to venous structures being segmented and classified as a neighboring artery. Finally, as marked in blue, there appears to be over-segmentation of smaller arteries in the pseudo-labeled image. Conversely, larger arteries tended to be slightly under-segmented, though to a lesser extent. Close-up illustrations of these findings are provided in Figures B2–B4 in Appendix B.

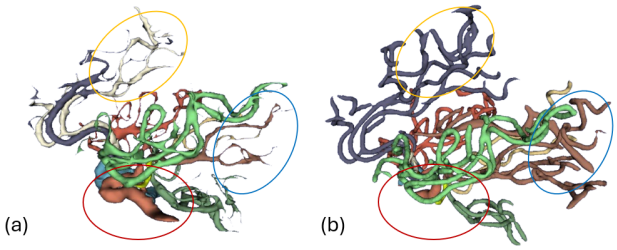

Figure 9. A comparison of the multi-class ground truth segmentation (a) and corresponding pseudo-labeled image (b). The three main differences between the pseudo-labeled image and the ground truth scan are failure to segment the entire length of the R-ICA and L-ICA (red), failure to separate the R-ACA and L-ACA (yellow) and over-segmentation of smaller vessels (blue).

C. Internal test set performance

There were no significant differences in age, gender or the mean number of segments present in the images between the training and test sets, as can be seen in Table 3. The model was initially trained for 1000 epochs while continuously monitoring the learning curve to stop model training prior to overfitting. After 1000 epochs the model had not yet clearly converged, as can be seen in Figure B6 in Appendix B. Training was therefore resumed for a further 1000 epochs. After 2000 epochs, no clear pattern of overfitting was visible, but the model was considered to have converged as performance seemed to have reached a plateau and did not appear to improve significantly as training progressed. Total training time was approximately 28 hours.

Table 3. The demographic characteristics of the training and test sets. The mean age and gender ratios do not take the 130 TopCoW CTA scans into consideration, since no demographic characteristics were made available for this dataset.

Demographic	Training	Test	Difference (p-value)
Subjects [N]	1367	342	-
Age* (STD) [years]	59.94 (13.52)	59.72 (14.07)	0.53
Gender* [% male]	43.64	48.15	0.17
Segments [N]	9.39	9.45	0.41

* Based on 1,579 subjects with complete demographic data.

The cross-validated model was used to perform inference on the internal test set and its predictions were subsequently evaluated, the results of which can be found in Table 4. The proposed model achieved a mean Dice score of 62% across all classes, with scores ranging from 51% for the left posterior cerebral artery to 78% for the right internal carotid artery. The mean clDice score was 40% and the mean IoU was approximately 51%. The model also yielded a mean HD of 17 voxels and a mean ASD of 4 voxels. Mean precision and recall were 61% and 71%, respectively. Centerline metrics presented a mean OV of 72%, a mean ASCD of 4.26 voxels and FP and

FN rates of 31% and 21%, respectively, as can be seen in Table 5.

Table 4. Performance segmentation metrics after evaluation of the model on the internal test set.

Class	DSC (STD) [%]	clDice (STD) [%]	IoU (STD) [%]	HD (STD) [voxels]	ASD (STD) [voxels]	Precision (STD) [%]	Recall (STD) [%]
BA	65.08 (28.29)	41.69 (20.01)	53.48 (25.35)	20.23 (19.07)	5.41 (12.64)	64.26 (27.94)	72.66 (29.33)
R-PCA	53.83 (31.53)	36.46 (23.22)	42.68 (27.39)	24.09 (25.21)	6.83 (13.94)	54.96 (30.34)	66.52 (34.74)
L-PCA	50.85 (31.19)	34.31 (22.40)	39.56 (26.41)	28.85 (27.61)	6.08 (12.14)	55.80 (29.21)	59.38 (36.69)
R-ICA	77.72 (28.03)	51.86 (24.25)	69.63 (26.82)	6.88 (20.07)	2.75 (11.54)	75.62 (28.33)	82.73 (27.66)
R-MCA	66.50 (25.64)	43.19 (17.15)	54.02 (22.13)	24.26 (22.39)	5.56 (13.09)	61.26 (24.57)	80.02 (22.40)
L-ICA	76.61 (27.97)	50.04 (24.84)	68.09 (26.89)	9.95 (26.39)	2.65 (8.33)	74.42 (28.30)	81.08 (28.44)
L-MCA	65.62 (26.25)	42.98 (17.78)	53.17 (22.39)	25.61 (21.52)	5.62 (12.99)	61.54 (25.67)	78.79 (22.48)
R-Pcom	59.56 (28.64)	36.94 (24.86)	47.48 (25.34)	3.28 (4.05)	1.35 (2.98)	57.85 (27.95)	70.33 (34.16)
L-Pcom	55.13 (30.06)	36.78 (26.17)	43.27 (25.61)	2.92 (3.36)	1.37 (2.76)	52.57 (29.71)	64.69 (35.56)
Acom	53.80 (27.79)	24.06 (34.81)	41.25 (23.91)	1.38 (0.57)	0.65 (0.83)	52.80 (30.84)	66.89 (35.96)
R-ACA	59.00 (29.42)	39.10 (21.61)	47.19 (25.90)	15.56 (21.57)	2.76 (6.93)	59.11 (26.40)	66.17 (34.30)
L-ACA	63.13 (24.04)	38.71 (15.48)	49.74 (20.84)	39.17 (23.83)	5.64 (13.48)	62.85 (23.90)	68.20 (26.02)
Mean	**62.24**	**39.68**	**50.80**	**16.85**	**3.89**	**61.09**	**71.46**

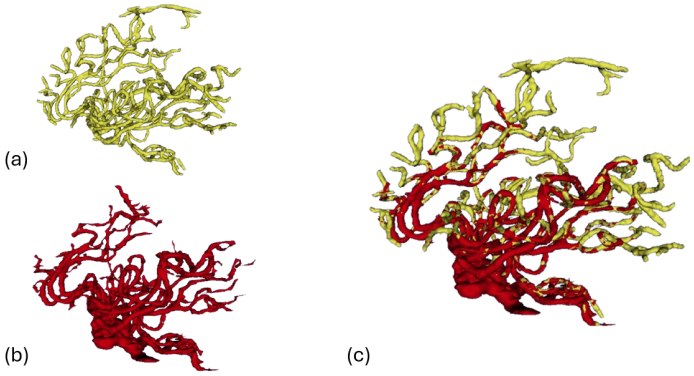

Figure 10. A comparison of the binarized predicted segmentation (a) and the expert-labeled ground truth segmentation (b). In (c), the predicted segmentation shows a similar pattern of over- and under-segmentation relative to the manual segmentation as was observed in the pseudo-labeled images.

Table 5. Performance centerline and structural metrics after evaluation of the model on the internal test set.

Class	OV Mean (STD) [%]	False Negatives [%]	False Positives [%]	ASCD Mean [voxels]
BA	65.31 (37.02)	28.56 (51.96)	23.30 (33.82)	5.51 (10.98)
R-PCA	60.49 (37.58)	38.78 (84.68)	29.05 (59.53)	7.20 (12.17)
L-PCA	58.06 (36.96)	50.35 (95.11)	19.92 (40.38)	7.96 (11.45)
R-ICA	81.80 (31.23)	17.86 (56.58)	12.16 (24.50)	3.09 (11.16)
R-MCA	69.44 (36.90)	11.18 (27.98)	30.62 (41.01)	4.62 (11.49)
L-ICA	80.31 (30.74)	36.69 (151.29)	12.87 (24.47)	3.73 (10.98)
L-MCA	71.39 (34.93)	13.34 (19.82)	26.97 (37.70)	4.34 (10.08)
R-Pcom	84.86 (30.20)	24.59 (59.19)	15.65 (46.39)	1.34 (2.75)
L-Pcom	85.58 (30.01)	28.54 (67.37)	11.07 (40.67)	1.22 (1.88)
Acom	65.70 (46.23)	37.15 (82.38)	38.19 (83.65)	0.83 (1.06)
R-ACA	69.01 (35.72)	53.56 (142.19)	11.58 (17.91)	4.97 (9.82)
L-ACA	71.04 (27.62)	25.38 (28.86)	17.89 (19.81)	6.27 (11.51)
Mean	**71.92**	**30.50**	**20.77**	**4.26**

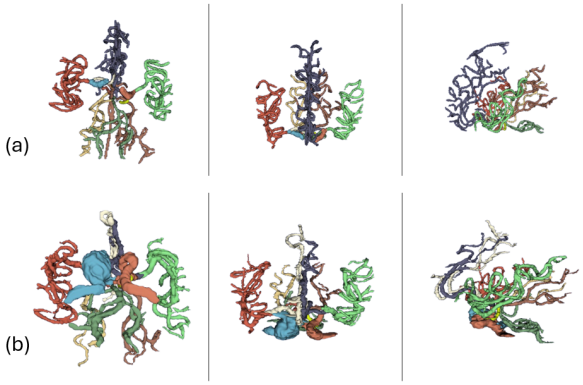

Figure 11. A comparison of the multi-class predicted segmentation (a) and the expert-labeled ground truth segmentation (b) for three different orientations. Notable are the same three main issues identified in the pseudo-labeled image for the same case.

