

Automated Labeling of Intracranial Arteries with Uncertainty Quantification Using Deep Learning

Javier Bisbal^{1,2,3,4,8,**}, Patrick Winter^{8,10}, Sebastian Jofre⁵, Aaron Ponce^{4,6}, Sameer A. Ansari^{10,11,12}, Ramez Abdalla^{10,11}, Michael Markl¹⁰, Oliver Welin Odeback¹, Sergio Uribe⁷, Cristian Tejos^{2,3,4}, Julio Sotelo⁵, Susanne Schnell^{8,10,*}, David Marlevi^{1,9,*}

Abstract

Accurate anatomical labeling of intracranial arteries is essential for cerebrovascular diagnosis and hemodynamic analysis but remains time-consuming and subject to interoperator variability. We present a deep learning-based framework for automated artery labeling from 3D Time-of-Flight Magnetic Resonance Angiography (3D ToF-MRA) segmentations ($n=35$), incorporating uncertainty quantification to enhance interpretability and reliability. We evaluated three convolutional neural network architectures: (1) a UNet with residual encoder blocks, reflecting commonly used baselines in vascular labeling; (2) CS-Net, an attention-augmented UNet incorporating channel and spatial attention mechanisms for enhanced curvilinear structure recognition; and (3) nnUNet, a self-configuring framework that automates preprocessing, training, and architectural adaptation based on dataset characteristics. Among these, nnUNet achieved the highest labeling performance (average Dice score: 0.922; average surface distance: 0.387 mm), with improved robustness in anatomically complex vessels. To assess predictive confidence, we implemented test-time augmentation (TTA) and introduced a novel coordinate-guided strategy to reduce interpolation errors during augmented inference. The resulting uncertainty maps reliably indicated regions of anatomical ambiguity, pathological variation, or manual labeling inconsistency. We further validated clinical utility by comparing flow velocities derived from automated and manual labels in co-registered 4D Flow MRI datasets, observing close agreement with no statistically significant differences. Our framework offers a scalable, accurate, and uncertainty-aware solution for automated cerebrovascular labeling, supporting downstream hemodynamic analysis and facilitating clinical integration.

Keywords: Intracranial artery labeling, 3D ToF-MRA, Deep learning, UNet, Intracranial 4D Flow MRI

1. Introduction

The intracranial arterial system plays a critical role in brain perfusion to maintain normal cognitive function. Occlusion or stenosis of these blood vessels can cause vascular alterations that contribute to the development of cerebrovascular or neurodegenerative diseases [1, 2].

*Authors contributed equally.

**Corresponding author.

Email address: jebisbal@uc.cl (Javier Bisbal)

¹Dept. of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden

²Biomedical Imaging Center, Pontificia Universidad Católica de Chile, Santiago, Chile

³Department of Electrical Engineering, School of Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

⁴Millennium Institute for Intelligent Healthcare Engineering (iHEALTH), Santiago, Chile

⁵Departamento de Informática, Universidad Técnica Federico Santa María, Santiago, Chile

⁶Escuela de Ingeniería Civil Informática, Universidad de Valparaíso, Valparaíso, Chile

⁷Department of Medical Imaging and Radiation Sciences, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia

⁸Department of Medical Physics, Institute of Physics, University of Greifswald, Greifswald, Germany

⁹Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

¹⁰Department of Radiology, Northwestern University, Chicago, IL, USA

¹¹Department of Neurological Surgery, Northwestern University, Chicago, IL, USA

¹²Department of Neurology, Northwestern University, Chicago, IL, USA

Three-dimensional Time-of-Flight Magnetic Resonance Angiography (3D ToF-MRA) is the clinical gold standard for non-invasive imaging of the intracranial vasculature. Recently, 4D Flow MRI has emerged as a new modality, adding valuable functional hemodynamic data including regional blood flow variations. In particular, intracranial 4D Flow MRI has shown promise in assessing a variety of vascular pathologies, including aneurysms [3, 4, 5], arteriovenous malformations [6, 7], and intracranial atherosclerotic disease (ICAD) [8, 9].

Accurate quantification of 4D Flow MRI data is highly dependent on both segmentation and precise anatomical labeling of the intracranial arteries [10, 11]. Although several methods have been proposed for automated segmentation of the vascular tree [12, 13, 14, 15], our study focuses on automated labeling of the major vascular structures. This process remains one of the most labor intensive and clinically critical tasks in cerebrovascular imaging and is often significantly affected by interoperator variability [11].

To achieve automated vessel labeling, traditional graph-based approaches model the vascular tree using relational graphs derived from the centerlines [16, 17, 18]. However, their performance is often compromised in cases of severe stenosis or disconnected vasculature. Moreover, these methods typically overlook contextual information embedded in the 3D image space.

Deep learning (DL) models offer an alternative by learning vessel features directly from image data. These models leverage structural and spatial context to produce anatomically consistent labeling results. Most of the current literature focuses on variants of the UNet architecture [19, 20, 21, 22, 23], which have shown strong baseline performance. However, the field lacks standardized procedures for selecting and designing network architectures, which directly affect performance and generalization [24]. Additionally, advanced architectural components, such as spatial attention mechanisms, which have shown benefits in segmenting curvilinear structures [13], have not been explored in intracranial artery labeling. Furthermore, current approaches do not incorporate uncertainty quantification, limiting explainability and adoption in clinical settings.

To address these challenges, we performed an evaluation of UNet-based architectures to label intracranial arteries using 3D ToF-MRA, focusing on three key objectives. First, to address the lack of standardization, we leverage the self-configuring nnUNet framework, which automatically tailors the network architecture and training pipeline to the dataset [25]. Second, we investigate architectural refinements by adapting and assessing spatial attention mechanisms, previously applied to the segmentation of curvilinear structures, to intracranial artery labeling and evaluate their potential to improve anatomical labeling [13]. Third, to enable uncertainty-aware predictions, we incorporate test-time augmentation (TTA) [26] and introduce a novel coordinate-guided strategy to reduce interpolation errors during inference, thus improving the reliability of uncertainty estimates.

Together, these contributions aim to build a more robust and scalable framework for automated labeling of intracranial arteries, which is a critical step for subsequent advanced analyzes, including hemodynamic assessments using 4D Flow MRI.

2. Methods

2.1. Study cohort

We retrospectively selected 25 patients (11 females) from an IRB-approved ICAD study at Northwestern Memorial Hospital. Fourteen cases exhibited severe stenosis, defined as constriction > 70%, while the remaining cases showed moderate stenosis, with constriction ranging between 50% and 70%. The affected vessels included the middle cerebral arteries (MCAs), internal carotid arteries (ICAs), and the basilar artery (BA). Two interventional neuroradiologists (RA, SA) reviewed the clinical electronic medical records from MRI/MRA and MR vessel wall imaging to confirm ICAD-related stenoses.

We also included data from ten healthy volunteers (six females). Informed written consent was obtained from all participants in this study. A summary of demographic and physiological data for patients and volunteers is shown in Table 1.

2.2. MRI Acquisitions

Patients and volunteers were scanned using a clinical MRI protocol designed for intracranial vascular assessment. This protocol included a gradient-echo 3D ToF-MRA sequence to visualize vascular anatomy, and an ECG-triggered

Table 1: Median age, BMI, and average heart rate (maximum and minimum values) for the Control and ICAD groups.

