Machine Learning with Applications 23 (2026) 100843

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Analysis of major segmentation models for intracranial artery time-of-flight magnetic resonance angiography images☆

Mekhla Sarkar [a], Yen-Chu Huang [b,c], Tsong-Hai Lee [c,d], Jiann-Der Lee [b,c], Prasan Kumar Sahoo [a,b,*]

[a] Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan
[b] Department of Neurology, Chang Gung Memorial Hospital Chiayi, Chiayi City, Taiwan
[c] College of Medicine, Chang Gung University, Taoyuan, Taiwan
[d] Department of Neurology, Chang Gung Memorial Hospital, Linkou Medical Center, Fu Hsing Street, Guishan, Taoyuan 333, Taiwan

ARTICLE INFO

Keywords:
Semantic segmentation
Backbones
UNet
LinkNet
FPN
PSPNet
Intracranial arterial stenosis

ABSTRACT

Intracranial arterial stenosis (ICAS) is a leading cause of cerebrovascular accidents, and accurate morphological assessment of intracranial arteries is critical for diagnosis and treatment planning. Complex vascular structures, imaging noise, and variability in time-of-flight magnetic resonance angiography (TOF-MRA) images are challenging issues for the manual delineation that motivates the use of deep learning (DL) for automatic segmentation of the intracranial arteries. DL based automatic segmentation offers a promising solution by providing consistent and noise-reduced vessel delineation. However, selecting an optimal segmentation architecture remains challenging due to the diversity of network designs and encoder backbones. Therefore, this study presents a systematic benchmarking of five widely used DL segmentation architectures, UNet, LinkNet, Feature Pyramid Networks (FPN), Pyramid Scene Parsing Network (PSPNet), and DeepLabV3+, each combined with nine backbone networks, yielding 45 model variants, including previously unexplored configurations for intracranial artery segmentation in TOF-MRA. Models were trained and cross-validated on four datasets: in-house, CereVessMRA, IXI and ADAM, and evaluated on held-out independent test set. Performance metrics included Intersection over Union (IoU), Dice Similarity Coefficient (DSC), and a Stability Score, combining the coefficient of variation of IoU and DSC to quantify segmentation consistency and reproducibility. Experimental results demonstrated highest DSC score was achieved with UNet–SE-ResNeXt50, LinkNet-SE-ResNeXt50, FPN-DenseNet169, FPN-SENet154. The most stable configurations were LinkNet–EfficientNetB6, LinkNet–SENet154, UNet–DenseNet169, and UNet–EfficientNetB6. Conversely, DeepLabV3+ and PSPNet variants consistently underperformed. These findings provide actionable guidance for selecting backbone–segmentation pairs and highlight trade-offs between accuracy, robustness, and reproducibility for complex intracranial artery TOF-MRA segmentation tasks.

1. Introduction

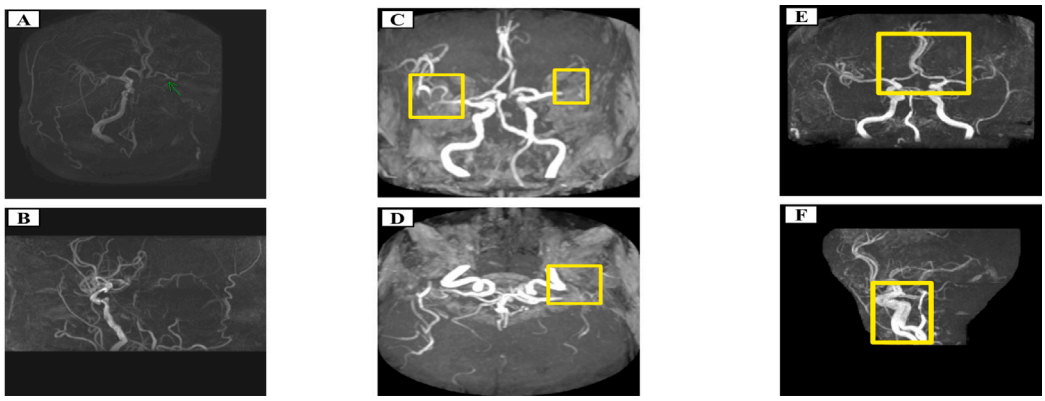
Intracranial atherosclerosis (ICAS), also referred to as intracranial arterial stenosis, is a major contributor to ischemic stroke and transient ischemic attack, accounting for approximately 9%–33% of global ischemic stroke cases (Shi et al., 2020). Accurate morphological assessment of intracranial arteries is therefore critical for diagnosis, risk stratification, and treatment planning. Automated segmentation of intracranial arteries (IA) from medical imaging has emerged as a key enabling step for visualization, extraction, and measurement of stenotic regions, which can subsequently serve as computational biomarkers for disease severity, treatment efficacy, and risk prediction.

Several imaging modalities are used for intracranial vascular assessment, including digital subtraction angiography (DSA), computed tomography angiography (CTA), and magnetic resonance angiography (MRA). While DSA and CTA provide high spatial resolution, their invasive nature, radiation exposure, and contrast dependency limit routine and longitudinal use (Tian et al., 2020). In contrast, time-of-flight

☆ This work is supported in part by the National Science and Technology Council (NSTC), Taiwan grant number 114-2221-E-182-022-MY3 and in part by the Chang Gung Memorial Hospital, Taiwan research grants number CORPG6L0151.

* Corresponding author at: Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan.
E-mail addresses: d0829005@cgu.edu.tw (M. Sarkar), deepblue@cgmh.org.tw (Y.-C. Huang), thlee@cgmh.org.tw (T.-H. Lee), jdlee540908@cgmh.org.tw (J.-D. Lee), pksahoo@mail.cgu.edu.tw (P.K. Sahoo).

https://doi.org/10.1016/j.mlwa.2026.100843
Received 2 November 2025; Received in revised form 9 January 2026; Accepted 11 January 2026
Available online 19 January 2026



(a) A-B represents varied visible intensity

(b) C-D depicts shadow-like effects around arteries

(c) E-F shows inter-wined and overlapped arteries

Fig. 1. Examples of TOF-MRA slices in different orientations.

magnetic resonance angiography (TOF-MRA) enables non-invasive visualization of intracranial vasculature without contrast agents, making it particularly suitable for screening and follow-up examinations. The modality also offers high spatial resolution, three-dimensional coverage, and fast acquisition, facilitating the mapping of both major arteries and smaller perforating branches (Ge et al., 2019; Min et al., 2024). However, TOF-MRA presents unique segmentation challenges due to flow-related signal variability, intensity inhomogeneity, vessel overlap, and limited boundary contrast, especially in maximum intensity projection (MIP) representations.