Finally, a qualitative analysis was conducted by comparing the predicted segmentation with the expert-annotated ground truth. First, both were binarized in order to compare whether the predicted segmentation contained the correct number of arteries and whether the topology of the predicted segmentation matched that of the ground truth. This comparison can be seen in Figure 10. Here, it became clear that the model correctly segmented the majority of arteries, but that larger arteries appeared to be under-segmented and smaller arteries appeared to be over-segmented, as was also the case for the pseudo-labeled images when compared to the ground truth.

A qualitative analysis was also performed for the multi-class segmentation, as can be seen in Figure 11. It is apparent that the model suffers from the same limitations that the pseudo-labeled images also suffered from. The ICAs were truncated prematurely, resulting in the model failing to segment the aneurysm in the R-ICA. The model also failed to correctly separate the L-ACA and R-ACA. Despite these issues, the overall segmentation appeared visually consistent with the expected anatomy.

D. CFD results

The ground truth segmentation required more manual post-processing than the predicted segmentation, as the predicted segmentation contained less defects and extrusions. Consequently, the additional effort needed to prepare it for CFD simulation led to longer overall preparation times for the ground truth segmentation compared to the predicted segmentation. Postprocessing of the predicted model required approximately 10 minutes, as opposed to 25 minutes for the ground truth segmentation.

Velocity, pressure and WSS were analyzed across 10 corresponding cross-sections in the predicted and ground truth models as shown in Figure B7 in Appendix B. The model's predictive performance varied across these parameters, as can be seen in Figures 12, 13 and 14. For velocity, the scatterplot demonstrated a weak correlation between the values predicted for the automated and ground truth models (Pearson's r = 0.25, R^2 = -0.99), with many points near the identity line and

a noticeable cluster of over-predictions in the lower velocity range (100–125 mm/s). The corresponding Bland-Altman plot indicated a slight positive bias and a uniform distribution of differences within the 95% limits of agreement. Pressure predictions showed a strong correlation (Pearson's r = 0.91, $R^2 = 0.72$) with predicted values closely aligned to the identity line in the scatterplot. The Bland-Altman plot for pressure revealed minimal bias and narrow 95% limits of agreement, with most data points near zero difference. WSS predictions displayed poor correlation (Pearson's r = 0.04, $R^2 = -1.84$), with the scatterplot showing wide dispersion away from the identity line. The Bland-Altman plot for WSS indicated a near-zero mean bias but broad limits of agreement, reflecting high variability. Mean absolute errors were 48.96 ± 30.69 mm/s for velocity, 7.47 ± 6.07 Pa for pressure and 1.22 ± 0.81 Pa for WSS (all $p < 0.05$). Plots of pressure and WSS over the full meshes are shown in Figure 12, while Figure 13 displays velocity profiles and associated error heatmaps.

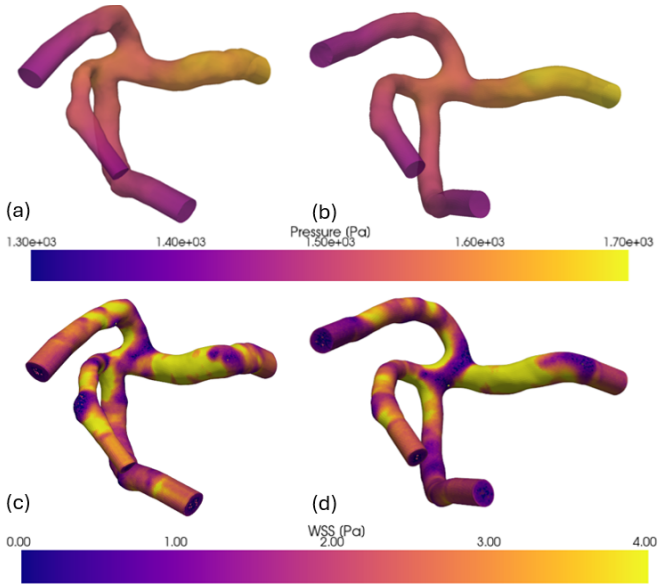

Figure 12. Qualitative comparison of the pressure (a and b) and WSS (c and d) for the ground truth (a and c) and predicted segmentations (b and d). Overall patterns are similar for both parameters, but subtle discrepancies are observed between the predicted and ground truth segmentations, particularly in terms of WSS. These differences are most noticeable in the M2 region, where deviations occur in multiple localized areas along the superior and inferior branches.

4. Discussion

The primary aim of this study was to develop a deep learning segmentation model capable of producing cerebral vessel segmentations suitable for CFD analysis of the CoW and its branching arteries. An accurate vessel segmentation is a key requirement for CFD, as even small anatomical errors can significantly affect downstream simulations. The main challenge in training such a model was the lack of a large,

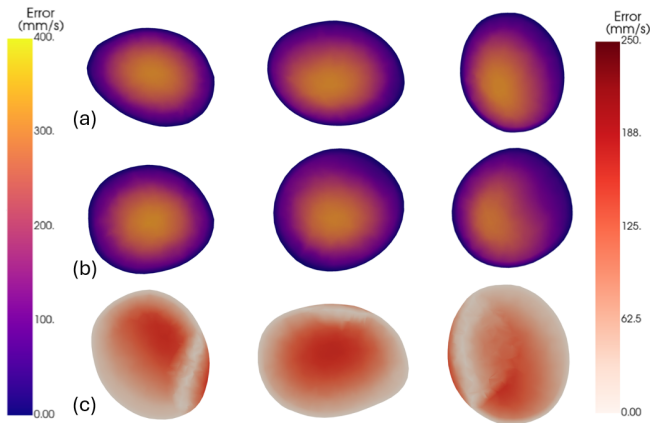

Figure 13. Qualitative comparison of the velocity for the ground truth (a) and predicted (b) segmentations. A heatmap (c) shows the magnitude and location of errors in the predicted segmentation compared to the ground truth. Velocity profiles appear to be relatively similar, with larger absolute errors in the center of the vessel lumen.

expert-annotated CTA dataset for this application. Given the technical, ethical and practical difficulties of obtaining high-quality labeled data in medical imaging [50, 51], a pseudo-labeling strategy was used in order to leverage a large dataset of unlabeled CTA images. An nnU-Net V2 model with a topology-aware loss function was trained and its predictions were evaluated against expert-labeled ground truth segmentations for a single case, with additional CFD validation performed by comparing velocity, pressure and WSS in the MCA.

In this study, pseudo-labels were generated using a region-growing approach to combine binary and multi-class segmentations produced by pretrained models. While this strategy was efficient, it introduced a number of errors which impacted label quality. Quantitative and qualitative evaluations revealed that pseudo-label quality varied, with a mean Dice score of 48% against thresholded CTA images and under- and over-segmentation in large and small vessels, respectively. Visual inspection also revealed that segmentations with fewer than nine labeled classes often suffered from labeling inaccuracies, especially in complex vascular regions such as the ACAs, where spatial proximity and image resolution limitations led to overlap errors, as can be seen in Figure 15. While some errors were filtered out by discarding pseudo-labeled images with outliers in terms of segment volume and segment count, the remaining errors in the dataset may have contributed to the model converging to a suboptimal solution.