Parameter	Controls (n=10)	ICAD (n=25)
Age (years)	27 (19–35)	64 (34–85)
BMI (kg/m ²)	30.30 (19.37–38.73)	27.99 (21.91–41.29)
Heart Rate (bpm)	90.6 (72.6–121.2)	73.8 (63.0–115.2)

intracranial 4D Flow MRI sequence to capture blood flow dynamics. The 4D Flow MRI sequence used a dual-velocity encoding acquisition (dual-VENC) and was accelerated using PEAK GRAPPA with an acceleration factor of $R = 5$, as described in [27]. All scans were performed at 3T (Siemens MAGNETOM Skyra, Erlangen, Germany; and Siemens MAGNETOM Prisma Fit) using a 20-channel head/neck coil (Siemens, Erlangen, Germany). Detailed scan parameters for 3D ToF-MRA and 4D Flow MRI are provided in Table 2.

Table 2: Scan parameters for 3D ToF-MRA and dual-VENC 4D Flow MRI for Control and ICAD groups.

3D ToF-MRA Parameters		
Parameters	Controls	ICAD
TR [ms]	22	21
TE [ms]	3.42	3.4
Voxel size [mm]	0.52	0.52
Slice thickness [mm]	0.50	0.50
Flip angle [°]	17	17
Scan time [min]	4–5	4–5
Siemens MR system	Prisma Fit	Skyra
Dual-VENC 4D Flow MRI Parameters		
Parameters	Controls	ICAD
TR [ms]	5.9	6.1–6.2
TE [ms]	3.25	3.4
Temporal resolution [ms]	82.6	42.7 - 86.8
Voxel size [mm]	0.982	0.978–1.146
Slice thickness [mm]	1.0	1.0–1.2
Number of slices	44	40–60
Number of cardiac phases	5–9	5–18
Flip angle [°]	15	15
Low venc/high venc [m/s]	0.5–0.6 / 1.0–1.2	0.5–0.6 / 1.0–1.2
Siemens MR system	Prisma Fit	Skyra

2.3. Automated labeling

2.3.1. Data preparation

We generated binary masks of the cerebral vasculature using semi-automatic thresholding applied to the 3D ToF-MRA images. As in previous work [8], an in-house algorithm was employed to automatically extract the centerlines of the cerebral vasculature. The identified centerlines were then manually labeled according to the anatomical section of the vessels. We focused our study on annotating nine major vessels of the Circle of Willis: the Basilar Artery (BA), Right and Left Internal Carotid Arteries (RICA and LICA), right and left middle cerebral arteries (RMCA and LMCA), Right and Left Anterior Cerebral Arteries (RACA and LACA), and Right and Left Posterior Cerebral Arteries (RPCA and LPCA). These vessels are illustrated in Figure 1.

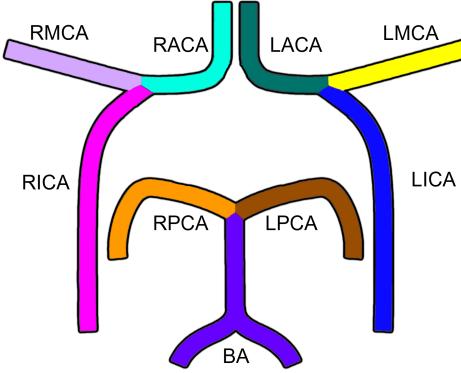


Figure 1: Schematic representation of the nine intracranial artery segments included in this study: basilar artery (BA); right and left internal carotid arteries (RICA, LICA); right and left middle cerebral arteries (RMCA, LMCA); right and left anterior cerebral arteries (RACA, LACA); and right and left posterior cerebral arteries (RPCA, LPCA).

To generate voxel-wise ground truth labels, we used a $7 \times 7 \times 7$ voxel neighborhood around each centerline position. Each voxel within this neighborhood was assigned the label of its closest centerline point. This procedure produced multi-class masks for each binary ToF-MRA segmentations. Voxels without a centerline point within their $7 \times 7 \times 7$ neighborhood were classified as "non-annotated".

2.3.2. Network architectures

Our work evaluated three state-of-the-art UNet variants for semantic segmentation, each chosen with a distinct objective: 1) automate preprocessing, hyperparameter tuning, and training strategies; 2) integrate advanced architectural refinements; and 3) benchmark against previous work.

1. **Self-configuring UNet [25]** (nnUNet): nnUNet is a self-configuring framework that automatically optimizes the UNet with a residual encoder architecture and a training pipeline based on the input dataset. Unlike the other variants, nnUNet does not introduce new architectural innovations, but instead focuses on automating preprocessing, hyperparameter tuning, and training strategies. This approach ensures that the model is tailored to the specific characteristics of the dataset, often achieving state-of-the-art performance without manual intervention. [23]
2. **Channel and Spatial Attention Network [13]** (CS-Net): CS-Net enhances the standard UNet architecture by integrating channel and spatial attention mechanisms. These mechanisms enable the network to better capture fine-grained details and contextual information, making this variant particularly well suited for tasks requiring precise segmentation of curvilinear structures. Originally designed for segmentation, we adapted its implementation for our labeling task.
3. **UNet [28]** (baseline): This variant replaces the standard UNet encoder with a residual encoder, incorporating skip connections at each encoder layer to improve gradient flow and mitigate the vanishing gradient problem [29]. As it represents the architecture used in recent published work on automated intracranial labeling [20, 22], this UNet variant serves as our baseline model.

2.3.3. Preprocessing

Two preprocessing pipelines were implemented, ensuring that the input images aligned with the input formats required for the UNet, CS-Net, and nnUNet setups, respectively.

1. **Scaling with zero-padding and cropping (UNet and CS-Net):** Convolutional networks typically require input images of a fixed size. To handle images of varying dimensions, we first extracted a bounding box that encapsulated each segmentation, adding an empirically chosen 15% zero-padding margin along each dimension. Each image was also scaled to match the largest dimension of the target dimension, while preserving the original aspect ratio. This ensured that the anatomical structures were not distorted. Subsequently, we applied cropping to adjust the image to the exact target shape. For this application, we used a target dimension of $128 \times 256 \times 256$ pixels.

2. Patch inference (nnUNet): As part of the nnUNet preprocessing pipeline, the proposed patch-based approach was utilized directly, allowing for input images of different sizes. Patches of size equal to the median shape of the dataset ($80 \times 224 \times 160$ pixels) were extracted using a sliding window strategy [24]. This allowed the model to predict on images of varying dimensions without resizing.

2.3.4. Loss function

For C classes and N voxels, let $p_{i,c}$ be the predicted probability and $g_{i,c} \in \{0, 1\}$ the one-hot ground truth. The Dice coefficient per class is

$$\text{Dice}(c) = \frac{2 \sum_i^N p_{i,c} g_{i,c}}{\sum_i^N p_{i,c} + \sum_i^N g_{i,c}}, \quad (1)$$

and the Dice loss is the complement averaged over classes:

$$L_{\text{Dice}} = 1 - \frac{1}{C} \sum_{c=1}^C \text{Dice}(c). \quad (2)$$

The cross-entropy loss is the mean over voxels and classes:

$$L_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C g_{i,c} \log(p_{i,c}). \quad (3)$$

We utilized a hybrid loss that include cross-entropy (L_{CE}) and Dice (L_{Dice}) losses with dynamic weights [20], defined as

$$L = \begin{cases} L_{\text{CE}} & \text{epoch } \leq \beta \\ \alpha L_{\text{CE}} + (1 - \alpha)L_{\text{Dice}} & \beta < \text{epoch} \leq \gamma \\ 0.9L_{\text{Dice}} + 0.1L_{\text{CE}}, & \gamma < \text{epoch} \leq \text{total} \end{cases} \quad (4)$$

where

$$\alpha = 0.8 \left(1 - \frac{\text{epoch} - \beta}{\gamma - \beta} \right) + 0.1. \quad (5)$$

This loss and the accompanying weights ensure that, during the initial training epochs ($\leq \beta$), the networks focus on minimizing the cross-entropy loss, with more stable convergence. Between epochs β and γ , the contribution of the two losses is balanced by the weight factor α , which gradually decreases as training progresses, shifting the emphasis from cross-entropy to Dice loss. Finally, after epoch γ , the loss stabilizes with a fixed weighting of 0.9 for Dice and 0.1 for cross-entropy, so the networks prioritize Dice loss, which, although less stable, can lead to more accurate predictions [20].