Despite these advantages, accurate segmentation of intracranial arteries in TOF-MRA remains an open challenge. Several intrinsic imaging characteristics complicate the task. First, intensity inhomogeneity, arising from variable blood flow velocity, slice orientation, and scanner parameter discrepancies, produces inconsistent contrast levels between and within slices (Deshpande et al., 2021). For instance, in Fig. 1(a), slice B appears substantially brighter than slice A due to localized flow acceleration. Second, signal similarities between vascular and parenchymal tissues (e.g., between small vessels and adjacent white matter) often lead to segmentation ambiguity. Third, shadow-like distortions induced by gray–white matter boundaries, as shown in Fig. 1(b), can obscure true vessel contours, complicating boundary extraction. Fourth, the presence of numerous fine, overlapping arteries, can mask or fragment the primary arterial trunk (Fu et al., 2020). Fifth, orientation-dependent morphological variability, for example, the tortuosity of the ICA across axial slices, Fig. 1(c), poses additional difficulties for segmentation algorithms. Finally, flow-related artifacts and motion-induced noise can mimic stenoses or occlusions, leading to diagnostic misinterpretations. Manual annotation under these imaging conditions is labor-intensive, time-consuming, and highly operator-dependent, limiting its practicality in both clinical and research settings. Consequently, there is a growing imperative to develop robust, automated IA segmentation algorithms capable of handling these confounding factors while maintaining clinical-grade accuracy.

Recent advances in deep learning (DL) have significantly improved vascular segmentation performance across modalities. Numerous architectures have been proposed for vessel and aneurysm segmentation, including nnUNet-based frameworks (Nageler et al., 2023), attention-driven networks, e.g., ARU-Net (Mu et al., 2023), artery-aware pipelines, e.g., AASeg (Yao et al., 2024), and multi-scale learning models, e.g., MGLIA-Net (Hou et al., 2025). While these approaches demonstrate strong task-specific performance, most studies focus on proposing new architectures or optimizing a single model on a specific dataset. As a result, existing literature offers limited insight into how different segmentation architectures interact with encoder backbones, particularly for ICAS-specific TOF-MRA, and how such design choices affect robustness and reproducibility.

In particular, DL-based encoder–decoder architectures such as U-Net (Ronneberger et al., 2015), LinkNet (Chaurasia & Culurciello, 2017), Feature Pyramid Network (FPN) (Martinsson & Mogren, 2019), Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017), and DeepLabV3+ (Chen, Papandreou, Schroff, & Adam, 2017) have significantly improved segmentation performance. The effectiveness of these models critically depends on the ability of their encoders to extract meaningful features. While these architectures include their own encoders, replacing them with pre-trained classification models, such as VGG (Simonyan & Zisserman, 2014), Residual Networks (ResNet) (He et al., 2016), DenseNet (Huang et al., 2017), and EfficientNet (Tan & Le, 2019), has been shown to further enhance segmentation accuracy (Abdelrahman & Viriri, 2023b; Hussain et al., 2025; Li & Xie, 2025; Sharma et al., 2023; Sulaiman et al., 2024). Pre-trained backbones efficiently capture hierarchical features, accelerate convergence, and improve accuracy, particularly in medical imaging tasks with limited training data (Abdelrahman & Viriri, 2023a; Rayed et al., 2024). Although such backbones have demonstrated strong performance across various segmentation tasks (Bhatti et al., 2025; Chen, Papandreou, Kokkinos, et al., 2017; Long et al., 2015; Singh et al., 2025), their relative effectiveness is highly task and data-dependent, and systematic comparisons of backbone–segmentation combinations for IA segmentation in TOF-MRA remain scarce.

Moreover, most prior studies emphasize peak performance metrics such as Dice similarity coefficient (DSC) or Intersection over Union (IoU), while largely overlooking performance stability and reproducibility. In clinical settings, segmentation models must exhibit consistent behavior across patients, acquisition conditions, and validation folds. Models that achieve high average accuracy but exhibit high variability may produce unreliable segmentations, limiting their translational potential. Despite its importance, stability-aware benchmarking has received little attention in IA segmentation research.

To address this gap, the present study systematically evaluates nine widely used classification architectures, VGG Net, ResNet, SE-ResNet, Residual Networks with Cardinality (ResNeXt), Squeeze-and-Excitation Residual Networks with Cardinality (SE-ResNeXt) (Hu et al., 2018), Squeeze-and-Excitation Networks (SENet) (Hu et al., 2018), DenseNet, Inception Net (Szegedy et al., 2015), EfficientNet (Tan & Le, 2019), as encoder backbones within five encoder–decoder based segmentation networks, UNet, LinkNet, FPN, PSPNet, and DeepLabV3+, resulting in 45 model variants. Each segmentation model was chosen to represent distinct architectural strategies: UNet for its symmetric skip connections that preserve fine-grained spatial information, LinkNet for residual connections enabling efficient encoder–decoder feature transfer, FPN for top-down multiscale feature fusion that captures vessels of varying sizes, and PSPNet for pyramid pooling to integrate global contextual information. Furthermore, the evaluation of