Evaluation of the model on the internal test set revealed a nuanced picture. The mean Dice score of 62% and IoU of 51% suggest moderate overlap with the pseudo-labeled test set but substantial class-wise variability was observed. Smaller vessels and the posterior cerebral arteries generally displayed reduced segmentation accuracy, likely due to a combination of anatomical complexity and the higher sensitivity of overlap-

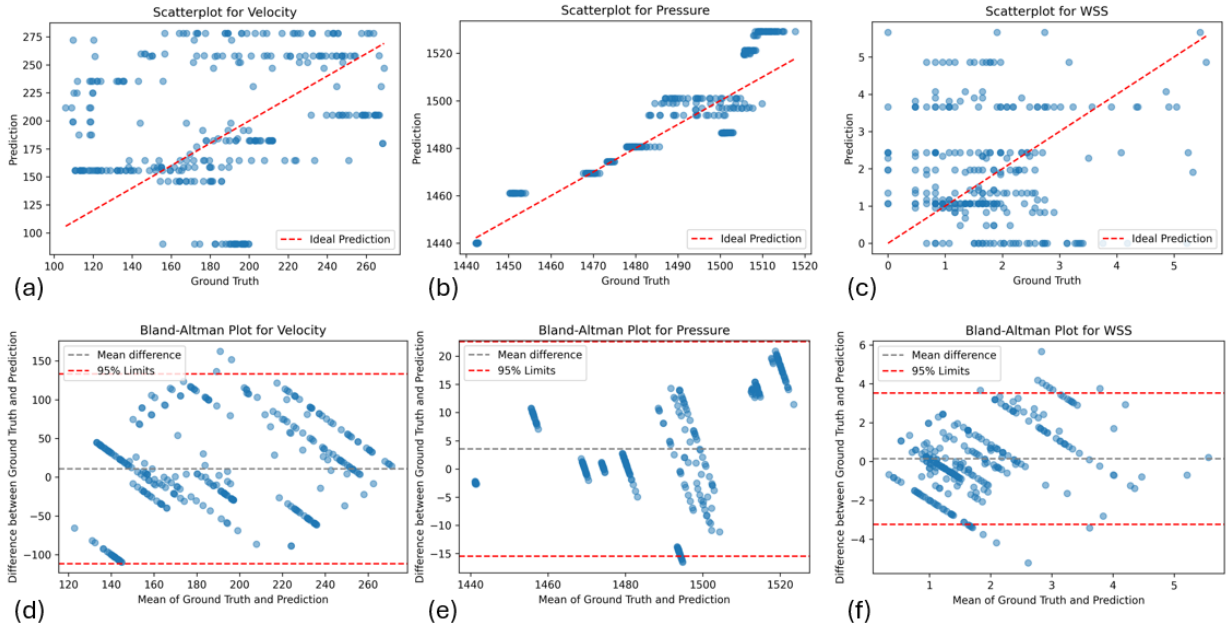

Figure 14. Scatter plots comparing (a) velocity, (b) pressure and (c) WSS values between the ground truth and predicted models. The corresponding Bland–Altman plots are shown in (d) for velocity, (e) for pressure, and (f) for WSS, in order to assess error distributions and identify potential systematic biases. The 95% limits of agreement are shown to indicate the magnitude of the spread of the errors for each parameter.

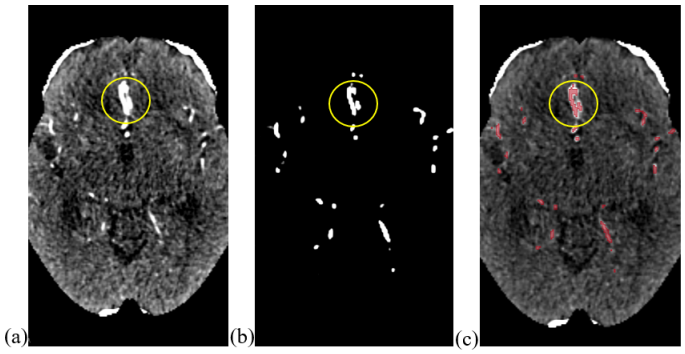

Figure 15. An example of how the L-ACA and R-ACA appear to merge in (a) the original CTA scan, and (b) the DTUNet segmentation. In (c) the voxel-wise overlap of the L-ACA and R-ACA in the DTUNet segmentation is overlaid ontop of the original CTA image.

based metrics to small absolute errors in small volumes. Topological accuracy, as measured by a clDice of 40%, was also limited when compared to the pseudo-labeled images, potentially indicating discontinuities or misaligned centerlines in the predicted vessel trees. Interestingly, visual analysis did not suggest obvious disconnections, which implies that the low clDice score is more likely explained by a centerline mismatch rather than actual structural breaks. A key consideration is whether the predicted centerline is, in fact, more accurate than its pseudo-labeled counterpart; an outcome that is plausible if the segmentation model successfully learned to smooth out noise present in the training data, thereby producing higher-quality results compared to the labels it was trained with. This is possible if noise or errors present in the training data were completely random and did not contain a systematic bias [52].

Additional metrics provided further insight: the HD was relatively high at 17 voxels. Although the CTA images varied in resolution, voxel sizes across all images were comparable to those in the TopCoW dataset, which are all 0.45 mm in the x- and y-directions and 0.7 mm in the z-direction. This would indicate that the HD is between 7.8–11.9 mm. In contrast, the ASD was only 4 voxels, corresponding to 1.8–2.8 mm, which implies that the majority of boundary predictions were reasonably accurate. The poor HD therefore implies that the predictions contain a small number of large segmentation errors when compared to the pseudo-labeled images. Additionally, the model displayed a higher recall of 71% compared to a precision of 61%, which suggests that the model has a tendency to successfully identify and classify the majority of voxels in which a branch of the CoW is present at the expense of including a higher number of false positive voxels. This indicates that the model may produce a slight over-segmentation of the vessels. Visual inspection of the predicted segmentations showed smooth, plausible segmentations with three main errors: over-segmentation of smaller vessels, under-segmentation of larger vessels and misclassification of vessels due to spatial proximity.

Complementing voxel-wise segmentation metrics, centerline-based evaluation provides additional insight into the structural fidelity of the predicted vascular trees.

The mean OV of 72% suggests that the model successfully captures a substantial portion of the arterial network's topology. The ASCD of approximately 4.3 voxels, which corresponds to 1.94-3.01 mm, aligns well with the surface-level ASD metric, indicating reasonable spatial agreement between predicted and reference centerlines.

Importantly, the centerline false positive and false negative rates of 31% and 21% respectively, reveal nuanced error modes. The higher false positive rate corroborates the voxel-level precision-recall imbalance, confirming that over-segmentation is a key contributor to reduced accuracy. However, the presence of false negatives also highlights that some arterial branches are missed by the model, indicating incomplete vessel coverage.

Crucially, these segmentation errors had downstream consequences in the CFD validation stage. Discrepancies between the predicted and expert-annotated segmentations in the MCA affected computed values of blood velocity, pressure and WSS, though not all metrics were impacted to an equal extent. While visual analysis showed a relatively strong similarity between predicted and expert-annotated segmentations for all parameters, numerical analysis of the CFD results revealed more pronounced differences. In terms of velocity prediction, the model demonstrates moderate accuracy, with predicted values aligning with physiological velocity ranges reported in existing literature [53–55]. However, the presence of over-prediction in the lower velocity range suggests a possible bias in the training data or model behavior in low-flow regions. The Bland–Altman plot shows that most errors lie between –110 and +130 mm/s, and the observed homoscedasticity indicates a consistent error magnitude across the entire velocity range. Pressure is the most accurately and consistently predicted variable, though the values in both simulations seem to be lower than those reported in physiological conditions [55, 56], which is likely due to outlet boundary condition applied where a fixed fluid pressure of 0.00144 MPa was prescribed, which is equivalent to 10 mmHg. This value was chosen as it has been used in previous CFD analyses of the MCA [57], which allows for comparison of values achieved with those from previous literature. This comparison showed that the values for the three hemodynamic parameters predicted for the automated segmentation were in line with expected values given this fixed fluid pressure. The smooth spatial behavior of pressure and lower sensitivity to small-scale geometric variations likely contribute to making pressure easier for the model to approximate compared to velocity and WSS. The Bland-Altman plots shows that the absolute errors are low and that the spread of these errors is relatively narrow, with most ranging between -15 and +22 Pa. WSS predictions, on the other hand, seem to be less reliable. While the values seen in both simulations are similar to those reported in literature in physiological conditions [58, 59], the high variability and lack of a consistent trend suggest that WSS is particularly challenging to predict. The Bland–Altman plot reveals heteroscedasticity, with a large error spread ranging from –3.6 to +3.8 Pa. Given that WSS depends on the spatial gradient of velocity near the vessel wall, even small errors in velocity estimation or wall geometry can lead to significant inaccuracies. This emphasizes the need for improved model resolution or alternative techniques when targeting local wall-level hemodynamic metrics.