Table 3: Architecture details of UNet, CS-Net, and nnUNet. Stride and kernel size have the same value for all dimensions.

Network	Input size (pixels)	Layers	Channels size	Stride	Kernel size	Activation	Sliding window	Residual encoder	Spatial attention
UNet	$128 \times 256 \times 256$	6	(16, 32, 64, 128, 256, 512)	2	3	PReLU	No	Yes	No
CS-Net	$128 \times 256 \times 256$	5	(16, 32, 64, 128, 256)	2	3	ReLU	No	Yes	Yes
nnUNet	$80 \times 224 \times 160$	6	(32, 64, 128, 256, 320, 320)	2	3	LeakyReLU	Yes	Yes	No

2.3.5. Uncertainty quantification with modified test-time augmentation

We estimated the uncertainty using test-time augmentation (TTA) [26]. TTA generates multiple slightly different inputs, each representing a plausible variation of the original data. If the labeling varies significantly across different augmented versions, it suggests that the model is less confident in its prediction for that particular region, likely due to inherent data variability arising from factors such as anatomical abnormalities, imaging artifacts, or ambiguities in ground truth annotation.

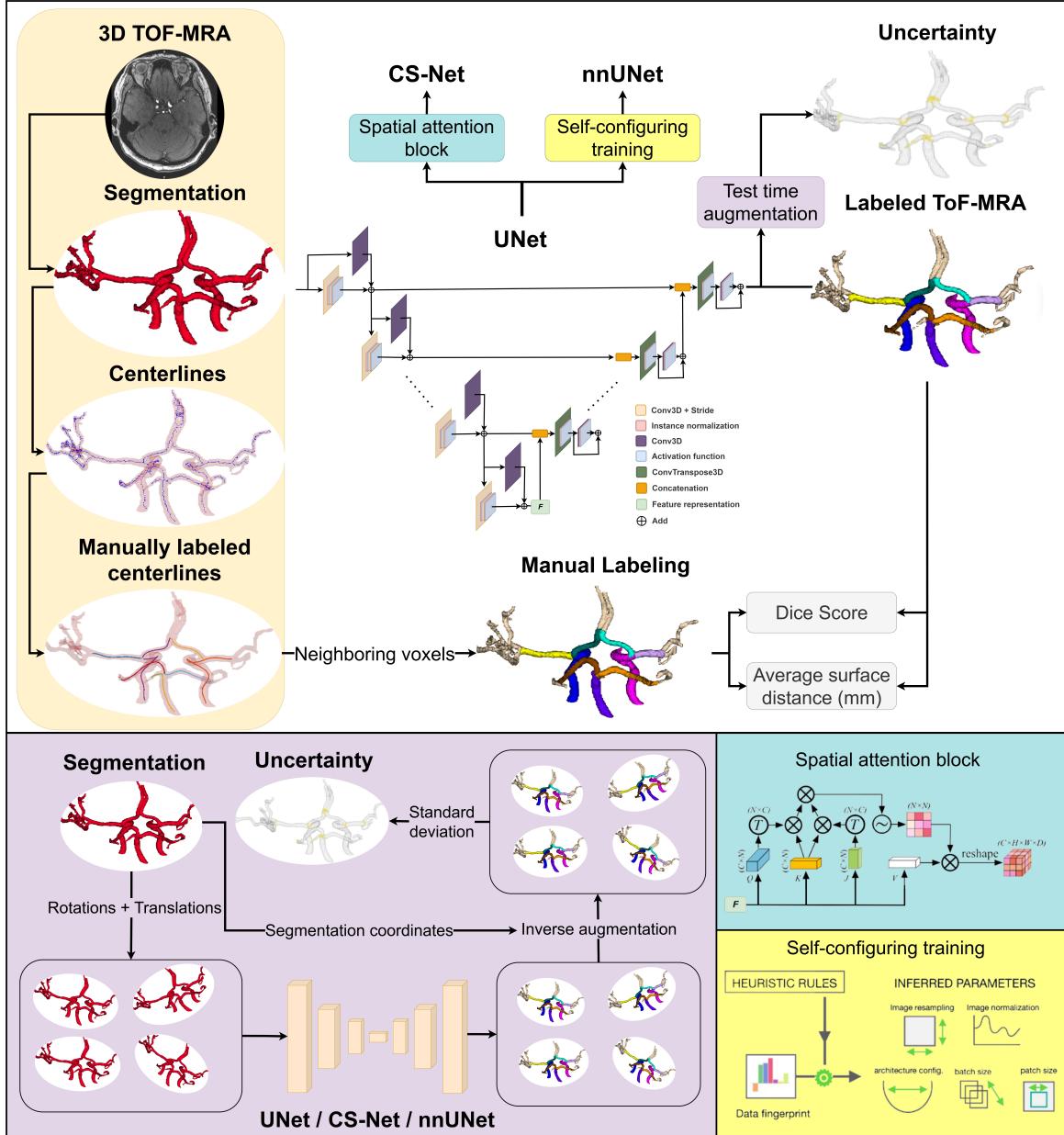


Figure 2: **Overview of the automated labeling framework.** 3D ToF-MRA images were used to segment intracranial arteries, followed by centerline extraction and manual labeling of nine arterial segments. Neighboring voxels of centerlines were labeled for voxel-wise classification. Three UNet variants (UNet, CS-Net, and nnUNet) were trained to perform voxel-wise classification directly from the segmentations. Test-time augmentation (TTA) was applied to estimate uncertainty. Labeling performance was evaluated using Dice Score and Average Surface Distance (ASD).

In our implementation, for each evaluation during inference, we applied random rotations (within $\pm 18^\circ$ per axis) and translations (within ± 5 voxels per axis), simulating variations in the positioning of the region of interest. For each acquisition, we generated seven predictions. Then, we inverted the transformations and computed the variance across predictions as a measure of uncertainty.

Although rotations and translations are inherently invertible, interpolations are not and can introduce errors, particularly at label boundaries, which may distort uncertainty estimates using TTA. To mitigate this, we developed a coordinate-guided transformation method. For each point in the original image, we defined X, Y, and Z grids, applied the TTA transformations to these grids, and then rounded the transformed coordinates to map them back to the original space. Labels were assigned based on the closest valid segmentation value, minimizing misassignments to the background. This approach significantly reduced interpolation errors, with only minor errors remaining at the label interfaces.

We performed an experiment to prove that our coordinate-guided strategy improves the estimates of TTA. More details can be found in [Appendix A](#).

An overview of the automated labeling framework including preprocessing, labeling, and uncertainty quantification is shown in [Figure 2](#).

2.4. Experimental setup

To estimate the accuracy and stability of the networks, we performed a 5-fold stratified cross-validation. For each iteration of cross-validation, 28 scans were used for training (80% of the data), and the remaining 7 scans were used for testing (20% of the data).

2.4.1. Network implementation

UNet and CS-Net were implemented using the MONAI open-source framework [30], while nnUNet was implemented using its public library¹³. All implementations were built on the PyTorch deep learning framework [31]. Training, testing, and uncertainty quantification scripts, as well as model weights, are publicly available¹⁴. Note that reproducing the training for nnUNet requires following the guidelines provided in [24].

UNet and CS-Net networks were trained using the Adam optimizer [32], while nnUNet was trained with stochastic gradient descent (SGD). A linear learning rate scheduler was used for all optimizers, with parameters set specifically for each optimizer. For Adam, the initial learning rate was set to 0.001 and the final learning rate to 0.0001, while for SGD, the initial learning rate was 0.01 and the final learning rate was 0.001.