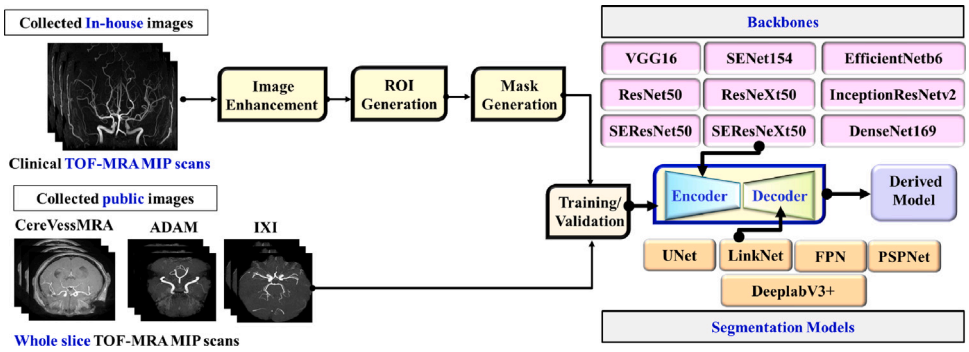


Fig. 2. Illustration of the adopted framework for automatic segmentation of ICAS from MIP sequence of TOF-MRA.

these 45 combinations are conducted across four heterogeneous TOF-MRA datasets, including an in-house clinical dataset and three public datasets (CereVessMRA (Guo et al., 2024), IXI (Information eXtraction from Images) (I.X.I. Project, 2025), and ADAM (Aneurysm Detection and Segmentation Challenge (Timmins et al., 2020)), enabling a multi-domain assessment of generalizability. In addition to conventional accuracy metrics, we introduce a coefficient-of-variation–based stability score to quantify segmentation consistency across cross-validation folds and test cohorts. By jointly analyzing accuracy, stability, and computational efficiency, this work provides practical guidance for selecting reliable backbone–segmentation model pairs for ICAS-oriented TOF-MRA segmentation.

The main contributions of this study are summarized below:

1. A systematic and controlled evaluation of 45 previously unexplored encoder–decoder combinations is conducted for IA segmentation in TOF-MRA. To the best of current knowledge, no prior study has jointly benchmarked segmentation architectures and backbone networks for this task.
2. A comprehensive benchmark is performed across segmentation accuracy, IoU, DSC, computational complexity, trainable parameters, and runtime efficiency, median inference time, enabling a precise characterization of performance–efficiency trade-offs among models.
3. The evaluation spans four datasets, an in-house TOF-MRA dataset and three public datasets (CereVessMRA, ADAM, IXI), providing a multi-domain assessment of generalizability across scanners, populations, and annotation protocols.
4. A Coefficient-of-Variation–based stability metric is introduced to quantify cross-fold and cross-subject variability in IoU and DSC, facilitating the identification of model configurations that exhibit both high accuracy and statistical robustness.

The rest of the paper is organized as follows. Section 2 describes the materials and methods, including TOF-MRA dataset acquisition, preprocessing steps, Region of Interest (ROI) generation, and mask creation. Section 3 outlines the training procedure, hyperparameters, and evaluation metrics. Section 4 presents the results and performance analysis of all backbone–segmentation model pairs, while Section 5 concludes with a summary of key findings.

2. Methods

This section describes the complete methodological pipeline adopted for intracranial arterial segmentation and cross-dataset evaluation. The primary objective of the study is to determine the most effective DL–based segmentation model for isolating major intracranial arteries from TOF–MRA. The analysis consists of two parallel components:

(i) a fully curated and ROI-specific segmentation pipeline developed exclusively for the in-house ICAS dataset, and

(ii) cross-dataset generalizability evaluation performed on three external TOF–MRA datasets (CereVessMRA (Guo et al., 2024), ADAM (Timmins et al., 2020)), and IXI (I.X.I. Project, 2025)) using their official annotations and preprocessing pipelines.

Unlike the in-house dataset, which required manual ROI definition, slice-level enhancement, and mask construction, the public datasets were used *as provided*, without additional annotation or modification. This separation ensures methodological transparency and preserves the integrity of each dataset. The overall framework is summarized in Fig. 2, and the key methodological components are outlined below.

- **Data acquisition:** Collection of TOF–MRA images from both clinical and public sources.
- **Image enhancement (in-house only):** Contrast improvement to accentuate ICAS-relevant arterial boundaries.
- **Automatic ROI generation (in-house only):** Delimiting clinically relevant arterial segments based on ICASMAP criteria, forming the basis for mask creation.
- **Mask generation (in-house only):** Constructing pixel-level ground-truth annotations for the predefined arterial regions.
- **Deep learning model training:** Training and cross-validating four state-of-the-art segmentation networks across all datasets.
- **Output generation:** Producing binary vessel masks along with visualization of segmentation results.

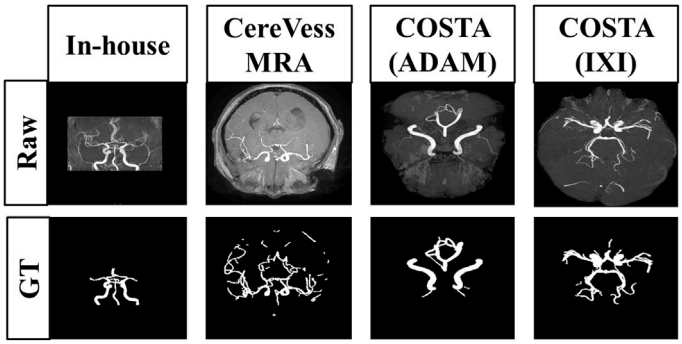
These stages are described in more detail below. The adopted framework has been schematically illustrated in Fig. 2.

2.1. Data acquisition

The primary dataset employed in this study consists of clinically acquired TOF–MRA MIP images specifically related to ICAS. ICAS can also be assessed using volumetric TOF–MRA, particularly whole-slice axial acquisitions. ue to the absence of publicly accessible ICAS-specific TOF–MRA datasets, publicly available aneurysm and healthy vessel datasets were incorporated, as the primary objective of this work is accurate arterial vessel segmentation rather than lesion-specific segmentation. Accordingly, the final dataset comprises two types of data sources: (i) in-house clinical TOF–MRA MIP images, and (ii) whole-slice TOF–MRA volumes obtained from three publicly available datasets: CereVessMRA (Guo et al., 2024), ADAM (Aneurysm Detection and Segmentation Challenge), and IXI (Information eXtraction from Images). The ADAM and IXI datasets were accessed through the COSTA platform (Mou et al., 2024), which also provided the corresponding ground-truth annotations. An overview of the dataset configuration is presented in Fig. 3.

