Towards improved decision making of unruptured intracranial aneurysms using automated segmentation from MRA-TOF with iterative pseudolabeling

Anouk S. Verschuur[1]

Jiaxin Zhang[1]

Maarten J. Kamphuis[1]

Chantal M.W. Tax[1,2,*]

Irene C. van der Schaaf[1,*]

*Shared last author

Affiliations

[1]Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands.

[2]CUBRIC, School of Physics and Astronomy, Cardiff University, CF10 3AT, Cardiff, United Kingdom.

Corresponding author

A.S. Verschuur, MSc

University Medical Center Utrecht

Division of imaging and oncology

Heidelberglaan 100

Tel: +31(0)887553172

E-mail: a.s.verschuur@alumnus.utwente.nl

Funding

This research was funded by TKI-LSH Health Holland (2021AI-TKI2203). CMWT is supported by a Vidi grant (21299) from the Dutch Research Council (NWO) and a Sir Henry Wellcome Fellowship (215944/Z/19/Z). The sponsors had no role in the design and conduct of the study,

the collection, management, analysis and interpretation of the data, the preparation, review and approval of the manuscript, or the decision to submit the manuscript for publication. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Conflicts of interest

None to declare

Abstract

Objective: To enable accurate 3D morphological assessment and support clinical decision making, DIVA-seg: a Deep learning-based method for Intracranial Vessel and Aneurysm segmentation from MRA-TOF using a pseudo labeling approach was developed and validated.

Methods: Three MRA-TOF datasets were used: 1) labeled data for training (n=57) and testing (n=14), 2) unlabeled data for pseudo labels (n=518), and 3) labeled data for external validation (n=82). An nnU-Net (Model 1) was iteratively trained for creating pseudo labels for Dataset 2. Cases with stable segmentation performance across iterations were selected for further training. Stable cases (n=484) were combined with Dataset 1 to train a second nnU-Net (Model 2). Performance testing on Dataset 1 and 3 comprised of Dice similarity coefficients (DSC), 95%-Hausdorff distances, 3D morphological measures, and a blinded qualitative evaluation.

Results: DIVA-seg achieved a mean (standard deviation) internal vessel and aneurysm DSC of 0.925 (±0.025) and 0.880 (±0.045), respectively. On the external test set the DSC were 0.899 (±0.028) and 0.861 (±0.114), respectively. Mean Hausdorff distances were 0.67mm for both test sets. Bland-Altman plots showed a high agreement between 3D morphological measures from ground truth and model segmentations; however, a proportional bias was observed for voxel volume, surface area, sphericity and shape index. The qualitative evaluation showed no clear preference for either ground truth or model segmentation.

Conclusion: The model achieved accurate and reliable segmentation of vessels and aneurysms internally and externally while also showing high agreement between 3D morphological measures from automatic and manual segmentations, indicating its potential clinical utility.

Key words: MRA-TOF, unruptured intracranial aneurysm, deep learning, artificial intelligence, segmentation

Summary section

Accurate intracranial aneurysm assessment is essential for treatment planning and risk stratification. Manual aneurysm segmentation is labor-intensive and subject to substantial inter- and intra-observer variability. Although automated segmentation approaches have been proposed, many suffer from limited accuracy, lack of robustness across datasets, or insufficient validation on heterogeneous, real-world data. As a result, reliable and generalizable tools for aneurysm segmentation and morphological analysis remain an unmet need. DIVA-seg, an nnU-Net-based model, achieved high aneurysm segmentation accuracy (DSC >0.86; HD <0.7mm) and close agreement with expert annotations in clinically relevant 3D morphological measures, demonstrating consistent performance across internal and external datasets. This work demonstrates a robust and generalizable approach for automated intracranial aneurysm segmentation, enabling reliable morphological analysis. The proposed method has the potential to streamline aneurysm monitoring, reduce observer variability, and support future automated tools for risk predictions and clinical decision making.

Introduction

It is estimated that the prevalence of unruptured intracranial aneurysms (UIAs) in the adult population is 3% [1]. While most UIAs remain asymptomatic, rupture leads to subarachnoid hemorrhage (SAH), which carries a high case fatality rate ranging from 13% to 71% [2]. Preventative treatment options such as endovascular coiling or neurosurgical clipping can reduce the risk of rupture; however, the decision to intervene must carefully balance the rupture risk against potential complications associated with treatment. Rupture risk prediction is therefore of importance in clinical decision making in UIA patients. Key determinants in current prediction models for rupture risk are hypertension, age, size, shape and location of the UIA [3,4]. Growth and shape changes are also important in clinical decision-making, as both have been linked to rupture risk [5–8].

Currently, aneurysm size and growth are generally measured manually in 2 dimensions (2D). Recent studies, however, suggest that 3D morphological measurements are more reproducible and better predictors of aneurysm rupture risk than 2D measurements [6,9]. Yet, obtaining these 3D measurements requires manual segmentation of aneurysms, which is time-consuming and prone to human error. Robust and reproducible segmentations are essential to ensure stable measurements and therewith accurate change detection and risk prediction in UIAs.

Automatic segmentation methods may provide faster, and more consistent results compared to manual approaches. Nonetheless, aneurysm segmentation remains a challenging task [10]. While several studies have applied deep learning techniques for aneurysm detection and/or segmentation, most primarily focused on detection, reporting sensitivity and specificity metrics [10]. In studies evaluating segmentation accuracy, reported Dice similarity coefficients (DSC) were moderate (<0.75) [11–15] and several models were

trained on relatively small sample sizes [11–17]. In addition, external testing was not always included, and where performed, generalizability across datasets remained limited [10].

This study addresses these shortcomings by developing and externally validating DIVA-seg: Deep learning for Intracranial Vessel and Aneurysm segmentation, an nnU-Net model for automated segmentation of intracranial vessels and UIAs on magnetic resonance angiography time-of-flight (MRA-TOF) data. Our approach combines high-quality manually corrected annotations with a large and diverse pseudo labeled dataset, to overcome the limitations of small training cohorts and improve generalizability. We further provide extensive internal and external validation and, beyond conventional segmentation metrics, evaluate clinically relevant 3D morphological parameters. Together, these contributions provide a strong foundation for future clinical applications, including aneurysm change detection and rupture risk prediction.

Methods

This study was performed following and in accordance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM).

Datasets

This study used MRA-TOF scans from three datasets 1) for training and testing [18,19], 2) for creating pseudo labels, and 3) for external validation [20] of the vessel and aneurysm segmentation model. A complete explanation of the used datasets is provided in the supplemental material. The inclusion process is illustrated in a flowchart (Figure 1), acquisition and preprocessing settings are provided (Table 1). This study retrospectively utilized existing data obtained from routine patient care. Local approval was obtained for conducting this study. Informed consent was not required, as the data were collected for clinical purposes. All patient records were reviewed for objections to the use of clinical data for research; patients who had registered objections were excluded from this study.