UNet and CS-Net were trained for 6000 epochs with $\beta = 2000$ and $\gamma = 3500$, while nnUNet was trained for 1000 epochs with $\beta = 400$ and $\gamma = 600$. The number of epochs was chosen to ensure stable convergence in the validation curves. All networks were trained on NVIDIA A100-SXM4-40GB GPU. The total training times were 23:30 (HH:MM) for UNet, 27:30 for CS-Net, and 38:05 for nnUNet.

Specific architectural details, including convolutional layers, activation functions, and additional refinements, can be found in [Table 3](#).

2.4.2. Automated labeling evaluation

For evaluation, we considered the models at the last epoch of each training session. This approach prevents any bias toward better performing models in the test set for each cross-validation iteration.

Two metrics were used to evaluate the performance of the automated labeling: the Dice score (Dice) (Eq. 1) and the average surface distance (ASD), defined as:

$$\text{ASD} = \frac{1}{|S_X| + |S_Y|} \left(\sum_{x \in S_X} d(x, S_Y) + \sum_{y \in S_Y} d(y, S_X) \right). \quad (6)$$

Here, X and Y represent the sets of manually and automatically labeled regions, respectively. The surfaces of X and Y are denoted by S_X and S_Y , and $d(a, S)$ denotes the minimum Euclidean distance between a point a and the

¹³<https://pypi.org/project/nunetv2/>

¹⁴<https://github.com/JavierBZ/IC-UNet>

surface S . $|.|$ computes the area of each surface. The Dice score measures the overlap between two sets, X and Y , and ranges from 0 (no overlap) to 1 (perfect overlap). On the other hand, ASD quantifies the average distance between the surfaces of the two sets, providing a measure of the spatial agreement between them: lower ASD values indicate better alignment between the surfaces, and vice versa.

2.4.3. 4D Flow MRI velocity analysis

To investigate whether anatomical labels could guide the analysis of corresponding 4D Flow MRI datasets, we evaluated the agreement between velocity measurements derived from manual reference labels and those generated by the best-performing automated labeling network. Specifically, we assessed whether the average velocity at peak systole within anatomically labeled regions was consistent between manual and network-based labels.

To co-register ToF-MRA images with 4D Flow MRI data, we used a semi-automatic MATLAB tool [8] that performed rigid registration with built-in functions from the SPM12 toolbox (Statistical Parametric Mapping 12) [33]. Following co-registration, both manually and automatically labeled ToF regions were downsampled by a factor of two to match the resolution of the 4D Flow MRI data.

Vessel-specific average velocities at peak systole, obtained from manual and automated labeling, were then compared using Bland–Altman analysis and the Wilcoxon signed-rank test for statistical evaluation.

3. Results

3.1. Automated labeling

[Figure 3](#) presents the automated labeling results for the best and worst cases in terms of average Dice across all networks.

In the best-case scenario all networks achieved accurate labeling of the vessels of interest with only minor errors occurring primarily at vessel bifurcations or connections with non-annotated segments. These mislabeled regions are highly correlated with areas of higher uncertainty.

In contrast, for the worst-case scenario, three regions showed some problems:

- **RPCA misclassification:** UNet and CS-Net incorrectly labeled the RPCA segment, probably due to its anatomical similarity and proximity to a variant vessel branching from the RICA (a common hyperplastic posterior communicating artery variant [34]). While UNet and CS-Net failed to capture this uncertainty, nnUNet correctly classified this region as non-annotated with a relatively high uncertainty, recognizing it as an anatomical variant.
- **BA stenosis:** All networks struggled to label the BA segment in a patient with severe stenosis. UNet and CS-Net partially captured the uncertainty associated with this error, while nnUNet showed high uncertainty throughout the stenotic BA.
- **ACA variability:** All networks misclassified upper ACA segments due to inconsistencies in manual labeling protocols, particularly to delineate terminal portions of smaller vessels. These variations were consistently reflected in the uncertainty estimates across all networks.

[Table 4](#) shows the quantitative evaluation of the labeling performance. Consistently, strong performance is observed across all networks, with nnUNet achieving the highest overall accuracy. For Dice scores, nnUNet demonstrated superior performance (average 0.922) compared to CS-Net (0.907) and UNet (0.904), with notable improvements in the LMCA (0.939 vs. 0.911/0.909) and RACA (0.880 vs 0.836/0.828) segments. Although all networks excelled at labeling larger vessels, such as LICA and RICA (Dice > 0.975), smaller vessels, including the ACA and PCA segments, proved more challenging, showing lower agreement between methods.

The lowest surface distance (which provided specific information about the boundary detection capacity) was for nnUNet (ASD = 0.387 mm) compared to UNet (0.480 mm) and CS-Net (0.521 mm). Although all networks performed well on the ICA segments (<0.2 mm error), nnUNet showed particular advantages in the regions RMCA (0.379 mm vs 0.638/0.758 mm) and LPCA (0.495 mm vs 0.908/0.821 mm). The BA segment showed consistent performance across networks (0.24-0.30 mm), suggesting that all architectures handled this structure effectively.