The details of each dataset are provided below

- **In-house dataset:** The in-house dataset includes TOF–MRA MIP examinations from 63 patients collected at the Chang Gung Medical Foundation, Taipei, Taiwan. Ethical approval was granted



Total Patients	66	271	107	170
Training	~53	~195	~86	~122
Validation	~6	~49	~9	~14
Testing	7	27	12	34

Fig. 3. Graphical representation of different datasets along with their patient-level distribution during training, cross-validation and testing.

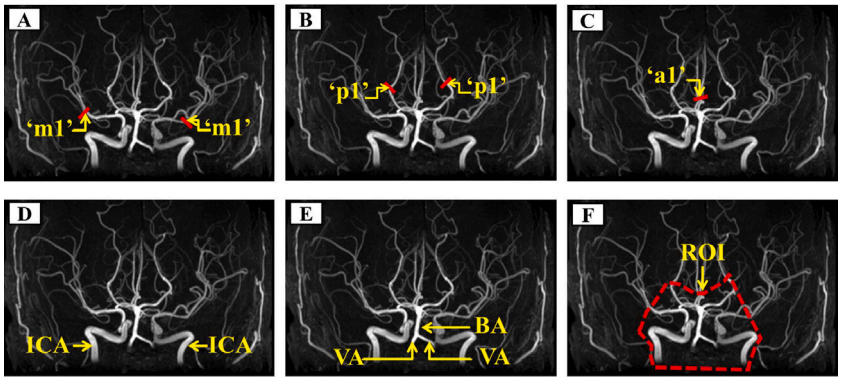


Fig. 4. Slices A–F illustrate the considered arterial regions. A–C show the M1 segment of MCA, P1 of PCA, and A1 of ACA, respectively. D–E depict ICA, BA, and VA, while F presents the complete ROI with all included arteries.

by the Institutional Review Board (IRB), license number 202402 052B0. All images were de-identified in accordance with IRB regulations, and the requirement for informed consent was waived. The dataset consists of 1003 axial 2D TOF–MRA clinical MIP projection slices (512 × 512 × 3) stored in JPEG format, acquired with a slice thickness of 5 mm. All annotations were manually generated for this study based on the arterial territories implicated in symptomatic ICAS, as defined by the multicenter ICASMAP study (Han et al., 2018). Accordingly, in-house ROI involves large cerebral arteries, including the ICA, M1 segment of the middle cerebral artery (MCA), A1 segment of the anterior cerebral artery (ACA), P1 segment of the posterior cerebral artery (PCA), vertebral arteries (VA), and the proximal basilar artery (BA), forming the ROI for this study. These ROI-specific annotations was generated for this in-house dataset only; no such ICAS-focused annotation exists in any public dataset. Due to variation in acquisition windowing, mild intensity enhancement was applied to improve vessel–background contrast (Fig. 4).

- **CereVessMRA dataset:** This dataset contains 271 manually annotated TOF–MRA volumes (150 healthy and 121 pathological), totaling 28,128 slices. Healthy cases were obtained from UK institutions within the UK-Hset cohort, while pathological cases originated from the DGH Hospital in Fujian Province, China. To standardize evaluation and facilitate comparison to the in-house MIP-based workflow, 20 representative maximum intensity projection (MIP) slices were generated for each subject, producing 5420 MIP images. The original annotations and preprocessing pipeline provided by the dataset were used without modification.
- **ADAM:** The ADAM dataset, sourced through the COSTA platform (Mou et al., 2024), comprises 107 TOF–MRA volumes acquired on a 3.0 T MRI scanner. Voxel spacing ranges from 0.20×0.20×0.40 mm^3 to 0.59×0.59×1.00 mm^3, with image sizes between 256 × 256 × 100 and 560 × 560 × 140. Of these subjects, 85 exhibit intracranial aneurysms, while 22 serve as healthy controls. Similar to the CereVessMRA processing strategy, 20 representative MIP projections were extracted for each subject, yielding 2140 MIP images. All ground-truth masks provided by COSTA were used directly without additional annotation.
- **IXI dataset**: The IXI dataset includes TOF–MRA scans from three UK institutions: IXI-Guys (Guy's Hospital), IXI-HH (Hammersmith Hospital), and IXI-IOP (Institute of Psychiatry). Using COSTA's standardized partitioning, the training set comprised 136 scans (48 IXI-Guys, 48 IXI-HH, 40 IXI-IOP), while 34 scans (12 IXI-Guys, 12 IXI-HH, 12 IXI-IOP) were reserved for testing. The native resolutions are as follows:
  - IXI-Guys: 512 × 512 × 100, voxel spacing 0.47×0.47×0.80 mm^3
  - IXI-HH: 512 × 512 × 100, voxel spacing 0.47×0.47×0.80 mm^3
  - IXI-IOP: 1024 × 1024 × 92, voxel spacing 0.26×0.26×0.80 mm^3

As with other public datasets, 20 MIP slices were generated per subject, producing 3400 projections in total. No additional preprocessing or annotation was performed beyond the official dataset pipeline.

To avoid slice-level data leakage and ensure fair evaluation, all training, cross-validation, testing was strictly performed at the patient level. Slices from the same patient were never allowed to appear simultaneously in the training, validation, or testing sets within any fold. The in-house dataset underwent 10-fold patient-wise cross-validation, while the CerevessMRA, ADAM and IXI datasets were evaluated using 5-fold patient-wise cross-validation. The official public test sets were kept untouched throughout all experiments.

2.2. Image enhancement

The in-house TOF-MRA slices exhibited noticeable intensity inhomogeneity due to variations in acquisition parameters, motivating the need for preprocessing to achieve consistent contrast throughout the dataset. Intensity enhancement was therefore applied to improve vessel–background separation and support downstream segmentation. Several commonly used enhancement methods were evaluated, including contrast stretching (CS), min–max scaling (MMS), max-absolute scaling (MAS), histogram equalization (HEq), gamma transformation (GT), and logarithmic transformation (LT).