A key strength of the method presented in this work lies in its development of a scalable pipeline capable of producing vessel segmentations suitable for downstream analysis. The use of pseudo-labels allowed for training on a large dataset that would otherwise have been unusable due to lack of expert annotations. Prior research has shown that pseudo-labeling can lead to good performance in medical image segmentation tasks [60], likely due to the fact that effective learning is achievable as long as the training data, pseudo-labeled or otherwise, accurately reflects the underlying distribution of the target anatomy [61]. It has also been shown that it is possible for a model that has been trained using a large volume of pseudo-labeled data to outperform a model trained on a small amount of expert-annotated data [62]. Furthermore, it has been determined that even in cases where training data contains substantially more erroneous data than accurate data, it is possible for deep learning models to successfully learn underlying patterns [52]. This shows that pseudo-labeling methods such as the one employed in this study, even when not entirely accurate, can be sufficiently reliable for practical segmentation tasks. The segmentations produced by the proposed model were also validated for CFD applications, providing a unique application-oriented evaluation of segmentation quality.

However, limitations of the current method remain. Although previous studies have shown that deep learning models can perform well even when trained on noisy and error-prone labels, in this case it is likely that labeling errors limited the model's segmentation performance. The region-growing method, while efficient, did not always resolve overlaps correctly, especially in arteries such as the ACAs. It should, however, be noted that even an experienced neuroradiologist also found it challenging to successfully separate and segment the R-ACA and the L-ACA, which indicates that this would be a difficult aspect to improve upon. Spatial resolution of the input CTA images may need to be improved for better segmentation of the ACAs. However, the resolution of the majority of CTA scans used in this study was relatively high. This means that in order to obtain better results, a more sophisticated region-growing approach would be necessary. Furthermore, the pipeline's performance is dependent on the accuracy of each individual step; any errors in pseudo-labeling, segmentation, or CFD simulation can propagate into subsequent processes and will affect the final outcome, making error attribution difficult.

One limitation of this study was the lack of an objective ground truth, which had several important consequences. First, it made it difficult to reliably assess pseudo-label quality, forcing reliance on visual inspection and indirect validation. Having a small, manually annotated dataset would have provided a more reliable basis for assessing pseudo-label quality and could have improved the overall training process. Second, due to the absence of ground truth labels, all evaluation

metrics were computed against pseudo-labels that themselves may contain noise and errors, potentially underestimating the true performance of the model. Lastly, the absence of a definitive ground truth posed challenges for validating the predicted segmentation for CFD analyses. In this study, a manual segmentation was created for a single case through expert annotation and treated as the ground truth. However, this reference segmentation was not necessarily the true vessel geometry but rather the best approximation achievable by a human expert, which may still contain inaccuracies or inconsistencies, particularly in complex regions. For example, the manual segmentation was somewhat coarse in parts, which may not capture the precise vessel shape needed for realistic flow simulations. The fact that the manual segmentation also required more postprocessing in order to be made suitable for CFD simulations also indicates that there were errors and inconsistencies present in this segmentation. This uncertainty makes it difficult to objectively assess how far the hemodynamic parameters in the predicted MCA model were from the genuine values. This means that the study can validate that the segmentation model can produce segmentations suitable for CFD but cannot confirm how accurate the predicted segmentations are.

While pseudo-labeling offers scalability, this study reinforces the idea that label quality can form a key bottleneck in achieving clinical-grade segmentation models. Future work could therefore focus on improving pseudo-label quality, for example by setting stricter requirements at the filtering step or by evaluating the reliability of individual pseudo-labels and filtering out low-confidence examples prior to training [63]. Methods should be sought which allow for better pseudo-labeling of the ACAs, perhaps by incorporating more advanced region-growing approaches such as marker-controlled watershed segmentation, graph-cut segmentation or active surfaces. Furthermore, evaluating the trained model against a small, expert-labeled test set could aid in establishing a more accurate benchmark for true performance and inform future iterations of the segmentation pipeline.

This study sought to address a critical gap in the existing literature by developing a segmentation model that not only achieves high voxel-wise classification accuracy but also aims to ensure that the resulting segmented structures are topologically correct and suitable for CFD analyses. Compared to existing approaches, particularly those relying on manual segmentation or traditional preprocessing, this method offers several advantages. By automating the segmentation pipeline, it significantly reduces the time burden on clinicians and researchers while enhancing objectivity, reproducibility and consistency. Additionally, the proposed pseudo-labeling strategy eliminates the need for labor-intensive manual annotation of large datasets, which is both time-consuming and error-prone. However, the results indicate that the segmentation model requires further refinement before it can be reliably used for CFD analysis. Notable discrepancies in blood velocity, pressure and WSS were observed between the predicted and ground truth MCA models, suggesting potential shortcomings in the pseudo-labeling process, model training, or both. Although it is difficult to pinpoint the primary source of error, it is plausible that low-quality pseudo-labels contributed to the model converging to a suboptimal solution. While pseudo-labeling enables the training of segmentation models in data-scarce environments, the error in these labels can limit both segmentation accuracy and downstream utility. This study underscores the delicate balance between dataset size, label quality and model performance, particularly in the high-precision context of medical imaging. Despite current limitations, the development of this segmentation model tailored for CFD analysis of the CoW represents a meaningful step toward enabling more accurate personalized simulations and ultimately improving clinical decision-making in the management of cerebrovascular disease.

5. Conclusions and recommendations

This study demonstrates the potential of using pseudo-labels to train a multi-class segmentation model of the CoW for downstream CFD analysis. While the model successfully generated anatomically plausible segmentations and enabled full CFD simulation, the validation against expert-annotated segmentations revealed significant differences in key hemodynamic metrics, namely blood flow velocity, pressure and wall shear stress. These discrepancies underscore the limitations of the current pseudo-labeling approach and its impact on downstream analysis. Improving the accuracy and consistency of pseudo-labels, particularly in regions with complex vascular geometry and regions where neighboring arteries are in close proximity to one another, will be essential for advancing the clinical utility of such models. Future work should also include evaluation against a smaller, expert-annotated test set to better assess model reliability.

Ethics statement

Ethical approval for this study was obtained from the Human Research Ethics Committee of the Technical University of Delft (HREC No. 5305). Informed consent, including consent for re-use of imaging data, was obtained by the respective institutions at the time of data collection.

Data availability statement

The imaging datasets used in this study include three publicly available datasets and one private dataset. Due to restrictions on patient data privacy and agreements with data providers, the combined imaging dataset used for this study has not been made publicly available. Access to imaging data from the MR CLEAN NO-IV trial can be requested directly from the CONTRAST consortium, with details and procedures available at https://www.contrast-consortium.nl/data-requests-consortium-members-and-trial-collaborators. However, the preprocessing pipeline and model code developed are available upon request.

Acknowledgements

The author would like to express sincere gratitude towards Dr. S. Pirola, Dr. T. van Walsum and F.G. te Nijenhuis for their guidance, feedback and unwavering support throughout the course of this thesis. Special thanks also go to F. Fontana, doctoral candidate in the Department of Biomechanical Engineering at TU Delft, for her assistance with the CFD simulations. The author also wishes to thank Drs. S.A.P. Cornelissen, interventional radiologist at the Department of Radiology and Nuclear Medicine at Erasmus Medical Center, for her support in creating the manual ground truth segmentation.

References

[1] GBD 2021 Stroke Risk Factor Collaborators, "Global, regional, and national burden of stroke and its risk factors, 1990-2021: A systematic analysis for the global burden of disease study 2021," *Lancet Neurol.*, vol. 23, no. 10, pp. 973–1003, Oct. 2024.

[2] V. L. Feigin, M. Brainin, B. Norrving, S. O. Martins, J. Pandian, *et al.*, "World stroke organization: Global stroke fact sheet 2025," *Int. J. Stroke*, vol. 20, no. 2, pp. 132–144, Feb. 2025.

[3] T. G. Shaw, K. F. Mortel, J. S. Meyer, R. L. Rogers, J. Hardenberg, *et al.*, "Cerebral blood flow changes in benign aging and cerebrovascular disease," *Neurology*, vol. 34, no. 7, pp. 855–862, Jul. 1984.