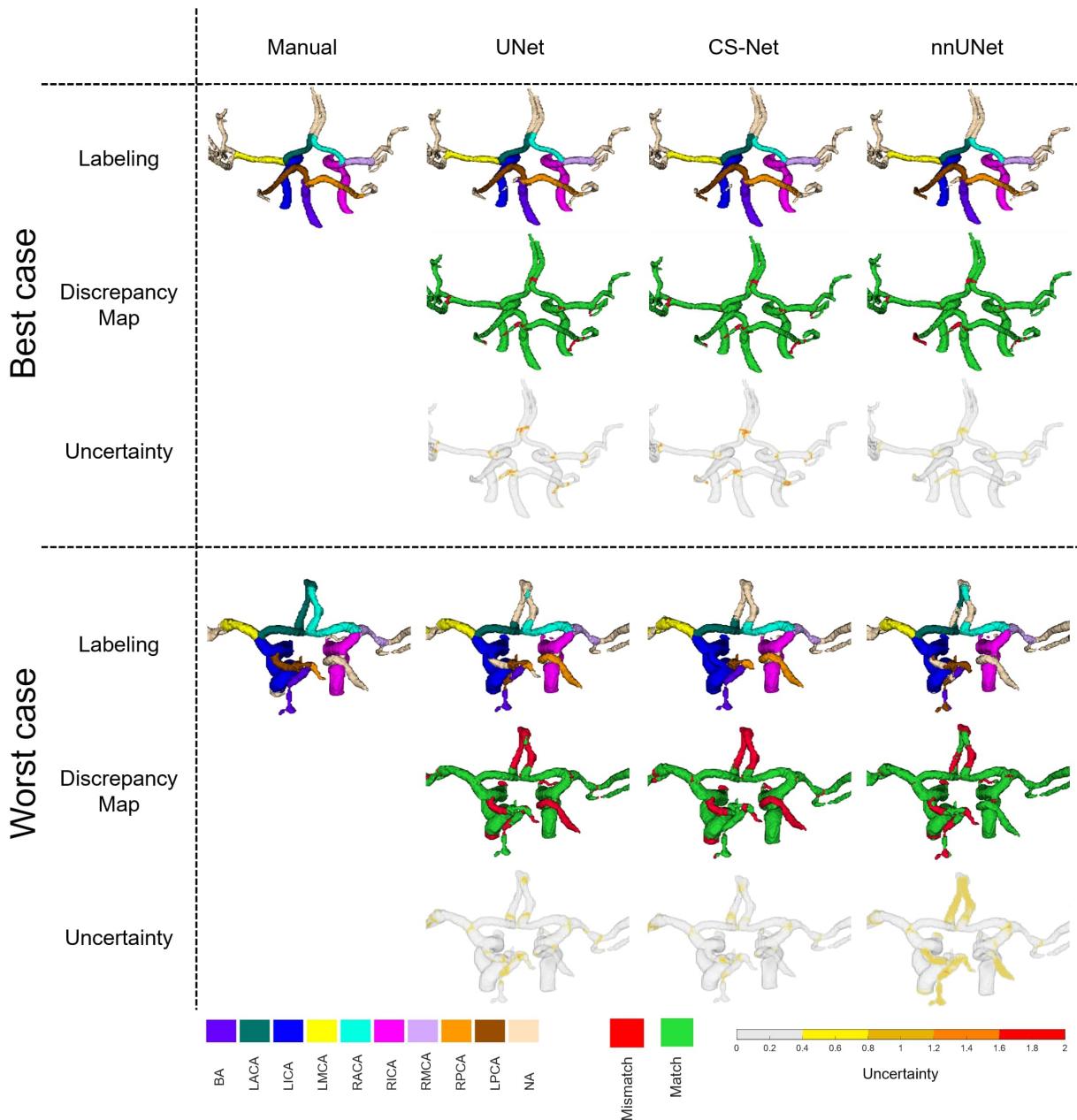


Figure 3: Automated labeling results for the best and worst-case patient data. For each case, the first row illustrates the manual labeling alongside the automated labeling predictions from UNet, CS-Net, and nnUNet. The second row highlights match and mismatch regions between the manual and automated labeling. The third row visualizes uncertainty, measured by the standard deviation of Test-Time Augmentation predictions.

Boxplots (Figure 4) illustrate differences between healthy volunteers and patients, as well as between networks. For both UNet and CS-Net, the median Dice scores and ASD showed no clear differences between the healthy and patient groups, although variability tended to be higher in patients, particularly in smaller or more complex vessels such as the RPCA and LPCA. In contrast, nnUNet tended to perform worse in patients, exhibiting higher variability and worse median values for both Dice scores and ASD. Despite this, nnUNet was the most robust model overall,

Table 4: Performance metrics of UNet, CS-Net, and nnUNet. The table presents the average Dice Score and Average Surface Distance values for different cerebrovascular segments across the three networks. Color intensity represents performance levels: lighter and darker colors indicate lower and higher performance, respectively.

	Dice Score									
	BA	LACA	LICA	LMCA	RACA	RICA	RMCA	RPCA	LPCA	Average
UNet	0.936	0.848	0.985	0.911	0.836	0.976	0.865	0.901	0.877	0.904
CS-Net	0.939	0.858	0.988	0.909	0.828	0.978	0.871	0.895	0.890	0.907
nnUNet	0.936	0.861	0.988	0.939	0.880	0.980	0.883	0.923	0.910	0.922

	Average Surface Distance (mm)									
	BA	LACA	LICA	LMCA	RACA	RICA	RMCA	RPCA	LPCA	Average
UNet	0.282	0.630	0.194	0.394	0.589	0.153	0.638	0.540	0.908	0.480
CS-Net	0.298	0.624	0.067	0.254	0.571	0.115	0.758	1.226	0.821	0.521
nnUNet	0.242	0.756	0.044	0.223	0.581	0.112	0.379	0.679	0.495	0.387

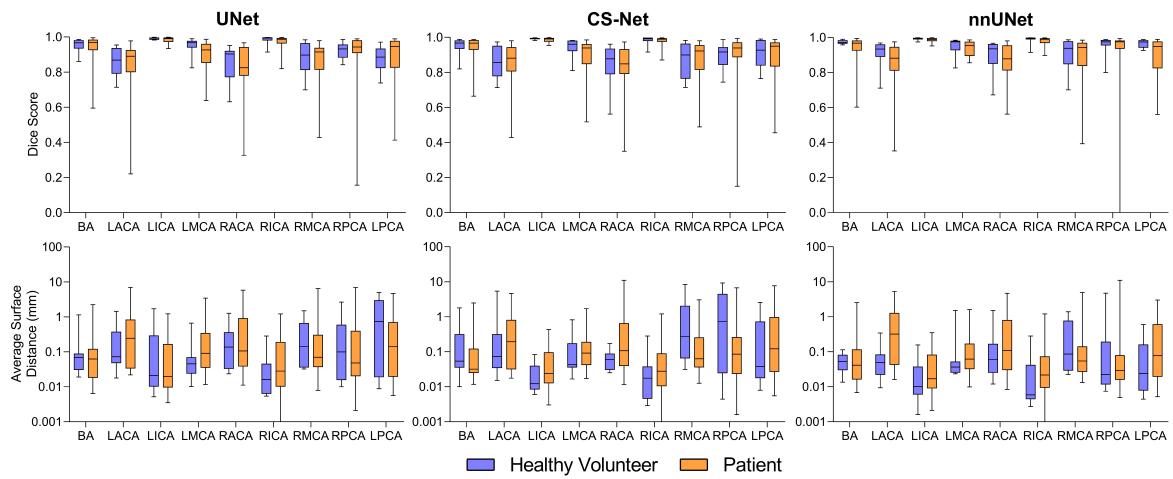


Figure 4: Boxplot analysis of Dice Score and Average Surface Distance distributions across networks comparing labeling performance between healthy volunteers (n=10) and patients (n=25). Whiskers represent the full data range (minimum to maximum values).

demonstrating the least variability across both groups.

3.2. 4D Flow MRI analysis

Figure 5 presents Bland-Altman plots for the 4D Flow MRI analysis, comparing the average velocity at the peak systole timeframe between manual and automated labeling using the best-performing network (nnUNet). ICAs demonstrated the smallest differences between manual and automatically labeled vessels, with agreement limits ranging from -0.34 cm/s to 0.43 cm/s (-2.1 % of the mean to 2.4 % of the mean, respectively). The BA also exhibited relatively small differences, although a few outliers widened its limits of agreement from -2.70 cm/s to 2.65 cm/s (-12.4 % of the mean to 12.8 % of the mean). Larger discrepancies were also observed in the smaller arterial groups including MCAs (limits of agreement: -3.23 cm/s to 3.25 cm/s, equivalent to -10.7 % to 9.4 % of the mean), PCAs (limits of agreement: -2.44 cm/s to 2.38 cm/s; equivalent to -15.5 % to 14.1 % of the mean), and ACAs (limits of agreement: -4.53 cm/s to 5.00 cm/s; equivalent to -15.9 % to 17.6 % of the mean).

Table 5 presents Bland-Altman individual limits of agreement and average differences for all intracranial segments. Overall, the average differences were less than 2 cm/s for all segments, with 7 out of 9 segments showing an average absolute difference below 1 cm/s. When expressed as a percentage of the mean velocity, the average absolute