Model and training

Model architecture

All models described in this study were trained using the nnU-NetV2 architecture (download November 18, 2024), with a 3D full resolution configuration and default settings, unless otherwise specified [21]. Inspired by the winning approach from Huang et al. [22] in the MICCAI FLARE22 challenge, we adopted the concept of iteratively training a model for generating pseudo labels and training a new model with this supplemented dataset (Figure 2). Model 1 was iteratively trained, starting with training and validation data from Dataset 1 for the first iteration, followed by Dataset 1 and 2 in the subsequent three iterations. Pseudo

labels for training were each time generated with the preceding iteration. After four iterations, the scans with the most stable pseudo labels over the four iterations were selected and combined with Dataset 1 for training Model 2.

Training configuration

Models were trained on NVIDIA RTX A5000 GPU servers using the 3D full resolution configuration and a 5-fold cross-validation. Batch size was automatically set at 2, patch size was dynamic; the first iteration of Model 1 used 64x224x160, all other models used patch size 56x224x192. 1000 epochs were used for training. Standard augmentations from nnU-Net were used, including rotation, scaling, Gaussian noise, Gaussian blur, brightness and contrast changes, resampling to lower resolution, gamma correction, mirroring and elastic deformations. The segmentation model DIVA-seg is publicly available at https://github.com/Verschuur95/DIVA-seg.

Pseudo labeling workflow

Iterative pseudo labels were created by running inference on preceding models using all five folds; nnU-Net ensembles probability maps from different folds using softmax averaging to achieve final segmentation. After four iterations, the most stable cases were selected by calculating uncertainty measures using the formula given in Huang et al. [22]:

$$u = \frac{1}{K-1}\sum_{i=3}^{K}\frac{\sum(y_i \neq y_{i-1})}{\sum(y_i > 0)}$$

Where u is the uncertainty measure, K the total number of iterations and y_i the binary voxel value from the pseudo label generated at iteration i. To determine an appropriate uncertainty threshold, uncertainty values were sorted and binned (bin size = 0.005) and visually inspected

for discontinuities using a scatter plot (Supplemental Figure 1). Three candidate thresholds, corresponding to observed sharp increase in bin number in the scatter plot, were initially selected and statistically tested to assess their effect on model performance (See Supplemental Table 1). For each threshold, images with uncertainty values exceeding the threshold were manually reviewed to confirm the validity of the exclusion criterion. Based on this process, a final threshold of 0.1 was selected. All images from Dataset 1 exceeding this threshold, caused by a slowly declining segmentation quality with each iteration, were still included in the training of Model 2, as these images had manually created segmentations that were quality-checked and deemed reliable.

Evaluation and analysis

Performance metrics

The test set from Dataset 1 (with automatic and manual center of mass separately) and Dataset 3 were used to assess model performance. Overlap between ground truth and model segmentations was evaluated by calculating the DSC (overall, vessel, and aneurysm scores) and the 95%-Hausdorff distance (95%-HD). The relation between those measures was visualized in a scatter plot. The outermost cases with highest and lowest scores were visually checked to assess segmentation quality of the aneurysm.

3D morphological measures

To add clinical value beyond standard performance measures such as DSC and HD, 3D morphological measures of the aneurysms, as previously described by Timmins et al. [23] and Kamphuis et al. [24] were calculated using PyRadiomics in Python (version 3.7) [25]. Included morphological measures were in compliance with the image biomarker standardization

initiative guidelines [26]: voxel volume, surface area, major axis length, sphericity, curvedness, shape index, elongation, and flatness. Curvedness and shape index are no radiomics measures and were therefore separately calculated [27]. Agreement between ground truth and model measurements were visualized with Bland-Altman plots for Dataset 1 (test set) and 3 combined. Proportional bias was defined as measure deviations proportional to the mean of the measure, assessed with linear regression. For measures showing proportional bias, a prediction interval was visualized in the Bland-Altman plots in addition to the 'classical' limits of agreement [28]. The highest deviating aneurysm from each measure was visualized as 3D mesh reconstruction in the Bland-Altman plots to provide understanding of differences in measures. Three examples with high DSC and low 95%-HD were additionally included. Regression significance was set at $p<0.05$.

Qualitative validation

Although considered the ground truth, manual segmentations are not without error; previous studies have reported inter-rater volumetric differences ranging from 3.6 to 20.6mm^3 [9]. Therefore, a post-hoc qualitative validation was performed on outlier aneurysms identified from the Bland-Altman plots. For each 3D morphological measure, scans were selected for review: on the one hand, cases that fell outside the 'classical' ±1.96 standard deviation limits (i.e. clear outliers; n=23), and on the other hand, a subset of cases that exhibited only small differences in morphological measures (control scans; n=7). On these scans a blinded evaluation was performed by ICvdS, selecting the preferred segmentation between the manual ground truth and the automated model output. In cases where the two segmentations appeared visually similar, the rater documented this similarity and additionally selected the segmentation judged to be more accurate. Segmentations were

presented side-by-side as image overlays in axial, coronal, and/or sagittal views using 3D slicer. Randomization of left/right positioning was performed using Python. Results were visualized in a bar-plot and group differences were assessed using an exact binomial test.

Results

Included aneurysms

An overview of basic aneurysm characteristics is provided (Table 2). Each dataset contained scans with multiple aneurysms. The primary aneurysms in Dataset 3 (72.7 [9.3-433.5] mm^3) were larger than Dataset 1 (55.8 [15.0-303.1] mm^3) and Dataset 2 (51.2 [0.77-389.7] mm^3).

Model performance metrics

Inference was performed on an NVIDIA RTX A5000 GPU, with an average 5-fold ensemble processing time of 55-130 seconds per 3D volume. An example of the segmentation output from the test set of Dataset 1 is shown in 2D-overlay and 3D (Figure 3). Model 2 achieved a high mean vessel and aneurysm DSC with low 95%-HD for both internal and external test sets (Dataset 1 and 3; Table 3) and the DSC remained stable over three inference runs. Overall DSC and HD values and results from experiments with manually annotated aneurysm center of mass are additionally provided (Table 3). A statistical comparison between Model 1 and 2, precision and recall box plots and an aneurysm volume stratification using box plots is provided in the supplemental materials (Supplemental Table 2, Figures 2, 3 and 4, respectively).

The relation between DSC and 95%-HD scores for both test sets is visualized (Figure 4). The lowest scoring segmentations from both test sets (Figure 4A and C), and best scoring segmentation from the test set of Dataset 1 (Figure 4B) are additionally shown in 2D-overlay and 3D. The lowest scoring segmentation from Dataset 3 shows severe under segmentation by the model in an aneurysm with inhomogeneous image intensities. Because this aneurysm

is a severe outlier compared to the other results, this image was removed from further analyses.

3D morphological measures

Bland-Altman plots of 3D morphological measures showed high agreement between manual ground truth and model segmentations (Figure 5 and Supplemental Figure 5). Mean differences were close to zero and limits of agreements were narrow. Voxel volume, surface area, sphericity and shape index showed proportional bias with a p-value ≤0.002 (Figure 5). 3D reconstructions of aneurysm segmentations show the interpretation of differences in measures from ground truth or model segmentations.