[4] World Stroke Organization, *"Impact of Stroke"*, https://www.world-stroke.org/world-stroke-day-campaign/about-stroke/impact-of-stroke, 2025 (accessed Apr. 2025).

[5] G. J. Rinkel, M. Djibuti, A. Algra, and J. van Gijn, "Prevalence and risk of rupture of intracranial aneurysms: A systematic review," *Stroke*, vol. 29, no. 1, pp. 251–256, Jan. 1998.

[6] C. P. Derdeyn, G. J. Zipfel, F. C. Albuquerque, D. L. Cooke, E. Feldmann, *et al.*, "Management of brain arteriovenous malformations: A scientific statement for healthcare professionals from the american heart association/american stroke association," *Stroke*, vol. 48, no. 8, e200–e224, Aug. 2017.

[7] C.-J. Chen, D. Ding, C. P. Derdeyn, G. Lanzino, R. M. Friedlander, *et al.*, "Brain arteriovenous malformations: A review of natural history, pathobiology, and interventions," *Neurology*, vol. 95, no. 20, pp. 917–927, Nov. 2020.

[8] P. B. Gorelick, K. S. Wong, H.-J. Bae, and D. K. Pandey, "Large artery intracranial occlusive disease: A large worldwide burden but a relatively neglected frontier," *Stroke*, vol. 39, no. 8, pp. 2396–2399, Aug. 2008.

[9] D. Benner, B. K. Hendricks, A. Benet, and M. T. Lawton, "Eponyms in vascular neurosurgery: Comprehensive review of 11 arteries," *World Neurosurgery*, vol. 151, pp. 249–257, 2021.

[10] J. Rosner, V. Reddy, and F. Lui, "Neuroanatomy, circle of willis," in *StatPearls*, Treasure Island, FL, USA: StatPearls Publishing, Jan. 2025. Available: https://www.ncbi.nlm.nih.gov/books/NBK534861/.

[11] R. D. Henderson, M. Eliasziw, A. J. Fox, P. M. Rothwell, and H. J. M. Barnett, "Angiographically defined collateral circulation and risk of stroke in patients with severe carotid artery stenosis," *Stroke*, vol. 31, no. 1, pp. 128–132, Jan. 2000.

[12] H. Lv, K. Fu, W. Liu, Z. He, and Z. Li, "Numerical study on the cerebral blood flow regulation in the circle of willis with the vascular absence and internal carotid artery stenosis," *Front. Bioeng. Biotechnol.*, vol. 12, p. 1 467 257, Aug. 2024.

[13] Royal College of Surgeons of Ireland, *"RCSI - Drawing Circle of Willis - English labels"*, https://anatomytool.org/content/rcsi-drawing-circle-willis-english-labels, 2024 (accessed Apr. 2025).

[14] J. W. Goldfarb, M. Mossa-Basha, K.-L. Nguyen, E. M. Hecht, and J. P. Finn, "Trends in magnetic resonance and computed tomography angiography utilization among medicare beneficiaries between 2013 and 2020," *Clin. Imaging*, vol. 107, p. 110 088, Mar. 2024.

[15] D. A. Katz, M. P. Marks, S. A. Napel, P. M. Bracci, and S. L. Roberts, "Circle of willis: Evaluation with spiral CT angiography, MR angiography, and conventional angiography," *Radiology*, vol. 195, no. 2, pp. 445–449, May 1995.

[16] E. Kalsoum, X. Leclerc, A. Drizenko, and J.-P. Pruvo, "Circle of Willis," in *Encyclopedia of the Neurological Sciences (Second Edition)*, M. J. Aminoff and R. B. Daroff, Eds., Second Edition, Oxford: Academic Press, 2014, pp. 803–805.

[17] S. A. Mayer, T. Viarasilpa, N. Panyavachiraporn, M. Brady, D. Scozzari, *et al.*, "CTA-for-all," *Stroke*, vol. 51, no. 1, pp. 331–334, Jan. 2020.

[18] A. T. Vertinsky, N. E. Schwartz, N. J. Fischbein, J. Rosenberg, G. W. Albers, *et al.*, "Comparison of multidetector CT angiography and MR imaging of cervical artery dissection," *AJNR Am. J. Neuroradiol.*, vol. 29, no. 9, pp. 1753–1760, Oct. 2008.

[19] G. K. Thakur, A. Thakur, S. Kulkarni, N. Khan, and S. Khan, "Deep learning approaches for medical image analysis and diagnosis," *Cureus*, vol. 16, no. 5, e59507, May 2024.

[20] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Front. Public Health*, vol. 11, p. 1 273 253, Nov. 2023.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[22] K. Yang, F. Musio, Y. Ma, N. Juchler, J. C. Paetzold, *et al.*, "Benchmarking the CoW with the TopCoW

challenge: Topology-aware anatomical segmentation of the circle of willis for CTA and MRA," Dec. 2023.

[23] P. D. Morris, A. Narracott, H. von Tengg-Kobligk, D. A. Silva Soto, S. Hsiao, *et al.*, "Computational fluid dynamics modelling in cardiovascular medicine," *Heart*, vol. 102, no. 1, pp. 18–28, Jan. 2016.

[24] M. J. Colebank, L. M. Paun, M. U. Qureshi, N. Chesler, D. Husmeier, *et al.*, "Influence of image segmentation on one-dimensional fluid dynamics predictions in the mouse pulmonary arteries," *J. R. Soc. Interface*, vol. 16, no. 159, p. 20 190 284, Oct. 2019.

[25] J. Liu, Z. Yan, Y. Pu, W.-S. Shiu, J. Wu, *et al.*, "Functional assessment of cerebral artery stenosis: A pilot study based on computational fluid dynamics," *J. Cereb. Blood Flow Metab.*, vol. 37, no. 7, pp. 2567–2576, Jul. 2017.

[26] X. Leng, F. Scalzo, H. L. Ip, M. Johnson, A. K. Fong, *et al.*, "Computational fluid dynamics modeling of symptomatic intracranial atherosclerosis may predict risk of stroke recurrence," *PLoS One*, vol. 9, no. 5, e97531, May 2014.

[27] M. Castro, C. Putman, and J. Cebral, "Computational fluid dynamics modeling of intracranial aneurysms: Effects of parent artery segmentation on intra-aneurysmal hemodynamics," *American Journal of Neuroradiology*, vol. 27, no. 8, pp. 1703–1709, 2006.

[28] D. Lesage, E. D. Angelini, I. Bloch, and G. Funka-Lea, "A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes," *Med. Image Anal.*, vol. 13, no. 6, pp. 819–845, Dec. 2009.

[29] N. Mu, Z. Lyu, M. Rezaeitaleshmahalleh, J. Tang, and J. Jiang, "An attention residual u-net with differential preprocessing and geometric postprocessing: Learning how to segment vasculature including intracranial aneurysms," *Med. Image Anal.*, vol. 84, p. 102 697, Feb. 2023.

[30] M. Bertolini, G. Luraghi, I. Belicchi, F. Migliavacca, and G. Colombo, "Evaluation of segmentation accuracy and its impact on patient-specific CFD analysis," *Int. J. Interact. Des. Manuf. (IJIDeM)*, vol. 16, no. 2, pp. 545–556, Jun. 2022.

[31] J. Montalt-Tordera, E. Pajaziti, R. Jones, E. Sauvage, R. Puranik, *et al.*, "Automatic segmentation of the great arteries for computational hemodynamic assessment," *J. Cardiovasc. Magn. Reson.*, vol. 24, no. 1, p. 57, Nov. 2022.

[32] G. Yankova, D. Tur, D. Parshin, A. Cherevko, and A. Akulov, "Cerebral arterial architectonics and cfd simulation in mice with type 1 diabetes mellitus of different duration," *Scientific Reports*, vol. 11, Feb. 2021.

[33] F. Dumais, M. P. Caceres, F. Janelle, K. Seifeldine, N. Arès-Bruneau, *et al.*, "EICaB: A novel deep learning pipeline for circle of willis multiclass segmentation and analysis," *NeuroImage*, vol. 260, p. 119 425, 2022.

[34] H. Bogunović, J. M. Pozo, R. Cárdenes, L. San Román, and A. F. Frangi, "Anatomical labeling of the circle of willis using maximum a posteriori probability estimation," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 1587–1599, 2013.