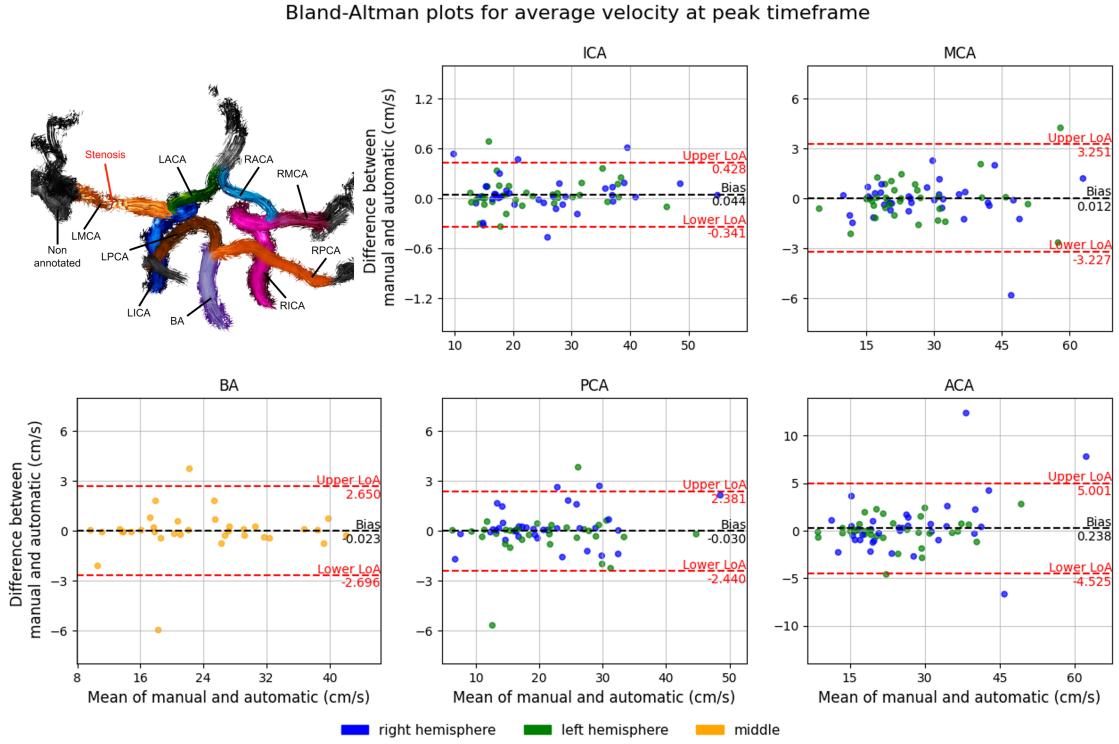


Figure 5: The upper left corner displays streamlines of different artery segments using automated labeling in a patient with severe stenosis in the left middle cerebral artery (LMCA). In the remaining figures, Bland-Altman plots compare the percentage differences in average velocity at peak systole between manual and automated labeling using nnUNet. Bias and limits of agreement are shown for each artery group: internal, middle, posterior, and anterior cerebral arteries (ICA, MCA, PCA, and ACA, respectively).

Table 5: Width of limits of Agreement (LoA), average absolute differences, and p-values using Wilcoxon signed-rank test for average velocity differences between manual and automated labeling.

Vessel	Average absolute difference (cm/s)	Average absolute difference (% of mean)	Width of limits of agreement (cm/s)	Width of limits of agreement (% of mean)	p-value
BA	0.585	2.490	4.863	20.704	0.259
LACA	1.693	6.306	12.065	44.931	0.942
LICA	0.155	0.605	0.805	3.142	0.056
LMCA	0.799	2.833	5.712	20.259	0.704
RACA	1.005	4.213	5.789	24.267	0.919
RICA	0.125	0.540	0.752	3.243	0.369
RMCA	1.000	3.711	6.209	23.043	0.968
RPCA	0.621	3.028	7.096	34.598	0.447
LPCA	0.591	2.828	4.041	19.342	0.158

differences ranged 0.54% (RICA) to 6.31% (LACA). Statistical analysis revealed no significant differences between manual and automatically labeled vessels for any of the derived average velocities.

4. Discussion

4.1. Highlights

This study investigated the application of deep learning, specifically UNet-based architectures, for automated anatomical labeling of the major intracranial arteries from 3D ToF-MRA data. Networks also incorporate uncertainty quantification for subsequent 4D Flow MRI analysis. Our findings demonstrate that, while architectural refinements such as channel and spatial attention blocks in CS-Net offered improved performance compared to the baseline UNet, the self-configuring nnUNet framework yielded the most accurate and robust labeling results. Furthermore, we introduced a novel coordinate-guided test-time augmentation (TTA) strategy for estimating uncertainty, which effectively highlighted regions prone to labeling errors. Validation through 4D Flow MRI velocity analysis confirmed that the automated labels generated by the best performing network (nnUNet) provide a reliable basis for subsequent hemodynamic assessment.

During the drafting and revision of our manuscript, Colombo et al. (2025) [23] published a nnUNet-based pipeline that simultaneously segments intracranial aneurysms and their parent vessels in 3D ToF-MRA. Their model achieved a median Dice score of 0.86 in 21 classes and an aneurysm detection sensitivity of 0.80. These results corroborate our choice of nnUNet as a strong baseline for cerebrovascular applications. However, their work did not compare to other UNet baselines or assess uncertainty quantification. By systematically benchmarking attention-based architectures and introducing coordinate-guided test-time augmentation, our study extends the nnUNet paradigm toward greater anatomical consistency and interpretability.

4.2. Uncertainty quantification and clinical utility

Beyond accuracy, the integration of uncertainty quantification is crucial for the clinical translation of deep learning models. Our implementation of TTA successfully captured regions where the model predictions were less confident. High uncertainty arises when an input image differs substantially from the types of image to which the model was exposed during training, for example, due to anatomical variations or disease-related changes. In these cases, applying geometric transformations, such as rotations or translations during TTA, results in highly variable predictions. This occurs because the models are trained to be invariant to such transformations only within the training data distribution. When the input deviates from that distribution, their predictions become less stable.

As illustrated in [Figure 3](#), regions with higher uncertainty often corresponded to anatomical variations (e.g. hyperplastic posterior communicating artery variant affecting RPCA labeling), vascular alterations (e.g., BA stenosis), or inconsistencies arising from the manual labeling protocol itself (e.g., defining distal ACA segments). This provides clinicians with valuable information on the trustworthiness of automated labels in specific regions.

Moreover, our proposed coordinate-guided TTA strategy addresses interpolation errors during the inverse transformation step. By mapping transformed coordinates back to the original grid space, we minimized these artifacts, reducing label misassignments from near 5% to <0.1% compared to standard affine inversion and obtained cleaner uncertainty maps at the vessel boundaries (see [Appendix A](#)).

4.3. Labeling considerations

For nnUNet, we also highlighted performance differences between healthy volunteers and patients with ICAD ([Figure 4](#)). For patients, nnUNet tended to obtain lower labeling performance and greater variability, particularly in smaller or more complex vessels. This is likely to be attributable to the increased anatomical complexity and variability introduced by cerebrovascular disease, such as stenoses or altered vessel morphologies, which pose a greater challenge to automated methods.

In contrast to previous approaches that operate directly on image intensity data, our pipeline performs anatomical labeling using pre-segmented vascular structures. This design choice reflects a deliberate focus on the spatial and geometric features of the vasculature, which are more relevant for accurate anatomical identification. Intensity-based features in ToF-MRA are often susceptible to artifacts, such as flow-related signal loss, inhomogeneities, or non-vascular enhancements, which can mislead the network and reduce its generalizability across diverse datasets [35]. By decoupling the labeling task from the raw image intensities and instead using binary segmentations as input, the network is encouraged to learn from the morphology and topology of the vessels themselves. Furthermore, excluding intensity features helps prevent the propagation of undesired biases from intensity features in the uncertainty estimates, resulting in more meaningful and spatially grounded uncertainty maps.

4.4. Validation for downstream 4D Flow MRI analysis

The successful application of automated labels for 4D Flow MRI velocity analysis ([Figure 5](#), [Table 5](#)) demonstrates the practical utility of our approach. Bland-Altman analysis revealed good agreement between velocity measurements derived from manual labels and those from nnUNet's automated labels, with no statistically significant differences found for any vessel segment using the Wilcoxon signed-rank test. Although agreement was excellent for larger vessels such as the ICAs, slightly larger discrepancies were noted for smaller arteries (ACAs, PCAs, MCAs). This may reflect the slightly lower labeling accuracy in these segments, potential partial-volume effects, or minor residual co-registration inaccuracies. Nonetheless, the overall small average differences (<2 cm/s absolute difference for all segments, <1 cm/s for 7 out of 9 segments) suggest that the labeling of nnUNet provides a robust foundation for automated hemodynamic quantification, reducing the need for time-consuming manual delineation.