Qualitative validation

Of the 30 assessed scans for qualitative validation by an experienced rater, the trained model's segmentations were selected as more accurate in 17 (56.7%) cases, while the ground truth was selected as more accurate in 13 (43.3%) cases (Figure 6). In the control group, the model segmentation was selected as more accurate in five scans compared versus two for the ground truth (orange). Similarly, among the 6 (26%) outlier scans that were judged to be visually similar, the model was selected as more accurate in five cases versus one (green). In contrast, among the outlier cases (blue), the ground truth segmentation was preferred in 10 out of 17 scans (58.9%). However, neither for the entire test (p=0.585), nor for only the outlier cases (p=0.648), did the binomial test indicate a significant preference for ground truth or model segmentation.

Discussion

In this study, we trained and evaluated DIVA-seg, an nnU-Net model for segmentation of intracranial vessels and aneurysms on MRA-TOF using a pseudo labeling approach. The model achieved high DSC and 95%-HD performance measures and a high agreement between 3D morphological measures derived from the model's and ground truth annotations. These results demonstrate the potential of our model for accurate and reliable segmentation of intracranial vasculature and aneurysms in a fully automated manner.

Since 2019, efforts have been made to develop artificial intelligence methods for the detection and segmentation of UIA from MRA-TOF images [10]. However, reported DSC have generally remained low; <0.750 [10,18]. One study, using a 3D convolutional neural network, reported a peak test DSC of 0.980, suggesting near-perfect agreement between model predictions and ground truth [16]. A direct comparison to our results however remains challenging due to differences in evaluation design and reporting clarity. The described evaluation appears to have been conducted on a limited test set of six scans from two datasets [16]. The extent to which these scans represent the full variability of aneurysm presentations, such as differences in size, location, shape and image quality, is not clearly discussed. Two previous studies also reported a high aneurysm DSC on internal test sets. One study, utilizing nnU-Net, achieved a median aneurysm DSC of 0.85, but reported a larger median 95%-HD of 2.33mm and larger median aneurysm volumes (100.5mm^3) compared with the current study [17]. Another study, using a transformer-based network, reported an average aneurysm DSC of 0.811 on an in-house dataset and 0.559 on the ADAM dataset (both also used for training) [29]. The in-house dataset contained larger aneurysms (4-28mm) than the ADAM dataset (1-17mm). Focusing on the ADAM results, our internal test set (also ADAM data) achieved a substantially higher DSC. Notably, both studies did not include external

testing. Taken together, compared with previous work, DIVA-seg achieved higher DSC scores across both internal and external test sets.

When comparing DSC results from the internal test set using manual versus automated center of mass annotations, vessel DSC remained identical and aneurysm DSC differed only marginally (0.019). These findings indicate that new datasets with manually annotated aneurysm centers can also be expected to achieve accurate segmentation results, thereby supporting feasibility and ease of application in practice.

In the external test set, we observed one scan with low aneurysm DSC score and high *95%*-HD (Figure 4), which was caused by insufficient (intensity) filling of the aneurysm as seen on the MRA-TOF image. These inconsistencies in scan quality from MRA-TOF data may influence segmentation results and should be carefully monitored. Because this scan was a severe outlier, we removed the scan from subsequent analyses.

An important added value of our study is the translation of aneurysm segmentation into 3D quantification of aneurysm morphology, as accurate estimation of 3D measures is essential for change detection and prediction modeling. To the best of our knowledge, this is the first study to systematically evaluate differences in 3D morphological measurement from MRA-TOF segmentations between the model and ground truth. While similar clinically oriented evaluations have been performed for CTA and 3D rotational angiography segmentation methods [30–37], such analyses have not yet been applied to MRA-TOF models. One exception is a prior MRA-TOF segmentation model that assessed maximum diameter; however, ground truth measurements in that work were obtained from DSA images rather than MRA-TOF [38].

The small deviation from 0 of the mean difference in 3D morphological measures indicates that there was no systematic bias between the ground truth and model

segmentations. To put the measures into perspective, assume a spherical aneurysm with a diameter of 3mm and 6mm. A mean volume difference of 8.38mm^3 corresponds to adding a uniform layer of 0.25mm thickness around an aneurysm of 3mm, and a uniform layer of 0.07mm all around an aneurysm of 6mm. Even for very small aneurysms, these distances are all <1 voxel distance (0.36mm) difference in segmentation and hold no clinical relevance. In a previous study, researchers found an average 9.1mm^3 difference between raters, which is slightly higher than DIVA-seg compared to the ground truth [9]. These findings indicate that the model's error is similar to the error between raters. However, the Bland-Altman plots showed a proportional bias for four morphological measures. This type of bias for voxel volume and surface area is expected; one additional layer of voxels at the edge of a larger aneurysm results in a larger increase in volume and surface area. For sphericity and shape index, the proportional bias is mainly determined by the definition of the neck (Figure 5 reconstructed inserts), which remains difficult for clinical experts as well. Though the model segmentations of outlier cases generally appear more spherical than the ground truth.

Because manually generated aneurysm segmentations are not always a perfect ground truth, we evaluated our own ground truth using a qualitative expert preference analysis. The results indicated no consistent preference between the ground truth and model outputs. Interestingly, for aneurysms rated as visually similar or included as control, the model segmentation was more frequently judged as more accurate. For other cases, there was no clear preference. These findings highlight that even carefully curated ground truths may be imperfect, and in some instances, model segmentations can be equally accurate or even preferred. This aligns with broader recommendations in the medical imaging literature suggesting that ground truth should be critically evaluated, for example by expert review or blinded comparison, to ensure reliability and support clinical translation [39,40].

Strengths and limitations

Compared to previous MRA-TOF models, DIVA-seg was trained on a larger dataset; 541 scans compared to 62-217 in other models [10]. As outlined in the introduction, our approach aimed to overcome the limitations of small training cohorts by combining high-quality manual annotations with a larger pseudo labeled dataset to improve generalizability. The use of pseudo labeling allowed the inclusion of a larger and more diverse dataset without requiring time consuming manual labeling for all scans. Although the internal test DSC decreased slightly from Model 1 to 2, validation and external DSC significantly increased and *95%*-HD decreased, suggesting improved model generalizability from incorporating a large pseudo labeled dataset. In addition, the use of the nnU-Net architecture promotes standardized preprocessing and data augmentation to improve data quality, which has been a source of variability in previous models [10]. An external validation on a heterogeneous dataset with MRA-TOF scans from different vendors, scanner models, imaging protocols and aneurysm sizes, confirmed the generalizability of the trained model for other data sources. Furthermore, a challenge in many segmentation models is the clinical interpretability of technical segmentation metrics, such as the DSC. The DSC is highly dependent on the size of the segmented structure, which can vary greatly within aneurysms. A new strategy for evaluating aneurysm segmentation models may be required to better understand the clinical implications of segmentation differences. In this study we have provided a first step towards this different strategy by including 3D morphological measures in the analysis of the model, allowing a more clinically meaningful interpretation of segmentation differences.