[35] D. Robben, S. Sunaert, V. Thijs, G. Wilms, F. Maes, *et al.*, "Anatomical labeling of the circle of willis using maximum a posteriori graph matching," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 16, Springer, 2013, pp. 566–573.

[36] M. Song, S. Wang, Q. Qian, Y. Zhou, Y. I. Luo, *et al.*, *CMHA: Intracranial aneurysm CTA image & 3D model dataset with clinical, morphological, hemodynamic data*, Nov. 2024.

[37] Z.-H. Bo, H. Qiao, C. Tian, Y. Guo, W. Li, *et al.*, "Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network," *Patterns*, vol. 2, no. 2, p. 100 197, 2021.

[38] N. E. LeCouffe, M. Kappelhof, K. M. Treurniet, L. A. Rinkel, A. E. Bruggeman, *et al.*, "A randomized trial of intravenous alteplase before endovascular treatment for stroke," *N. Engl. J. Med.*, vol. 385, no. 20, pp. 1833–1844, Nov. 2021.

[39] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, *et al.*, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, e230024, 2023.

[40] B. B. Avants, N. J. Tustison, G. Song, and J. C. Gee, "Ants: Advanced open-source normalization tools for neuroanatomy," *Penn Image Computing and Science Laboratory*, 2009.

[41] Y. Kirchhoff, M. R. Rokuss, S. Roy, B. Kovacs, C. Ulrich, *et al.*, "Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures," in *Lecture Notes in Computer Science*, Cham: Springer Nature Switzerland, 2024, pp. 218–234.

[42] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[43] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, *et al.*, "NnU-Net revisited: A call for rigorous validation in 3D medical image segmentation," Apr. 2024.

[44] S. Liu, R. Su, J. Su, W. H. van Zwam, P. J. van Doormaal, *et al.*, "Segmentation-assisted vessel centerline extraction from cerebral CT angiography," en, *Med. Phys.*, Apr. 2025.

[45] L. B. Hindenes, A. K. Håberg, L. H. Johnsen, E. B. Mathiesen, D. Robben, *et al.*, "Variations in the circle of willis in a large population sample using 3D TOF angiography: The tromsø study," *PLoS One*, vol. 15, no. 11, e0241373, Nov. 2020.

[46] K. Sommer, R. L. Izzo, L. Shepard, A. R. Podgorsak, S. Rudin, *et al.*, "Design optimization for accurate flow simulations in 3D printed vascular phantoms derived from computed tomography angiography," in *Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, T. S. Cook and J. Zhang, Eds., Orlando, Florida, United States: SPIE, Mar. 2017.

[47] T. M. Inc., *Matlab version: 9.13.0 (r2022b)*, Natick, Massachusetts, United States, 2022. [Online]. Available: https://www.mathworks.com.

[48] S. A. Maas, B. J. Ellis, G. A. Ateshian, and J. A. Weiss, "FEBio: Finite elements for biomechanics," en, *J. Biomech. Eng.*, vol. 134, no. 1, p. 011 005, Jan. 2012.

[49] D. H. P. C. C. (DHPC), *DelftBlue Supercomputer (Phase 2)*, https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2, 2024.

[50] K. Magudia, C. P. Bridge, K. P. Andriole, and M. H. Rosenthal, "The trials and tribulations of assembling large medical imaging datasets for machine learning applications," *J. Digit. Imaging*, vol. 34, no. 6, pp. 1424–1429, Dec. 2021.

[51] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, *et al.*, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020.

[52] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, *"Deep Learning is Robust to Massive Label Noise"*, 2018. arXiv: 1705.10694. [Online]. Available: https://arxiv.org/abs/1705.10694.

[53] K. F. Lindegaard, T. Lundar, J. Wiberg, D. Sjøberg, R. Aaslid, *et al.*, "Variations in middle cerebral artery blood flow investigated with noninvasive transcranial blood velocity measurements," *Stroke*, vol. 18, no. 6, pp. 1025–1030, Nov. 1987.

[54] W. A. Kofke, P. Brauer, R. Policare, S. Penthany, D. Barker, *et al.*, "Middle cerebral artery blood flow velocity and stable xenon-enhanced computed tomographic blood flow during balloon test occlusion of the internal carotid artery," *Stroke*, vol. 26, no. 9, pp. 1603–1606, Sep. 1995.

[55] L. A. Lipsitz, S. Mukai, J. Hamner, M. Gagnon, and V. Babikian, "Dynamic regulation of middle cerebral artery blood flow velocity in aging and hypertension," *Stroke*, vol. 31, no. 8, pp. 1897–1903, Aug. 2000.

[56] T. Shima, Y. Okada, S. Matsumura, M. Nishida, T. Yamada, *et al.*, "Cortical arterial pressure and anastomotic blood flow measurements during STA-MCA anastomosi," en, *Neurol. Med. Chir. (Tokyo)*, vol. 28, no. 4, pp. 340–345, 1988.

[57] S. E. Razavi, V. Farhangmehr, and N. Zendeali, "Numerical investigation of the blood flow through the middle cerebral artery," *BioImpacts*, vol. 8, no. 3, pp. 195–200, May 2018.

[58] H. G. Woo, H.-G. Kim, K. M. Lee, S. H. Ha, H. Jo, *et al.*, "Wall shear stress associated with stroke occurrence and mechanisms in middle cerebral artery atherosclerosis," *J. Stroke*, vol. 25, no. 1, pp. 132–140, Jan. 2023.

[59] M. Shojima, "Magnitude and role of wall shear stress on cerebral aneurysm: Computational fluid dynamic study of 20 middle cerebral artery aneurysms," *Stroke*, Oct. 2004.

[60] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2019, pp. 810–818.

[61] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, *et al.*, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Appl. Sci. (Basel)*, vol. 11, no. 2, p. 796, Jan. 2021.

[62] J. A. Fries, P. Varma, V. S. Chen, K. Xiao, H. Tejeda, *et al.*, "Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences," *Nat. Commun.*, vol. 10, no. 1, p. 3111, Jul. 2019.

[63] J. Su, Z. Luo, S. Lian, D. Lin, and S. Li, "Mutual learning with reliable pseudo label for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 94, p. 103 111, May 2024.

APPENDIX

A. *Supplementary Methods*

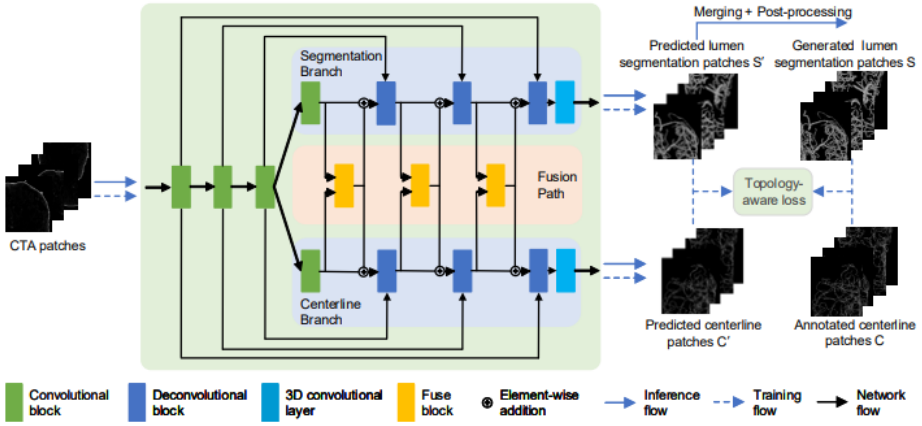

Figure A1. A graphical illustration of the DTUNet architecture. The figure highlights the segmentation and centerline branches, which are subsequently integrated by a fusion branch to enable simultaneous lumen segmentation and centerline prediction from CTA image patches. Adapted from Liu et al. [44].