4.5. Limitations

Despite promising results, our study has some limitations. First, our method relies on an initial semi-automated segmentation step; any vessels missed or incorrectly segmented in this initial mask are not correctly labeled by the network. Errors in the initial segmentation will inevitably propagate to the final labeling. Second, the validation involving 4D Flow MRI is subject to potential co-registration errors between the ToF-MRA (used for labeling) and the 4D Flow MRI datasets, which could influence the velocity comparison. Third, the number of datasets (n=35) is relatively limited, covering healthy volunteers and ICAD patients from a single vendor. Although a 5-fold cross-validation provides a measure of robustness, evaluation on larger, more diverse datasets is needed to fully assess generalizability. However, large amounts of intracranial 4D Flow MRI datasets are rare, thus our study represents an already acceptable cohort.

4.6. Future directions

Several promising directions can be pursued to extend our work. Building on previous studies that investigated architectural modifications within the nnUNet framework [[36](#)], future research could explore the integration of the attention mechanisms used in CS-Net, into the self-configuring nnUNet pipeline. This approach may further improve performance by combining the robust automated pipeline of nnUNet with architectural innovations. The resulting vessel labels could also serve as reliable inputs for downstream tasks, including automated centerline labeling, contributing to a fully automated pipeline for cerebrovascular hemodynamic analysis. Furthermore, applying transfer learning to adapt the labeling model directly to 4D Flow phase-contrast MRA (PC-MRA) data could eliminate the need for co-registration and potentially improve the accuracy of velocity quantification within automatically labeled vascular regions. In parallel, future validation efforts should include data from multiple scanner vendors and clinical sites to ensure broader applicability and to understand how scanner-specific differences may impact model generalization.

5. Acknowledgments

J.B. was funded by the National Agency for Research and Development (ANID) / Scholarship Program / DOC-TORADO BECAS CHILE/2022 – 21220454, and by ANID - Millennium Science Initiative Program - ICN2021_004. JS thanks to ANID Millennium Science Initiative Program ICN2021_004 and Department of Medical Imaging and Radiation Sciences at Monash University. C.T. thanks Fondecyt 1231535 and ANID - Millennium Science Initiative Program - ICN2021_004. S.S and P.W thanks to the National Institute of Health under R01HL115828 and R21NS106696.

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the author(s) used GPT-5 in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] J. D. Hinman, N. S. Rost, T. W. Leung, J. Montaner, K. W. Muir, S. Brown, J. F. Arenillas, E. Feldmann, D. S. Liebeskind, Principles of precision medicine in stroke, *Journal of Neurology, Neurosurgery & Psychiatry* 88 (1) (2017) 54–61.
- [2] A. G. Morgan, M. J. Thrippleton, J. M. Wardlaw, I. Marshall, 4d flow mri for non-invasive measurement of blood flow in the brain: a systematic review, *Journal of Cerebral Blood Flow & Metabolism* 41 (2) (2021) 206–218.
- [3] T. A. Hope, M. D. Hope, D. D. Purcell, C. von Morze, D. B. Vigneron, M. T. Alley, W. P. Dillon, Evaluation of intracranial stenoses and aneurysms with accelerated 4d flow, *Magnetic resonance imaging* 28 (1) (2010) 41–46.
- [4] S. Schnell, S. A. Ansari, P. Vakil, M. Wasielewski, M. L. Carr, M. C. Hurley, B. R. Bendok, H. Batjer, T. J. Carroll, J. Carr, et al., Three-dimensional hemodynamics in intracranial aneurysms: influence of size and morphology, *Journal of Magnetic Resonance Imaging* 39 (1) (2014) 120–131.
- [5] J. Liu, L. Koskas, F. Faraji, E. Kao, Y. Wang, H. Haraldsson, S. Kefayati, C. Zhu, S. Ahn, G. Laub, et al., Highly accelerated intracranial 4d flow mri: evaluation of healthy volunteers and patients with intracranial aneurysms, *Magnetic Resonance Materials in Physics, Biology and Medicine* 31 (2) (2018) 295–307.
- [6] M. Hope, D. Purcell, T. Hope, C. Von Morze, D. Vigneron, M. Alley, W. Dillon, Complete intracranial arterial and venous blood flow evaluation with 4d flow mr imaging, *American journal of neuroradiology* 30 (2) (2009) 362–366.
- [7] M. Aristova, A. Vali, S. A. Ansari, A. Shaibani, T. D. Alden, M. C. Hurley, B. S. Jahromi, M. B. Potts, M. Markl, S. Schnell, Standardized evaluation of cerebral arteriovenous malformations using flow distribution network graphs and dual-venc 4d flow mri, *Journal of Magnetic Resonance Imaging* 50 (6) (2019) 1718–1730.
- [8] A. Vali, M. Aristova, P. Vakil, R. Abdalla, S. Prabhakaran, M. Markl, S. A. Ansari, S. Schnell, Semi-automated analysis of 4d flow mri to assess the hemodynamic impact of intracranial atherosclerotic disease, *Magnetic resonance in medicine* 82 (2) (2019) 749–762.
- [9] A. El Ahmar, S. Schnell, S. A. Ansari, R. N. Abdalla, A. Vali, M. Aristova, M. Markl, P. Winter, D. Marlevi, Non-invasive quantification of pressure drops in stenotic intracranial vessels: using deep learning-enhanced 4d flow mri to characterize the regional haemodynamics of the pulsing brain, *Interface Focus* 15 (1) (2025) 20240040.
- [10] B. J. Alpers, R. G. Berry, R. M. Paddison, Anatomical studies of the circle of willis in normal brain, *AMA Archives of Neurology & Psychiatry* 81 (4) (1959) 409–418.
- [11] L. Chen, M. Mossa-Basha, J. Sun, D. S. Hippe, N. Balu, Q. Yuan, K. Pimentel, T. S. Hatsukami, J.-N. Hwang, C. Yuan, Quantification of morphometry and intensity features of intracranial arteries from 3d tof mra using the intracranial artery feature extraction (icafe): a reproducibility study, *Magnetic resonance imaging* 57 (2019) 293–302.
- [12] P. Winter, H. Berhane, J. E. Moore, M. Aristova, T. Reichl, J. Wollenberg, A. Richter, K. B. Jarvis, A. Patel, F. Z. Caprio, et al., Automated intracranial vessel segmentation of 4d flow mri data in patients with atherosclerotic stenosis using a convolutional neural network, *Frontiers in Radiology* 4 (2024) 1385424.
- [13] L. Mou, Y. Zhao, H. Fu, Y. Liu, J. Cheng, Y. Zheng, P. Su, J. Yang, L. Chen, A. F. Frangi, et al., Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging, *Medical image analysis* 67 (2021) 101874.
- [14] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, P.-A. Heng, 3d deeply supervised network for automated segmentation of volumetric medical images, *Medical image analysis* 41 (2017) 40–54.
- [15] H. Chen, Q. Dou, L. Yu, J. Qin, P.-A. Heng, Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images, *NeuroImage* 170 (2018) 446–455.
- [16] J. Sobisch, Ž. Bizjak, A. Chien, Ž. Špiclin, Automated intracranial vessel labeling with learning boosted by vessel connectivity, radii and spatial context, in: *Geometric Deep Learning in Medical Image Analysis*, PMLR, 2022, pp. 34–44.
- [17] L. Chen, T. Hatsukami, J.-N. Hwang, C. Yuan, Automated intracranial artery labeling using a graph neural network and hierarchical refinement, in: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, Springer, 2020, pp. 76–85.
- [18] Y. Zhu, P. Qian, Z. Zhao, Z. Zeng, Deep feature fusion via graph convolutional network for intracranial artery labeling, in: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 467–470.
- [19] A. Hilbert, J. Rieger, V. I. Madai, E. M. Akay, O. U. Aydin, J. Behland, A. A. Khalil, I. Galinovic, J. Sobesky, J. Fiebach, et al., Anatomical labeling of intracranial arteries with deep learning in patients with cerebrovascular disease, *Frontiers in Neurology* 13 (2022) 1000914.
- [20] Y. Lv, W. Liao, W. Liu, Z. Chen, X. Li, A deep-learning-based framework for automatic segmentation and labelling of intracranial artery, in: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2023, pp. 1–5.
- [21] F. Dumais, M. P. Caceres, F. Janelle, K. Seifeldine, N. Arès-Brunneau, J. Gutierrez, C. Bocti, K. Whittingstall, eicab: A novel deep learning pipeline for circle of willis multiclass segmentation and analysis, *Neuroimage* 260 (2022) 119425.
- [22] T. Chen, W. You, L. Zhang, W. Ye, J. Feng, J. Lu, J. Lv, Y. Tang, D. Wei, S. Gui, et al., Automated anatomical labeling of the intracranial arteries via deep learning in computed tomography angiography, *Frontiers in Physiology* 14 (2024) 1310357.
- [23] E. Colombo, M. de Boer, L. Bartels, L. Regli, T. van Doornmaal, Accuracy of an nnunet neural network for the automatic segmentation of intracranial aneurysms, their parent vessels, and major cerebral arteries from mri-tof, *American Journal of Neuroradiology* 46 (5) (2025) 956–963.
- [24] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature methods* 18 (2) (2021) 203–211.
- [25] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, P. F. Jaeger, nnu-net revisited: A call for rigorous validation in 3d medical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 488–498.
- [26] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, *Neurocomputing* 338 (2019) 34–45.
- [27] S. Schnell, S. A. Ansari, C. Wu, J. Garcia, I. G. Murphy, O. A. Rahman, A. A. Rahsepar, M. Aristova, J. D. Collins, J. C. Carr, et al., Accelerated dual-venc 4d flow mri for neurovascular applications, *Journal of Magnetic Resonance Imaging* 46 (1) (2017) 102–114.