Nevertheless, some limitations should be addressed. First, the proposed model requires the aneurysm location information as input channel. The model does therefore not allow for

detection of the aneurysms, manual annotation of the location or an independent detection should still precede segmentation. Second, the dataset for pseudo labels (Dataset 2) included aneurysms of all sizes to further finetune the model. Although the other datasets also contained tiny aneurysms as non-primary aneurysm, the performance of the model on these aneurysms was not separately tested, and inclusion of these aneurysms may have slightly altered the performance of the model. Model performance might have improved if tiny aneurysms were excluded, but they were retained to better reflect clinical practice, where such cases are common. Third, the ground truth segmentation from the external dataset was differently generated than the ground truth from the in-house dataset. Although we kept segmentation approaches as similar as possible by using the open access COSTA segmentations for Dataset 1 and the CESAR model, trained on COSTA data, for Dataset 3, the difference in approach may have affected our quantitative assessment for vessel segmentations. Nevertheless, aneurysm segmentations were all manually annotated by ASV and checked by ICvdS. Fourth, a formal stratified analysis of model performance (e.g. by aneurysm size or location) was limited by the heterogeneity of the test sets, uneven subgroup distributions and missing information on secondary aneurysm locations. Nonetheless, a stratified evaluation by aneurysm volume is provided in the supplemental data. Future work could explore additional stratified approaches, such as normalized surface distance or volume-specific DSC reporting, to further characterize model performance. Finally, hyperparameters were primarily selected using nnU-Net's automated data-driven optimization strategy. In addition, a limited empirical assessment was performed by evaluating training behavior under alternative settings, including variations in patch size, batch size and number of epochs; however, these adjustments did not improve the model's performance. More extensive optimization strategies, such as automated hyperparameter

tuning using tools like Optuna [41] and early stopping, could be explored in future work to further refine model performance.

Future directions

In future studies, a systematic comparison with other segmentation models and architectures should be performed to further evaluate performance differences, robustness and generalization across datasets. Such comparisons should, however, be interpreted with caution. Furthermore, this study focused on the development of an MRA-TOF segmentation model. However, in clinical practice, some patients may undergo computed tomography angiography (CTA) instead of, or in addition to MRA-TOF. Future work should aim to extend segmentation algorithms to CTA, ideally through modality-agnostic approaches validated on paired datasets from patients who have undergone imaging from multiple modalities within a short time interval. Following the development of segmentation models, accurate automated segmentations will enable comprehensive 3D morphological analyses of aneurysms, providing a stronger and more reproducible basis for change detection and prediction models for aneurysm growth and rupture risks. The next step is therefore to leverage these segmentations in the development of such prediction models, with the goal of supporting personalized treatment decisions.

Beyond clinical decision-making, accurate automated vessel and aneurysm segmentations also enable technical applications. For example, the proposed model could be used for fine-tuning to generate segmentations for 4D flow analyses aimed at improving our understanding of cerebral blood flow and its role in aneurysm development and rupture [42].

Conclusion

This study demonstrates that an nnU-Net deep learning model, leveraging the use of pseudo labeling, provides accurate and reliable segmentation of intracranial vessels and aneurysms across internal and external test sets, with a high agreement between 3D morphological measures derived from automated and manual segmentations. Future work should aim to extend segmentation models to other modalities and further investigate the use of 3D morphological measures from automated segmentations for change detection as well as growth and rupture risk prediction.

Acknowledgement

ChatGPT (OpenAI, GPT-5) was used for selective language refinement (e.g., phrasing and style). The authors integrated these suggestions into their own writing; all scientific content, interpretations, and conclusions remain the authors' original work.

Data and software availability

The clinical imaging data used in this study are not publicly available in order to protect patient privacy and comply with institutional and ethical regulations. The model weights and configuration files required for inference are available on GitHub: https://github.com/Verschuur95/DIVA-seg.

References

1. Vlak MH, Algra A, Brandenburg R, et al. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. Lancet Neurol. 2011;10(7):626–636. doi: 10.1016/S1474-4422(11)70109-0.

2. Asikainen A, Korja M, Kaprio J, et al. Sex Differences in Case Fatality of Aneurysmal Subarachnoid Hemorrhage: A Systematic Review. Neuroepidemiology. 2024;1–14. doi: 10.1159/000538562.

3. Greving JP, Wermer MJH, Brown RD, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. Lancet Neurol. 2014;13(1):59–66. doi: 10.1016/S1474-4422(13)70263-1.

4. Backes D, Rinkel GJE, Greving JP, et al. ELAPSS score for prediction of risk of growth of unruptured intracranial aneurysms. Neurology. 2017;88(17):1600–1606. doi: 10.1212/WNL.0000000000003865.

5. Van Der Kamp LT, Rinkel GJE, Verbaan D, et al. Risk of Rupture After Intracranial Aneurysm Growth. JAMA Neurol. American Medical Association (AMA); 2021;78(10):1228. doi: 10.1001/jamaneurol.2021.2915.

6. Kamphuis MJ, Timmins KM, Kuijf HJ, et al. Three-Dimensional Morphological Change of Intracranial Aneurysms Before and Around Rupture. Neurosurgery. 2024; doi: 10.1227/neu.0000000000002812.

7. Villablanca JP, Duckwiler GR, Jahan R, et al. Natural History of Asymptomatic Unruptured Cerebral Aneurysms Evaluated at CT Angiography: Growth and Rupture Incidence and Correlation with Epidemiologic Risk Factors. Radiology. 2013;269(1):258–265. doi: 10.1148/radiol.13121188.

8. Bor ASE, Rinkel GJE, Van Norden J, et al. Long-term, serial screening for intracranial aneurysms in individuals with a family history of aneurysmal subarachnoid haemorrhage: a cohort study. Lancet Neurol. 2014;13(4):385–392. doi: 10.1016/S1474-4422(14)70021-3.

9. Timmins KM, Kuijf HJ, Vergouwen MDI, et al. Reliability and Agreement of 2D and 3D Measurements on MRAs for Growth Assessment of Unruptured Intracranial Aneurysms. Am J Neuroradiol. 2021;42(9):1598–1603. doi: 10.3174/ajnr.A7186.

10. Mata-Castillo M, Hernández-Villegas A, Gordillo-Castillo N, et al. Systematic review of artificial intelligence methods for detection and segmentation of unruptured intracranial aneurysms using medical imaging. Med Biol Eng Comput. Springer Science and Business Media LLC; 2025; doi: 10.1007/s11517-025-03345-7.

11. Chen M, Geng C, Wang D, et al. A coarse-to-fine cascade deep learning neural network for segmenting cerebral aneurysms in time-of-flight magnetic resonance angiography. Biomed Eng OnLine. 2022;21(1):71. doi: 10.1186/s12938-022-01041-3.

12. Ham S, Seo J, Yun J, et al. Automated detection of intracranial aneurysms using skeleton-based 3D patches, semantic segmentation, and auxiliary classification for overcoming data imbalance in brain TOF-MRA. Sci Rep. 2023;13(1):12018. doi: 10.1038/s41598-023-38586-9.

13. Sichtermann T, Faron A, Sijben R, et al. Deep Learning-Based Detection of Intracranial Aneurysms in 3D TOF-MRA. AJNR Am J Neuroradiol. United States; 2019;40(1):25–32. doi: 10.3174/ajnr.A5911.

14. Yang X, Xia D, Kin T, et al. A two-step surface-based 3D deep learning pipeline for segmentation of intracranial aneurysms. Comput Vis Media. 2023;9(1):57–69. doi: 10.1007/s41095-022-0270-z.