For quantitative comparison of enhancement approaches, the original unprocessed slice $I(x, y)$ was used as the reference image when computing similarity-based measures, such as, mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) (Sara et al., 2019). These metrics were not used as indicators of diagnostic improvement; rather, they were used to ensure that enhancement did not alter or distort arterial morphology. Thus, lower MSE and higher PSNR/SSIM were interpreted as evidence of structural fidelity with respect to the original anatomy. Because similarity metrics alone do not directly quantify diagnostic relevance, qualitative assessment of vessel conspicuity was also conducted. Representative slices were visually examined to assess clarity of major intracranial arteries (ICA, MCA, ACA, PCA, BA, VA) with attention to lumen definition, contrast uniformity, and background suppression. This combined quantitative–qualitative evaluation ensured that the selected enhancement method preserved anatomical structure while improving visibility of vascular features.

Each enhancement technique was applied to all slices, and performance was assessed using MSE, PSNR, and SSIM. As summarized in Table 1, GT with $\gamma = 0.9$ yielded the best results, achieving the lowest MSE of 28.658, highest PSNR of 34.673, and highest SSIM of 0.9906, indicating optimal brightness and contrast. GT enhances each pixel of the normalized input image $I(x, y)$ on a pixel-by-pixel basis using a power-law transformation:

$$\hat{I}(x, y) = \alpha \cdot I(x, y)^{\gamma} \tag{1}$$

Here, $\hat{I}(x, y)$ is the enhanced output, α controls brightness, and γ adjusts contrast. In this study, α was set to 1, and $\gamma < 1$ was chosen to produce brighter images, as values $\gamma > 1$ tend to darken the output (Pedregosa et al., 2011; Rahman et al., 2016).

As illustrated in Fig. 5, varying γ produces markedly different visual characteristics:

- $\gamma = 0.1$ to 0.3: excessive brightening, loss of vessel boundaries, background washout.
- $\gamma = 0.4$ to 0.6: moderate enhancement but inconsistent contrast across thin vessels.
- $\gamma = 0.7$ to 0.8: improved arterial visibility but occasional oversharpening and lumen non-uniformity.
- $\gamma = 0.9$: optimal balance; clear delineation of intracranial arteries, improved vessel–background separation, and excellent preservation of morphology

Table 1
Performance of enhancement method with respect to image quality assessment metrics.

Enhancement methods		Image quality assessment metrics		
		MSE	PSNR	SSIM
CS		155.387	28.039	0.9218
MMS		565.338	22.0213	0.8522
MAS		825.256	21.667	0.8738
HEq		317.379	24.0431	0.3426
GT	γ=0.1	13672.0	7.567	0.5873
	γ=0.5	1631.67	16.9948	0.8073
	γ=0.9	**28.658**	**34.673**	**0.9906**
LT	log=0.5	510.751	21.4228	0.8940
	log=1.0	108.014	28.964	0.9780
	log=1.4	1496.49	16.945	0.8305

Quantitatively, $\gamma = 0.9$ produced the lowest MSE of 14.96 and highest PSNR of 36.38 among all tested values, indicating minimal structural distortion. Visual inspection further confirmed consistent enhancement of both large and small arterial segments without amplifying background noise.

2.3. Automated ROI generation

The automated ROI generation pipeline was developed to isolate clinically relevant intracranial arterial regions while suppressing non-informative background structures. This process, applied exclusively to the in-house dataset, consists of two sequential steps: (1) radiologist-guided polygonal cropping, followed by (2) automated bounding-box–based normalization. The full workflow is illustrated in Fig. 6.

2.3.1. Radiologist-guided polygonal cropping

Initial arterial demarcation was performed using polygonal cropping on the enhanced TOF-MRA image $\hat{I}(x, y)$ obtained from Section 2.2. Two board-certified neuroradiologists, each with more than 8 years of experience, independently delineated polygonal regions encompassing the major ICAS-relevant arteries (ICA, MCA–M1, ACA–A1, PCA–P1, BA, and VA–V4), following anatomical criteria described in Sartoretti et al. (2020). All polygon masks were created using the open-source *polygon-crop 0.0.3* tool (Chiang, 2020).

A slice-wise annotation strategy was adopted to ensure that subtle morphological variations across axial slices were accurately captured. To maintain consistency, a three-expert consensus protocol was implemented. Initially, two neuroradiologists independently annotated a subset of cases, and any discrepancies were resolved through joint discussion with a third senior neuroradiologist. Following this calibration phase, the remaining slices were annotated collaboratively by two raters, with every annotation subsequently reviewed and verified by the third rater for quality control. The resulting polygon-cropped image $\hat{I}_{p_c}(x, y)$ preserved the original spatial dimensions of $\hat{I}(x, y)$ while setting all non-ROI regions to zero intensity.

While this step effectively isolates vessel-rich regions, the resulting images often contained large zero-valued background areas. This introduced foreground–background imbalance which could negatively affect subsequent model training, motivating the second stage of automated normalization.

2.3.2. Automated bounding-box cropping

To reduce redundant background while preserving all vascular structures, an automated cropping stage was implemented using OpenCV 4.2.0. For every polygon-cropped slice, the software identified the smallest bounding box containing all nonzero pixels. These per-slice bounding boxes were collected across all patients and aggregated to determine the *union bounding region* that consistently enclosed ICAS-relevant arterial segments for the entire in-house dataset.

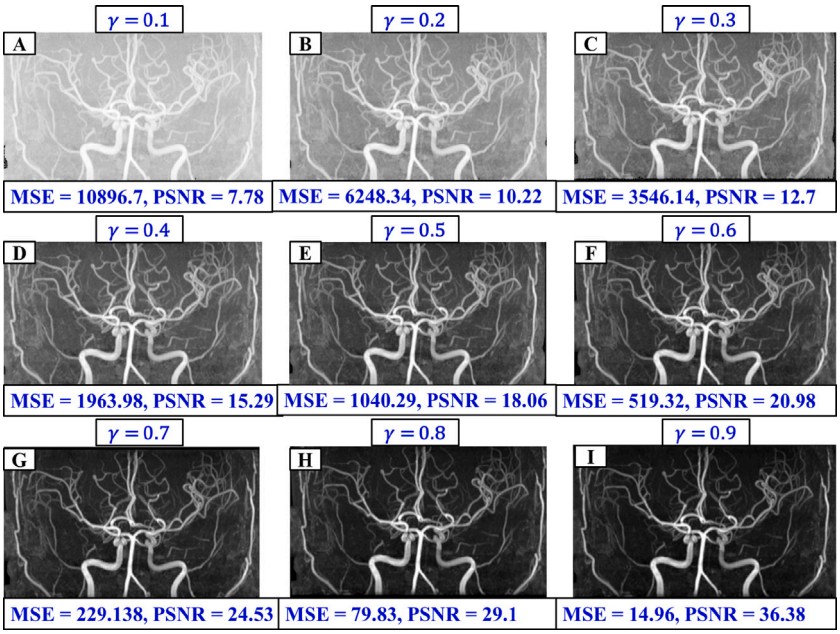


Fig. 5. Effect of gamma transformation on vessel visibility for a representative TOF-MRA slice.