- [28] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, J. A. Schnabel, Left-ventricle quantification using residual u-net, in: Statistical Atlases and Computational Models of the Heart, Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9, Springer, 2019, pp. 371–380.
- [29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [30] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al., Monai: An open-source framework for deep learning in healthcare, arXiv preprint arXiv:2211.02701 (2022).
- [31] A. Paszke, Pytorch: An imperative style, high-performance deep learning library, arXiv preprint arXiv:1912.01703 (2019).
- [32] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [33] J. Ashburner, K. J. Friston, Unified segmentation, neuroimage 26 (3) (2005) 839–851.
- [34] M. L. Rajan, Study of variations in the posterior part of the circle of willis using magnetic resonance angiography, Indian Journal of Clinical Anatomy and Physiology 8 (1) (2023) 11–14.
- [35] D. Wilcock, T. Jaspan, B. Worthington, Problems and pitfalls of 3-d tof magnetic resonance angiography of the intracranial circulation, Clinical radiology 50 (8) (1995) 526–532.
- [36] N. McConnell, N. Ndipenoch, Y. Cao, A. Miron, Y. Li, Exploring advanced architectural variations of nnunet, Neurocomputing 560 (2023) 126837.

Appendix A. Appendix A: Label Inversion and Uncertainty Quantification

To validate our coordinate-guided transformation method, we conducted an experiment using manual labels from one acquisition. We applied 100 random transformations (rotations: $\pm 18^\circ$ per axis; translations: ± 5 voxels per axis), then evaluated two key aspects.

First, we assessed the accuracy of label inversion by comparing standard inversion (using inverse affine matrices with nearest-neighbor interpolation) against our proposed coordinate-guided method. The proposed approach significantly reduced the inversion errors to $0.09\% \pm 0.04\%$ (mean \pm standard deviation) of misassigned voxels versus $5.10\% \pm 1.17\%$ of the standard inversion. Boxplots in [Figure A.1A](#) demonstrate that our method maintains the misassignment below 0.2%, while the standard inversion reaches up to 7%. The first row of [Figure A.1B](#) shows the location and amount of missassigned voxels. The second row of [Figure A.1B](#) shows the spatial distribution of these errors: standard inversion produces missassigned voxels scattered across borders and label interfaces, while our approach yields only minimal errors at label interfaces.

Second, we examined the effects on uncertainty quantification using the nnUNet labeling network. For seven augmented versions of the acquisition, we computed prediction uncertainty maps using both inversion methods. [Figure A.1](#) (second row) shows that our method produced more reliable uncertainty estimates, particularly at segmentation boundaries, where standard interpolation introduces artifacts. Although both methods can identify high-uncertainty regions, standard inversion fails to distinguish low-uncertainty areas due to interpolation artifacts.

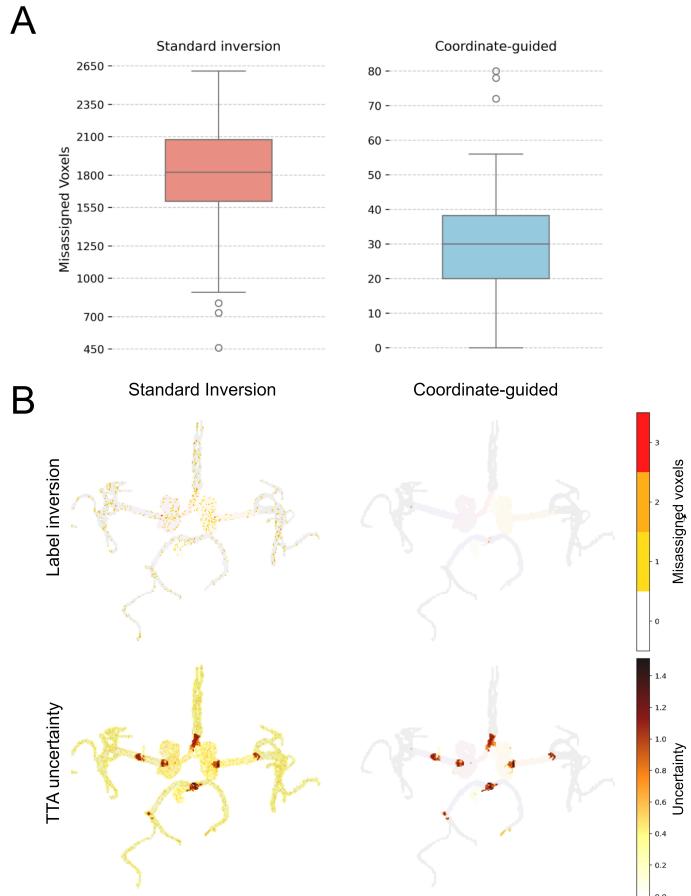


Figure A.1: **A:** Distribution of label inversion errors across 100 random transformations. Left boxplot shows errors using standard affine inversion (median 5.18%), while right boxplot demonstrates improved performance of our coordinate-guided method (median 0.09%). **B:** Comparison of inversion methods for one random rotation and translation. First row: Label inversion results with standard inversion (left) and the coordinate-guided approach (right). The sum of misassigned voxels along the slice direction is overlaid on the maximum intensity projection of the labeled image. Second row: Uncertainty maps using standard inversion (left) and the coordinate-guided approach (right), shown as maximum intensity projections along the slice direction.