15. Yuan W, Peng Y, Guo Y, et al. DCAU-Net: dense convolutional attention U-Net for segmentation of intracranial aneurysm images. Vis Comput Ind Biomed Art. Germany; 2022;5(1):9. doi: 10.1186/s42492-022-00105-4.

16. Ye J, Xu X, Li L, et al. OTO-Net: An Automated MRA Image Segmentation Network for Intracranial Aneurysms. Zhang G, editor. Comput Intell Neurosci. 2022;2022:1–14. doi: 10.1155/2022/5333589.

17. Colombo E, De Boer M, Bartels L, et al. Accuracy of an nnUNet Neural Network for the Automatic Segmentation of Intracranial Aneurysms, Their Parent Vessels, and Major Cerebral Arteries from MRI-TOF. Am J Neuroradiol. 2025;46(5):956–963. doi: 10.3174/ajnr.A8607.

18. Timmins KM, van der Schaaf IC, Bennink E, et al. Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge. NeuroImage. 2021;238. doi: 10.1016/j.neuroimage.2021.118216.

19. Mou L, Lin J, Zhao Y, et al. COSTA: A Multi-Center TOF-MRA Dataset and a Style Self-Consistency Network for Cerebrovascular Segmentation. IEEE Trans Med Imaging. Institute of Electrical and Electronics Engineers (IEEE); 2024;43(12):4442–4456. doi: 10.1109/tmi.2024.3424976.

20. De Nys CM, Liang ES, Prior M, et al. Time-of-Flight MRA of Intracranial Aneurysms with Interval Surveillance, Clinical Segmentation and Annotations. Sci Data. 2024;11(1):555. doi: 10.1038/s41597-024-03397-8.

21. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–211. doi: 10.1038/s41592-020-01008-z.

22. Huang Z, Wang H, Ye J, et al. Revisiting nnU-Net for Iterative Pseudo Labeling and Efficient Sliding Window Inference. In: Ma J, Wang B, editors. Fast Low-Resour

Semi-Supervised Abdom Organ Segmentation. Cham: Springer Nature Switzerland; 2022. p. 178–189. doi: 10.1007/978-3-031-23911-3_16.

23. Timmins KM, Kuijf HJ, Vergouwen MDI, et al. Relationship between 3D Morphologic Change and 2D and 3D Growth of Unruptured Intracranial Aneurysms. Am J Neuroradiol. American Society of Neuroradiology (ASNR); 2022;43(3):416–421. doi: 10.3174/ajnr.a7418.

24. Kamphuis MJ, Van Der Kamp LT, Van Eijk RPA, et al. Three-Dimensional Quantified Morphological Predictors of Intracranial Aneurysm Instability: A Longitudinal Study. Am J Neuroradiol. American Society of Neuroradiology (ASNR); 2025;ajnr.A8809. doi: 10.3174/ajnr.a8809.

25. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. American Association for Cancer Research (AACR); 2017;77(21):e104–e107. doi: 10.1158/0008-5472.can-17-0339.

26. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338. doi: 10.1148/radiol.2020191145.

27. Koenderink JJ, Van Doorn AJ. Surface shape and curvature scales. Image Vis Comput. Elsevier BV; 1992;10(8):557–564. doi: 10.1016/0262-8856(92)90076-f.

28. Ludbrook J. Confidence in Altman–Bland plots: A critical review of the method of differences. Clin Exp Pharmacol Physiol. Wiley; 2010;37(2):143–149. doi: 10.1111/j.1440-1681.2009.05288.x.

29. Guo L, Liang Y, Guo R, et al. Staged cluster transformers for intracranial aneurysms segmentation from structure fused 3D MRA. Int J Imaging Syst Technol. 2024;34(2). doi: 10.1002/ima.23039.

30. Ou C, Qian Y, Chong W, et al. A deep learning-based automatic system for intracranial aneurysms diagnosis on three-dimensional digital subtraction angiographic images. Med Phys. United States; 2022;49(11):7038–7053. doi: 10.1002/mp.15846.

31. Niemann A, Behme D, Larsen N, et al. Deep learning-based semantic vessel graph extraction for intracranial aneurysm rupture risk management. Int J Comput Assist Radiol Surg. Germany; 2023;18(3):517–525. doi: 10.1007/s11548-022-02818-6.

32. Nishi H, Cancelliere NM, Rustici A, et al. Deep learning-based cerebral aneurysm segmentation and morphological analysis with three-dimensional rotational angiography. J NeuroInterventional Surg. 2024;16(2):197–203. doi: 10.1136/jnis-2023-020192.

33. Nie X, Yang Y, Liu Q, et al. A deep-learning system to help make the surgical planning of coil embolization for unruptured intracranial aneurysms. Chin Neurosurg J. England; 2023;9(1):24. doi: 10.1186/s41016-023-00339-y.

34. Patel TR, Patel A, Veeturi SS, et al. Evaluating a 3D deep learning pipeline for cerebral vessel and intracranial aneurysm segmentation from computed tomography angiography-digital subtraction angiography image pairs. Neurosurg Focus. United States; 2023;54(6):E13. doi: 10.3171/2023.3.FOCUS2374.

35. Feng J, Zeng R, Geng Y, et al. Automatic differentiation of ruptured and unruptured intracranial aneurysms on computed tomography angiography based on deep learning and radiomics. Insights Imaging. Germany; 2023;14(1):76. doi: 10.1186/s13244-023-01423-8.

36. Yang Y, Chang Z, Nie X, et al. Integrated Deep Learning Model for the Detection, Segmentation, and Morphologic Analysis of Intracranial Aneurysms Using CT Angiography. Radiol Artif Intell. 2025;7(1):e240017. doi: 10.1148/ryai.240017.

37. Rezaeitaleshmahalleh M, Lyu Z, Mu N, et al. Using convolutional neural network-based segmentation for image-based computational fluid dynamics simulations of brain aneurysms: initial experience in automated model creation. J Mech Med Biol. 2023;23(04):2340055. doi: 10.1142/S0219519423400559.

38. Zhang M, Jin Q, Fu M, et al. Intracranial aneurysm segmentation using a deep convolutional neural network. Proc Intl Soc Mag Reson Med. Online; 2021. p. 2665.

39. Larroza A, Pérez-Benito FJ, Tendero R, et al. Three-Blind Validation Strategy of Deep Learning Models for Image Segmentation. J Imaging. 2025;11(5):170. doi: 10.3390/jimaging11050170.

40. Poel R, Rüfenacht E, Scheib S, et al. A comprehensive multifaceted technical evaluation framework for implementation of auto-segmentation models in radiotherapy. Commun Med. 2025;5(1):319. doi: 10.1038/s43856-025-01048-6.

41. Akiba T, Sano S, Yanase T, et al. Optuna: A Next-generation Hyperparameter Optimization Framework. Proc 25th ACM SIGKDD Int Conf Knowl Discov Data Min. Anchorage AK USA: ACM; 2019. p. 2623–2631. doi: 10.1145/3292500.3330701.

42. Zhang J, Verschuur A, Schrauben E, et al. Evaluating segmentation techniques for circle of willis in 4D flow MRI: a comparative study. Proc Int Soc Mag Reson Med. Honolulu, Hawaii; 2025. p. 1216.