B. *Supplementary Results*

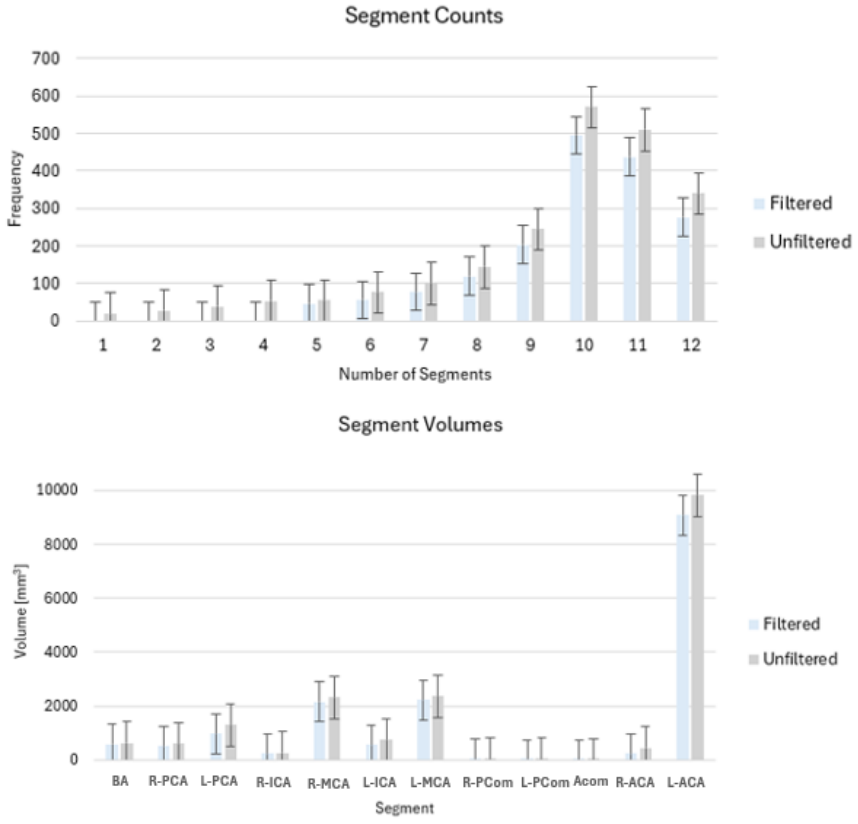

Figure B1. Comparison of segment counts and volumes between filtered and unfiltered pseudo-labeled images. For both metrics, data from filtered images are compared against unfiltered images which include volume and segment count outliers. Error bars indicate the standard deviation of the measurements.

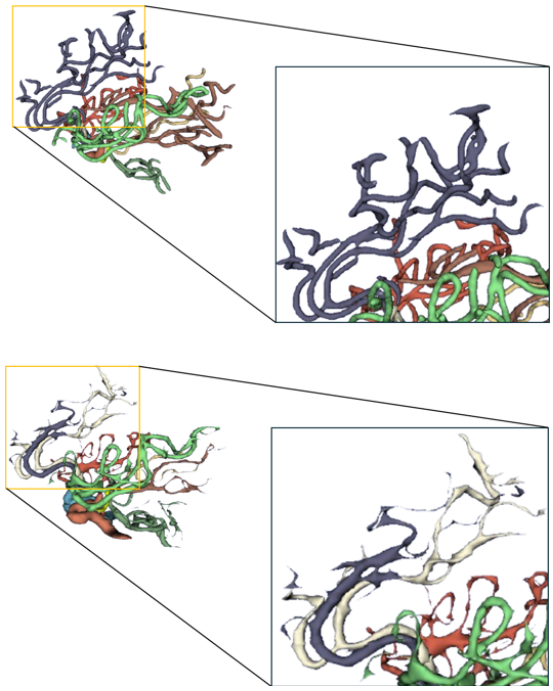

Figure B2. Pseudo-labeling discrepancies between L-ACA and R-ACA. A close-up evaluation highlights challenges in pseudo-labeling the L- and R-ACA, with the pseudo-labels in the top pane and the expert-labeled ground truth in the bottom. The magnified insets, corresponding to the yellow-boxed regions, visually demonstrate instances where the pseudo-labels fail to distinguish between these two segments, indicating misclassification or an inability to maintain accurate anatomical separation.

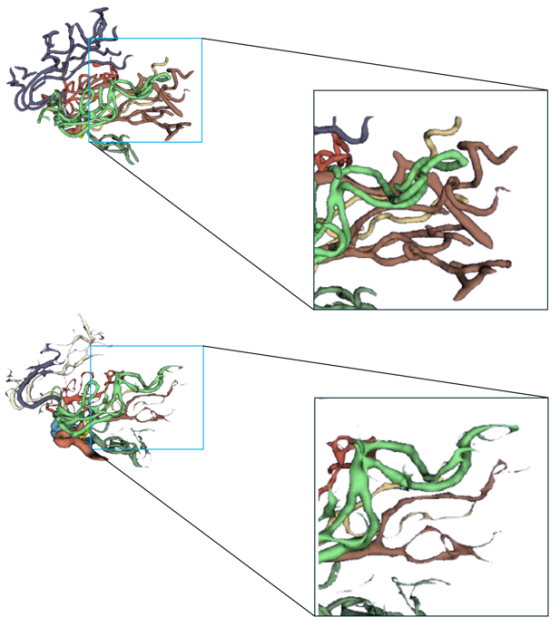

Figure B3. Over-segmentation in distal cerebral artery segments in pseudo-labels. A close-up visual assessment of the distal parts of the segmented cerebral arteries in the pseudo-labels in the top pane and the expert-labeled ground truth in the bottom pane. The areas indicated by the blue boxes reveal over-segmentation of smaller vessels, highlighted by the magnified insets. This illustrates a key challenge in accurately delineating the finer, more complex distal regions.

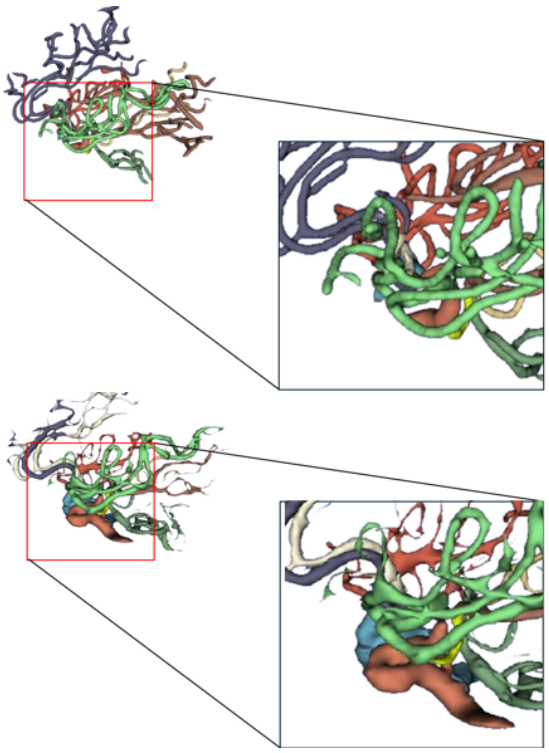

Figure B4. Truncation in pseudo-labeling of proximal ICA segments. A close-up evaluation of the proximal cerebral arteries shows that both ICAs are truncated prematurely in the pseudo-labels, shown in the top pane, compared to the expert-labeled ground truth in the bottom pane. The red-boxed insets highlight the missing portions of the segments, suggesting that the pseudo-labels fail to capture the full extent of the ICAs in these regions.

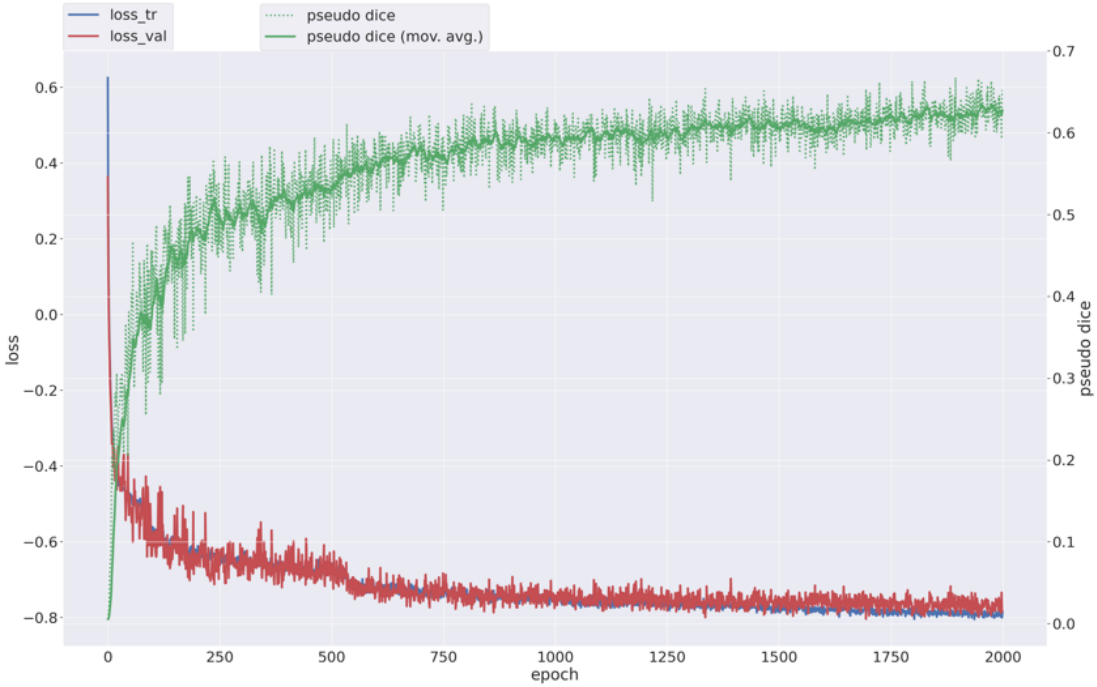

Figure B5. The learning curve for a single fold of the proposed model. This shows training (blue) and validation (red) losses over 2000 epochs, both decreasing over time. The pseudo dice score (green) improves overall, and appears to plateau around a pseudo dice score of 60%.