By analyzing the union of all bounding boxes across 1003 slices, we empirically observed that all polygon-defined arterial regions fell within a stable spatial zone. This led to the determination of fixed cropping coordinates: (100:350) along the y-axis and (150:350) along the x-axis. These coordinates represent the minimal rectangular region that fully contained the anatomical ROIs for *all* subjects in our in-house cohort during visual and quantitative verification. Applying this global crop resulted in the final ROI-normalized image, $\hat{I}_{A_c}(x, y)$, ensuring uniform input dimensions and facilitating efficient batch processing during model training. However, it is important to note that these coordinates are in-house dataset-specific and may not generalize to other TOF-MRA acquisitions, scanner geometries, or head orientations.

The visual summary of the preprocessing and ROI generation workflow is presented in Fig. 6. Row 1 and Row 2 represent anteroposterior (AP) and lateral TOF-MRA views, respectively. Column 1 shows the enhanced image $\hat{I}(x, y)$ obtained after gamma transformation ($\gamma = 0.9$), Column 2 depicts the polygonally cropped output $\hat{I_{p_c}}(x, y)$, and Column 3 displays the final automatically cropped ROI $\hat{I_{A_c}}(x, y)$ used for mask generation and model training.

2.4. Mask generation

For IA segmentation, the task is binary: distinguishing the ROI from the background. Accordingly, binary masks were generated where *class 0* represents the background and *class 1* corresponds to the ROI (Fig. 4). The process of generating these masks is a crucial step, as they serve as ground truth labels for model training, validation, and quantitative evaluation of segmentation accuracy.

To achieve consistent and anatomically accurate annotations, the Medical Image Processing, Analysis, and Visualization (MIPAV) software (version 10.0.0) was employed. MIPAV provides robust semi-automated segmentation tools with pixel-level precision, allowing radiologists and imaging experts to refine the ROI boundaries interactively while maintaining reproducibility across samples. This semi-automated approach reduces manual bias, enhances efficiency, and ensures uniformity in labeling across all TOF-MRA slices.

Each binary mask, denoted as $B\hat{I}_{A_c}(x, y)$, corresponds directly to the automatically cropped input image $\hat{I_{A_c}}(x, y)$ obtained from Section 2.3. The generated masks delineate only the arterial structures, with all non-ROI regions assigned zero intensity values. These curated masks form the foundational dataset for evaluating segmentation model performance through overlap-based metrics such as Dice coefficient, IoU, and pixel-wise accuracy. Representative examples of the binary masks alongside their respective input slices are illustrated in Fig. 7.

2.5. Deep neural network architectures

Convolutional neural networks (CNNs) have consistently demonstrated superior performance across a wide range of computer vision tasks, including vascular segmentation. Numerous segmentation architectures have been proposed, each tailored to specific applications and data characteristics, which makes the selection of an appropriate model for IA segmentation non-trivial. To systematically evaluate their suitability, a comparative analysis was conducted using several widely adopted deep learning-based encoder–decoder semantic segmentation architectures, including UNet, PSPNet, LinkNet, FPN, and DeepLabV3+.

Segmentation performance is strongly influenced by the choice of encoder, as modern decoders commonly benefit from feature representations learned by encoders pre-trained on large-scale classification datasets. Consequently, examining a diverse set of encoder architectures becomes essential. General representation of encoder families and the corresponding decoder structures are schematically expressed in Fig. 8. In typical hierarchical CNN design, the spatial resolution of feature maps is progressively reduced by factors of $1/2, 1/4, 1/8, 1/16, 1/32$ across five stages, which are denoted as Stage 1 through Stage 5, while the feature dimensionality increases accordingly.

All TOF-MRA slices are single-channel images. To ensure compatibility with ImageNet-pretrained encoders, each slice is replicated across three channels (1 to 3), followed by backbone-specific preprocessing, mean-subtraction and scaling, using the $get\_preprocessing\_fn$ function from the $segmentation\_models$ (Yakubovskiy, 2019) framework. This ensures consistent input normalization and preserves the integrity of pretrained feature statistics. Besides, batch-normalization layers included in the pretrained encoders remain trainable for all experiments. This configuration allows the normalization statistics to adapt to the target medical domain while maintaining stable initialization. Decoder blocks also retain batch-normalization layers as implemented in the standard architectures within the library.