Tables

Table 1: Acquisition and preprocessing settings MRA-TOF from the three datasets.

Data	Field	TR/TE	Voxel spacing (median; min; max)	Matrix size (median; min; max)	Intensity normalization	Resampling	BET
1.	1.5T - 3T	22/3.45 ms	036x0.36x0.5mm; 0.20x0.20x0.5mm; 0.63x0.63x0.8mm	370x461x138; 227x273x65; 719x940x180	Histogram standardization (by [21]). nnU-net default – z-score	nnU-Net default – anisotropic resampling [21]	Yes
2.	1.5T - 3T	22/3.45 ms	0.36x0.36x0.5mm; 0.20x0.20x0.20mm; 0.81x0.81x2mm	560x560x140; 256x256x42; 1024x1024x479	nnU-Net default – z-score	nnU-Net default	No
3.	1.5T - 3T	18.6 – 164.7 / 3.5 – 6.6	0.31x0.31x0.44mm; 0.22x0.22x0.30mm; 0.625x0.625x0.90mm	560x560x140; 260x245x72; 720x768x400	nnU-Net default – z-score	nnU-Net default	No

T: Tesla; TR/TE: repetition time/ echo time; BET: brain extraction; ms: millisecond; mm: millimeter.

562 *Table 2: Basic aneurysm characteristics per dataset.*

	Dataset 1	**Dataset 2**	**Dataset 3**
Number of patients, n	47	275	47
Number of unique aneurysms from first scan, n	66	356	63
Multiplicity on first scan, n	15	63	11
Primary* aneurysm volume in mm^3, median (range)	55.8 (15.0-303.1)	51.2 (0.77-389.7)	72.7 (9.3-433.5)
Non-primary aneurysm volume in mm^3, median (range)	18.5 (5.3-67.5)	15.2 (0.75-161.5)	34.8 (6.7-114.2)
Primary aneurysm location, MCA/ICA/ACA/Post/PCoA/unknown	18/12/9/4/4/0	88/56/49/28/12/42	26/20/12/4/1/0

563 *Primary = largest. MCA: middle cerebral artery; ICA: internal carotid artery; ACA: anterior
564 cerebral artery or anterior communicating artery; Post: posterior circulation; PCoA: posterior
565 communicating artery.

Table 3: Performance metrics of Model 2.

	Dataset 1 - Test set	Dataset 1 - manual aneurysm COM	Dataset 3 - External test set
Overall DSC, mean (95%-CI)	0.926 (0.910-0.941)	0.926 (0.910-0.941)	0.900 (0.894-0.907)
DSC Vessels, mean (95%-CI)	0.925 (0.910-0.940)	0.925 (0.910-0.940)	0.899 (0.893-0.906)
DSC Aneurysm, mean (95%-CI)	0.880 (0.853-0.906)	0.861 (0.819-0.902)	0.861 (0.836-0.886)
Aneurysm 95%-HD in mm, mean (95%-CI)	0.671 (0.446-0.895)	0.798 (0.511-1.08)	0.674 (0.529-0.819)

DSC: Dice similarity coefficient; 95%-HD: 95%-Hausdorff distance; SD: standard deviation; 95%-CI: 95% Confidence Interval (with 5000 bootstrapping).

Figure captions:

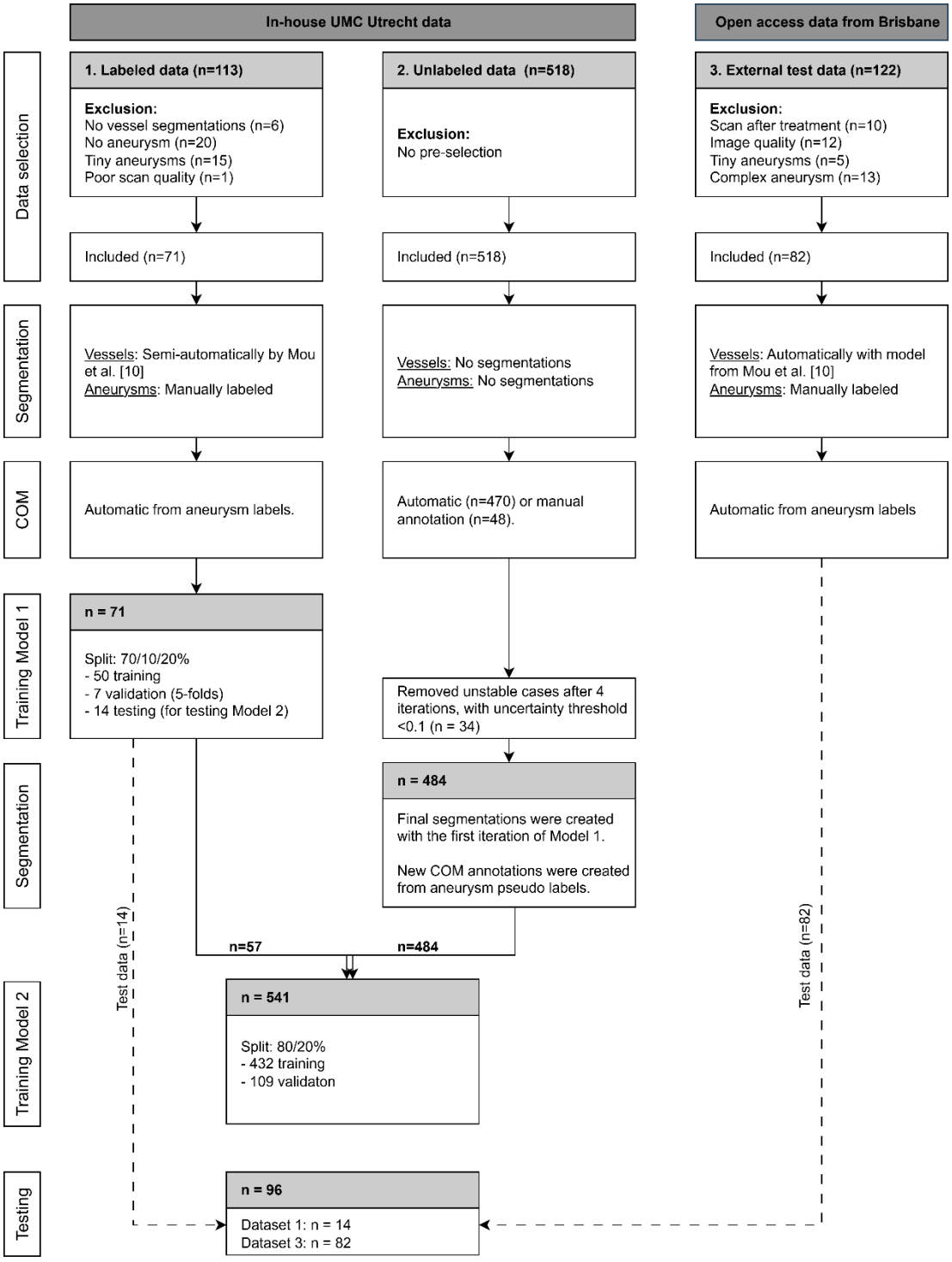


Figure 1: Inclusion flowchart for the three used datasets, providing exclusion criteria and numbers, segmentation method, center of mass estimation, final numbers for inclusion, and training/validation/testing split for each model.