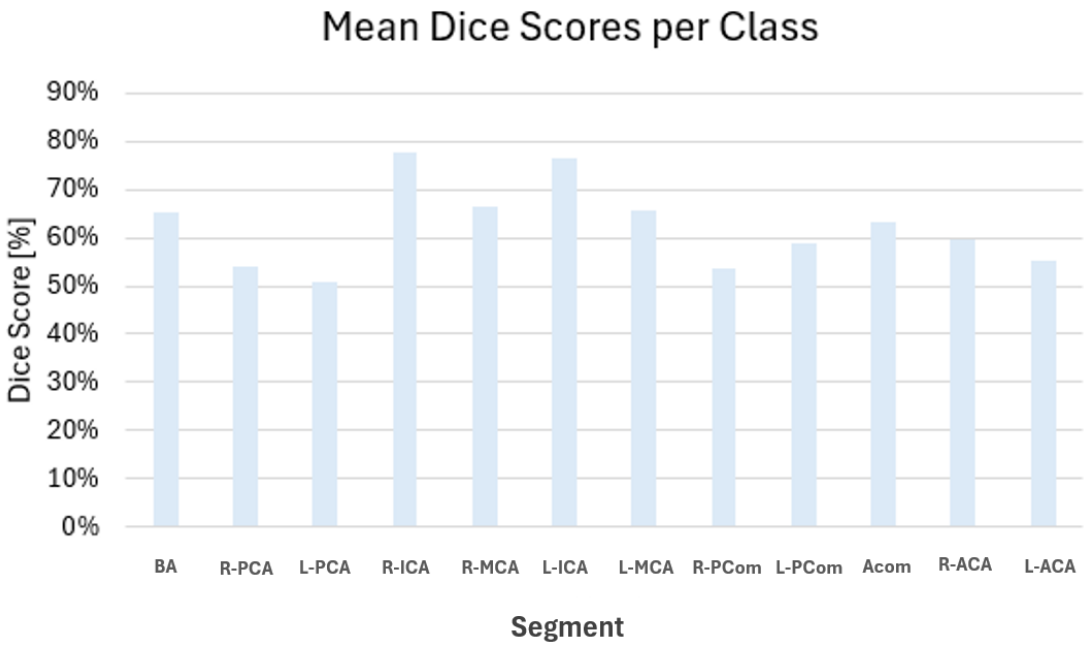

Figure B6. The mean Dice scores achieved per class. This figure illustrates the mean Dice scores obtained for the segmentation of each distinct anatomical vascular segment. This provides a detailed breakdown of segmentation accuracy across different parts of the vascular network.

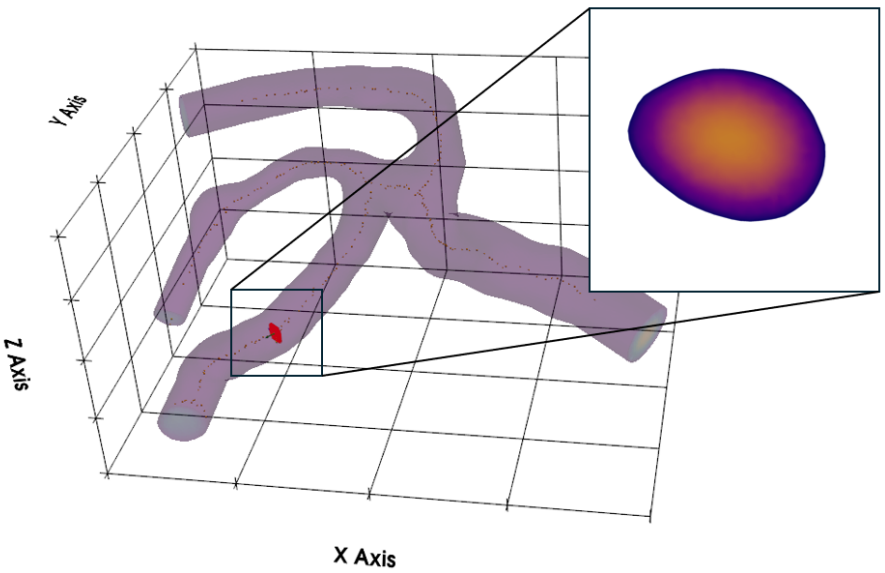

Figure B7. Cross-sectional analysis of the MCA. A representative cross-section is shown at a randomly selected centerline point in the ground truth model. The zoomed-in inset displays velocity across the lumen at that location, perpendicular to the centerline. The corresponding slice in the predicted model was identified by minimizing Euclidean distance between centerline points, and velocity, pressure and WSS were compared across slices for corresponding data points.

C. Derivation of formula for wall shear stress

By Cauchy's stress theorem, the traction vector $\mathbf{T}_n$ acting on a surface with unit normal vector $\mathbf{n}$ can be calculated using:

$$\mathbf{T}_n = \boldsymbol{\sigma} \cdot \mathbf{n}$$

where $\boldsymbol{\sigma}$ is the normal stress. This can be decomposed into two components via vector projection: one perpendicular to the surface, and one tangential to the surface, as shown:

$$\mathbf{T}_n = \sigma_n \mathbf{n} + \boldsymbol{\tau}$$

where $\boldsymbol{\tau}$ is the shear stress vector. This decomposition step can be seen in Figure C1. Rearranging this equation gives:

$$\boldsymbol{\tau} = \mathbf{T}_n - \sigma_n \mathbf{n}$$

To find the magnitude of $\boldsymbol{\tau}$, use the dot product property:

$$\|\boldsymbol{v}\|^2 = \mathbf{v} \cdot \mathbf{v}$$

In order to find the magnitude squared:

$$\|\boldsymbol{\tau}\|^2 = (\mathbf{T}_n - \sigma_n \mathbf{n}) \cdot (\mathbf{T}_n - \sigma_n \mathbf{n})$$

And expand:

$$\|\boldsymbol{\tau}\|^2 = \mathbf{T}_n \cdot \mathbf{T}_n - \sigma_n(\mathbf{T}_n \cdot \mathbf{n}) - \sigma_n(\mathbf{n} \cdot \mathbf{T}_n) + \sigma_n^2(\mathbf{n} \cdot \mathbf{n}).$$

Then simplify using the following additional properties:

$$\mathbf{T}_n \cdot \mathbf{n} = \sigma_n, \quad \mathbf{n} \cdot \mathbf{n} = 1,$$

In order to achieve:

$$\begin{aligned} \|\boldsymbol{\tau}\|^2 &= \|\mathbf{T}_n\|^2 - 2\sigma_n^2 + \sigma_n^2 \\ &= \|\mathbf{T}_n\|^2 - \sigma_n^2. \end{aligned}$$

Therefore, the magnitude of the shear stress vector is given by:

$$\|\boldsymbol{\tau}\| = \sqrt{\|\mathbf{T}_n\|^2 - \sigma_n^2}$$

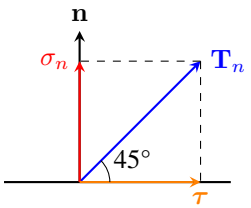

Figure C1. Decomposition of the traction vector $\mathbf{T}_n$ into normal stress σ_n and wall shear stress $\boldsymbol{\tau}$.