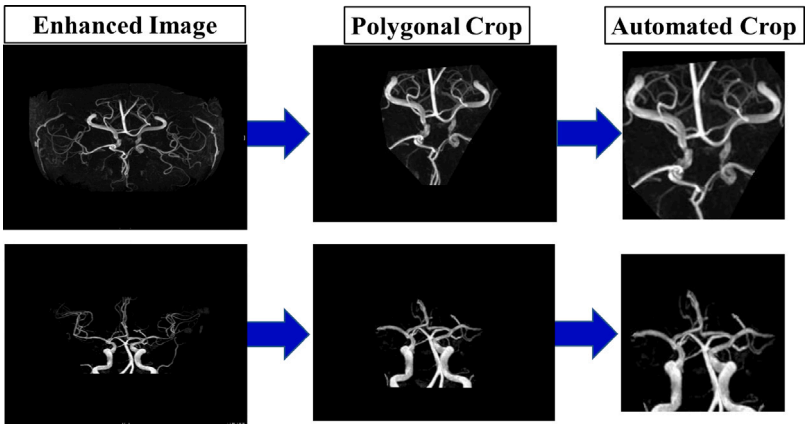
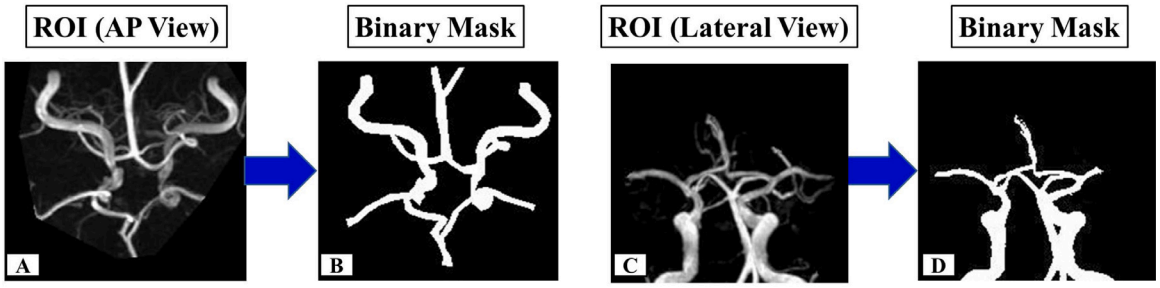


Fig. 6. Visual output of automated ROI generation phase where row 1 contains AP view slices and row 2 contains Lateral view slices. Here column 1 describes the enhanced raw TOF-MRA slices. Column 2 and column 3 descries the output of polygonal cropping and automated cropping respectively.



(a) B represents the binary mask of enhanced image A (b) D represents the binary mask of enhanced image C

Fig. 7. Examples of mask generated from respective ROI slices using MIPAV.

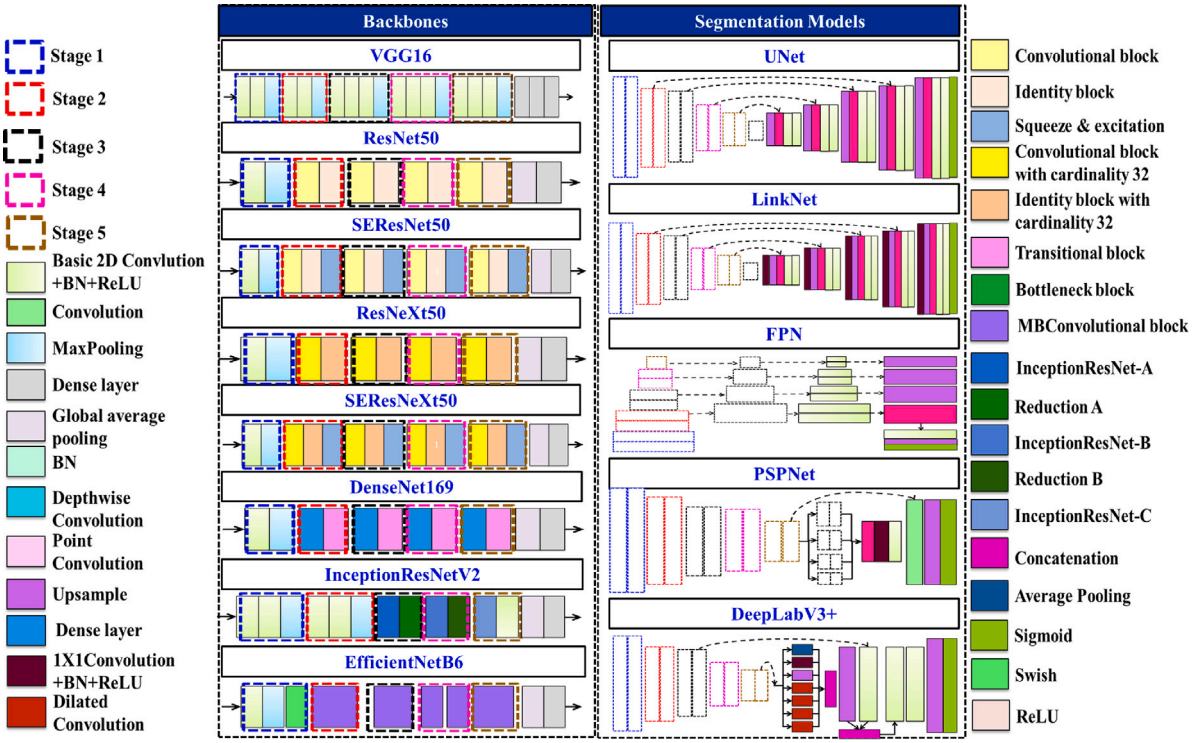


Fig. 8. Visual representation of different backbones (left panel) and segmentation models (right panel) used in this study.

Table 2
Summary of backbone networks used for segmentation.

Backbone	Depth	Key features	Distinct characteristics
vgg16	16 layers	Stacked 3 × 3 convolutions	Simple, sequential
resnet50	50 layers	Residual (skip) connections	Mitigates vanishing gradients
seresnet50	50 layers	Residual + SE blocks	Channel-wise attention, emphasizes informative features
resnext50	50 layers	Grouped convolutions	Increases feature diversity efficiently, higher capacity than ResNet
seresnext50	50 layers	Grouped convolutions + SE blocks	Combines diverse feature learning with adaptive channel weighting
densenet169	169 layers	Dense connectivity	Feature reuse, improved gradient flow, efficient parameter usage
senet154	154 layers	Deep network + SE blocks	Channel-wise attention at multiple stages, high capacity
irv2	152 layers	Inception modules + residual connections	Multi-scale feature extraction with residual learning
efficientnetb6	66 layers	Compound scaling (depth, width, resolution)+MBConv	High accuracy with balanced efficiency, optimized for both performance and computation

2.5.1. Backbone architecture

Backbone networks serve as the primary feature extractors within modern segmentation architectures. These networks are typically pre-trained on large-scale image datasets, enabling segmentation models to leverage well-established low- and mid-level representations through transfer learning. Such initialization is particularly advantageous in medical image segmentation, where annotated datasets are often limited, as it accelerates convergence and improves overall representational quality. Each backbone architecture embodies distinct design choices, including depth, connectivity patterns, scaling strategies, and computational efficiency, that directly influence segmentation performance (see Table 2). For instance, VGG16 (*vgg16*) provides a classical sequential baseline, characterized by stacked $3X3$ convolutional layer. Two successive $3X3$ convolutions achieve a receptive field comparable to a single $5X5$ convolution while requiring fewer parameters, making the architecture computationally efficient despite its simplicity.