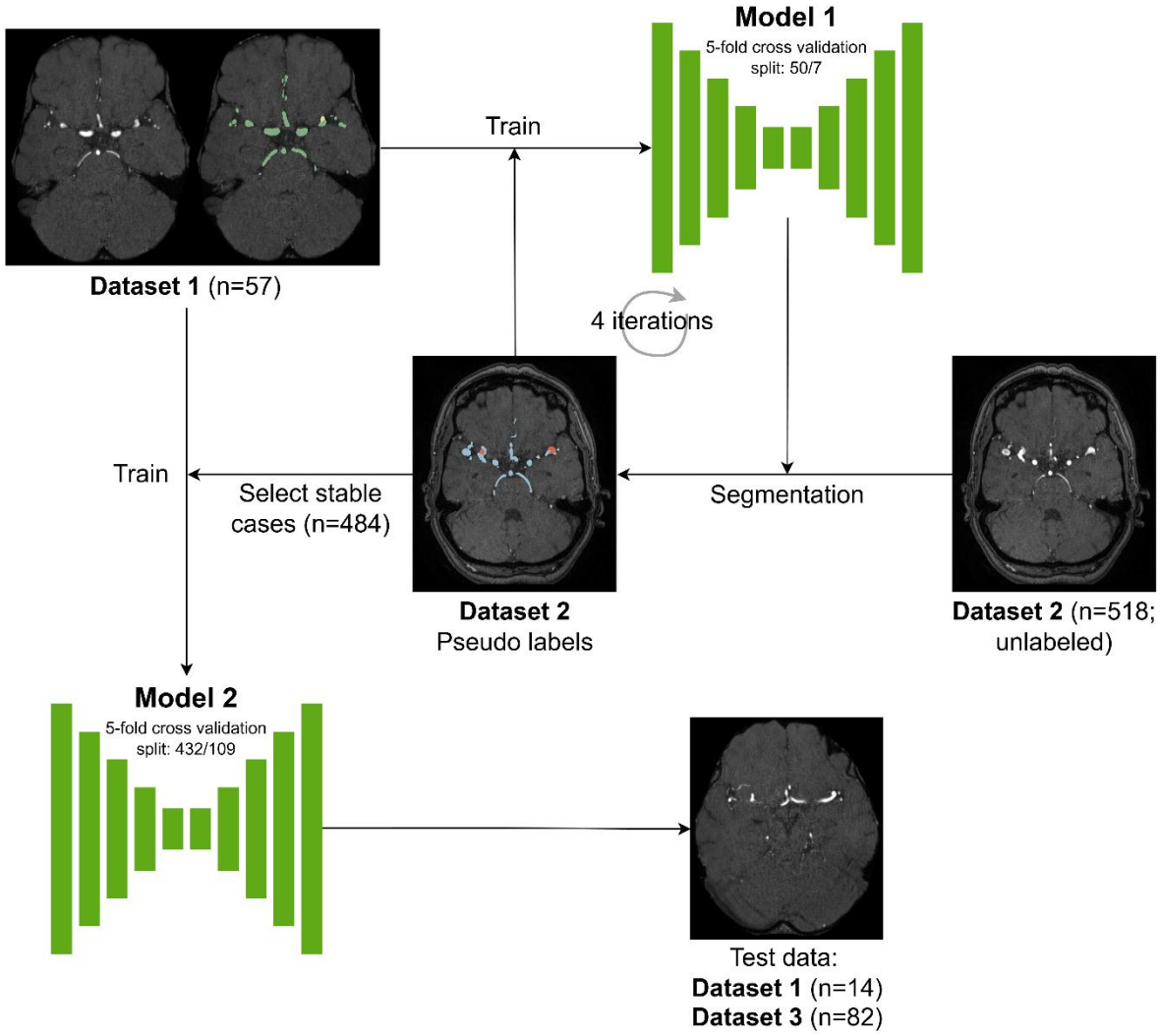


Figure 2: Overview of the DIVA-seg training pipeline and pseudo label generation process. Model 1 was iteratively trained four times using Dataset 1 (first iteration) or Dataset 1 and 2 with pseudo labels (iteration two to four), followed by training of Model 2 using Dataset 1 and selected stable scans from Dataset 2. Model 2 was tested on the test set from Dataset 1 and an external test set, Dataset 3.

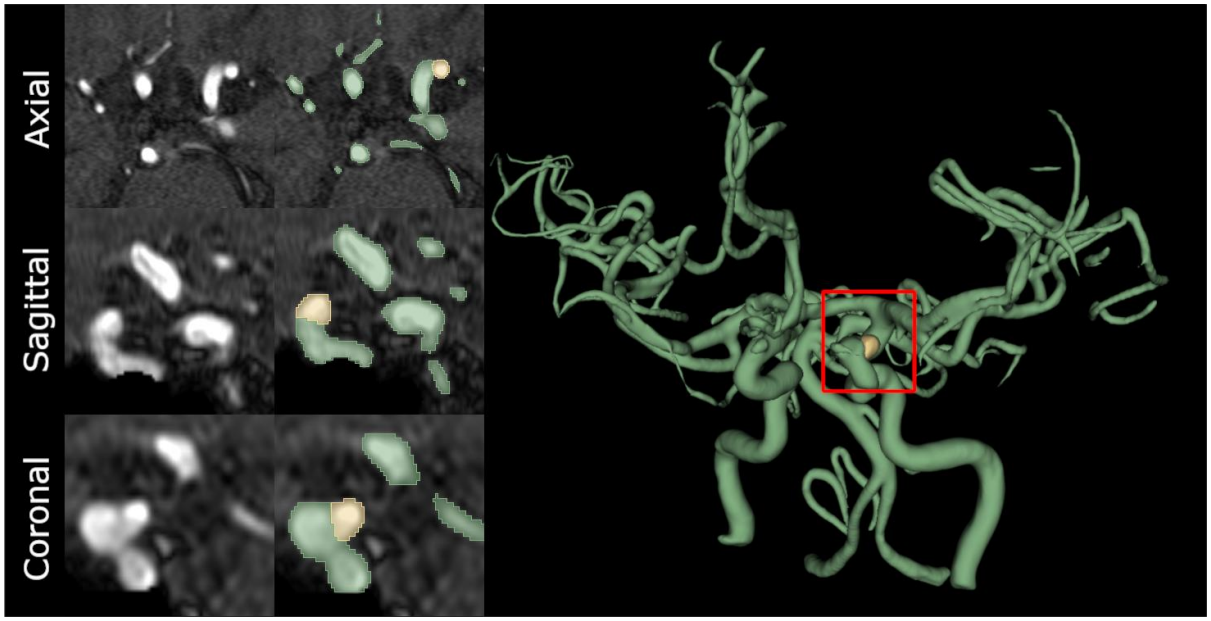


Figure 3: Segmentation example from the Dataset 1 test set. From left to right, zoomed area of an MRA-TOF image, segmentation overlay on MRA-TOF and 3D reconstruction of the vessel tree (green) with an internal carotid artery aneurysm (yellow). The red box is the visualized area in the 2D(-overlay) images.

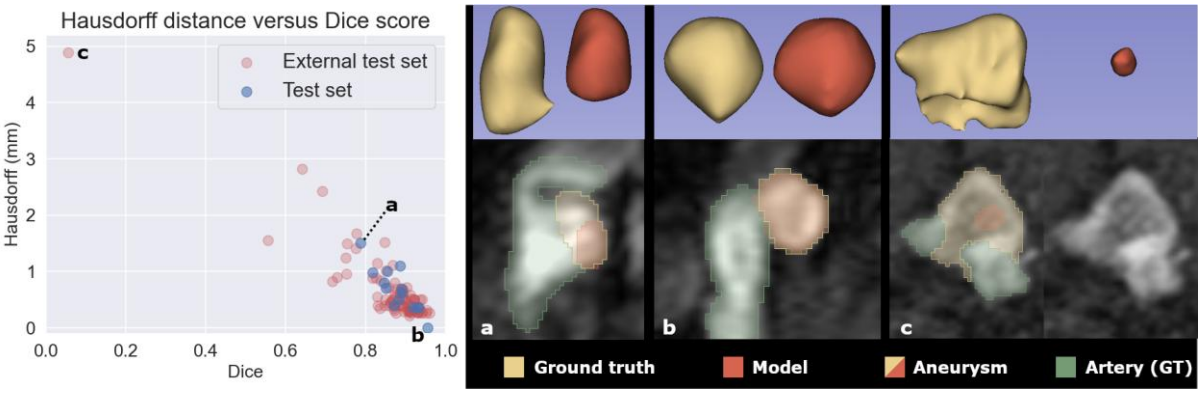


Figure 4: Scatter plot of performance metrics; 95%-Hausdorff distance versus Dice similarity coefficient scores. Annotated in the scatter plot and rendered in 3D and 2D-overlay are a) lowest scoring aneurysm from the test set of Dataset 1, b) best scoring aneurysm from the test set of Dataset 1, and c) lowest scoring aneurysm from Dataset 3. Insufficient filling of the aneurysm caused under segmentation in the model's results.

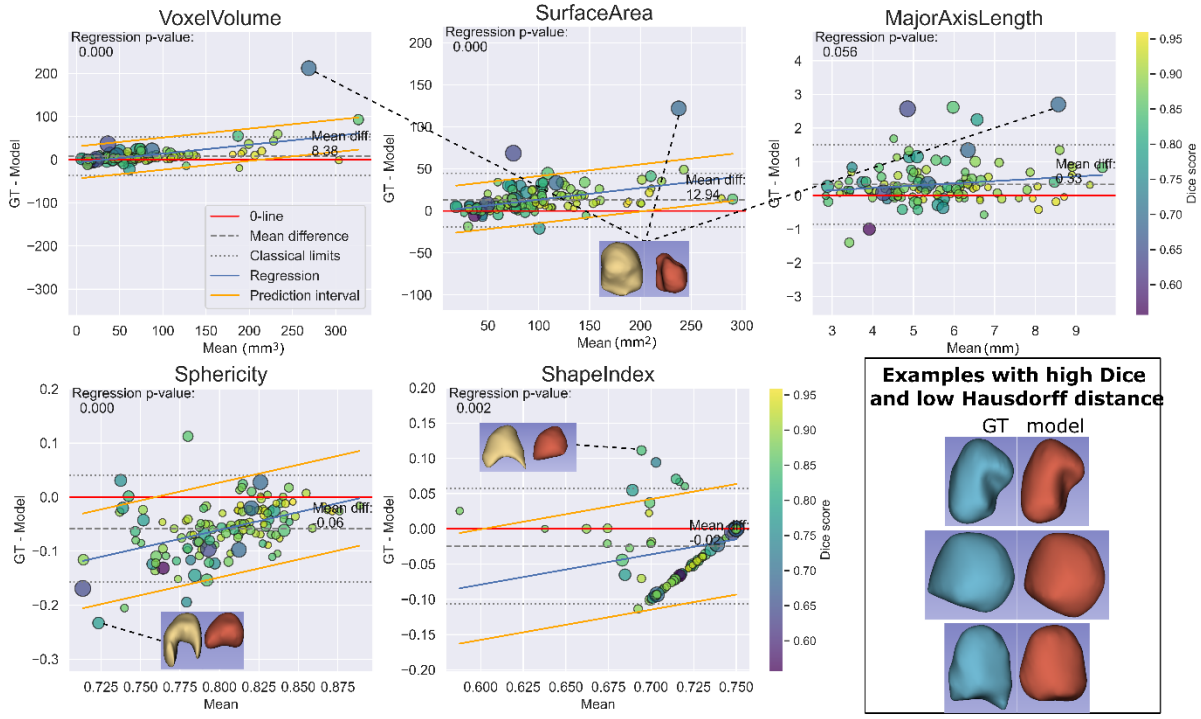
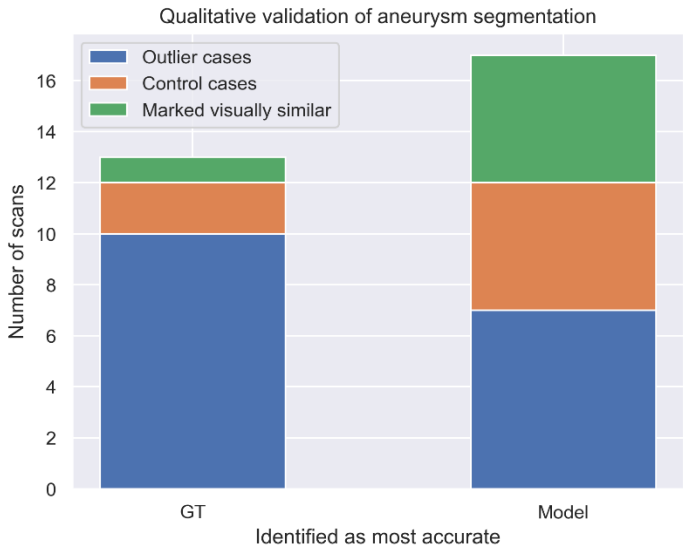


Figure 5: Bland-Altman plots assessing agreement between automated 3D morphological measurements from ground truth and deep learning-based segmentations across test sets from Dataset 1 and 3. Each plot displays the difference between ground truth and model-derived measurements plotted against their mean. Overall, the results demonstrate high agreement with low bias and narrow limits of agreement. Proportional bias was observed in 4 out of 8 measures (linear regression, p-value <0.05), though prediction intervals remained narrow. Scatter points are color coded by Dice Similarity Coefficient (DSC; see color bar) and sized by 95%-Hausdorff distance (95%-HD). Inserts show 3D reconstructions of ground truth (yellow) and model-predicted (red) segmentation corresponding to the largest outliers for each metric. For comparison, three representative examples with high DSC and low 95%-HD are shown in the bottom right corner of the figure. See also Supplemental Figure 2 for elongation, flatness and curvedness.

608



609

610 *Figure 6: Qualitative validation of aneurysm segmentation as assessed by a*

611 *neurointerventional radiologist who was blinded to segmentation method.*