Residual networks such as ResNet-50 (*resnet50*) introduce identity skip connections that alleviate the vanishing-gradient problem and facilitate feature reuse across layers. ResNeXt-50 (*resnext50*) extends this design by incorporating grouped convolutions; the use of cardinality (i.e., multiple parallel transformation paths) enhances representational capacity without proportional increases in complexity. The integration of squeeze-and-excitation (SE) blocks further augments feature discrimination by applying channel-wise recalibration through learned nonlinear interactions. This mechanism underpins architectures such as SEResNet-50 (*seresnet50*), SEResNeXt-50 (*seresnext50*), and the deeper SE-enhanced variant SENet-154 (*senet154*).

DenseNet-169 (*densenet169*) adopts dense connectivity wherein each layer receives feature maps from all preceding layers via channel-wise concatenation. This facilitates feature reuse, strengthens gradient propagation, and provides a highly parameter-efficient architecture. InceptionResNet-V2 (*irv2*) combines the multi-scale feature extraction capabilities of Inception modules, using parallel convolutional branches of varying kernel sizes, with residual connections that stabilize gradient flow. This hybrid design supports the training of very deep networks while maintaining strong multi-scale contextual encoding.

Finally, EfficientNet-B6 (*efficientnetb6*) employs compound scaling, uniformly adjusting network depth, width, and input resolution based on a principled scaling coefficient. Combined with mobile inverted bottleneck convolution (MBConv) blocks, this architecture achieves a favorable balance between accuracy, computational cost, and memory efficiency, making it suitable for high-resolution medical imaging tasks.

2.5.2. Decoder architecture

The decoder component of a semantic segmentation network is responsible for progressively reconstructing a high-resolution output from the compressed feature representations generated by the encoder. Its primary role is to recover spatial detail, refine object boundaries, and integrate multi-scale contextual cues to enable accurate pixel-wise classification. Typical decoder designs employ a combination of upsampling operations, convolutional refinement, and feature fusion strategies, often incorporating skip connections—to effectively restore fine-grained structural information while preserving the semantic richness learned at deeper layers (see Table 3).

For instance, the most popular medical image segmentation architecture, UNet, reconstructs high-resolution segmentation maps through a sequence of upsampling stages which involves bilinear upsampling or transposed convolution to double spatial resolution, followed by concatenation with encoder skip features, and two successive 3 × 3 convolutions + ReLU. This design preserves fine-grained anatomical boundaries and enables precise localization. LinkNet, on the other hand, employs a lightweight, efficient decoder designed around residual blocks. For every encoder stage, the decoder applies 1 × 1 convolution for channel reduction,followed by bilinear upsampling, and residual refinement block. Different from UNet, here skip connections from encoded are additive in nature rather than concatenative as seen in case of Unet, reducing computational load and memory usage while maintaining spatial detail.

FPN introduces a top-down pathway merged with lateral connections. High-level, semantically strong features are progressively upsampled and summed with spatially detailed lower-level features. Each pyramid level undergoes 1 × 1 lateral convolution, followed by upsampling, and 3 × 3 smoothing convolution. This produces a multi-scale feature pyramid effective for segmenting structures of varying sizes. PSPNet uses a pyramid pooling module (PPM) to aggregate global contextual information. The decoder pools features at multiple spatial scales (e.g., 1 × 1, 2 × 2, 3 × 3, 6 × 6 bins), applies 1 × 1 convolutions, upsamples each branch to the original size, and concatenates them. A final 3 × 3 convolution refines fused features. This allows the model to capture long-range dependencies and contextual relationships.

DeepLabV3+ integrates atrous (dilated) convolutions inside an Atrous Spatial Pyramid Pooling (ASPP) module, followed by a lightweight decoder. The decoder upsamples ASPP output by 4×, concatenates it with low-level encoder features (after 1 × 1 convolution to reduce channels), followed by application of two 3 × 3 convolutions and final upsampling. The combination of atrous convolution and multi-scale context yields robust segmentation in complex anatomical structures. Unlike UNet, LinkNet, DeepLabV3+ does not employ skip connections from all encoder stages, providing a computationally efficient yet effective decoder for complex anatomical structures. Additional decoder configurations are explained in Table A.12 in Appendix. Moreover, it is important to note that, in order to maintain consistency with the backbone skip connection of other segmentation models, low-level and high-level skip connection definitions used in DeepLabV3+ following the segmentation_models (Yakubovskiy, 2019) implementation as described in Table 4[1]

Table 3
Comparison of Decoder Architectures in Popular Segmentation Models.

Model	Decoder structure	Core components	Distinct characteristics
UNet	5 stages	Upsampling; concatenation-based skip connections	High-resolution recovery through dense skip connections; strong localization capability.
LinkNet	5 stages	1 × 1 conv for channel reduction; bilinear upsampling; residual refinement blocks; additive skip connections	Residual decoding improves gradient flow and reduces computational cost.
FPN	4 pyramid levels	Top-down pathway; 1 × 1 lateral convolutions; upsampling	Multi-scale pyramid representation; robust performance across object sizes; efficient feature fusion.
PSPNet	Pyramid pooling module + final refinement	Pyramid pooling (1 × 1, 2 × 2, 3 × 3, 6 × 6 bins); upsampling	Captures global context; models long-range dependencies; effective for large structural segmentation.
DeepLabV3+	1 lightweight decoder block	ASPP module; 1 × 1 low-level & high-level feature projection; progressive upsampling	Combines atrous convolutions with multi-scale context; precise boundary refinement with fewer parameters.

Table 4
Low-level and high-level skip connection definitions used in DeepLabV3+ models.

Backbone	Low-level Layer	High-level Layer
vgg16	block3_conv3	block5_conv3
resnet50	stage2_unit1_relu1	stage4_unit1_relu1
seresnet50	stage2_unit1_relu1	stage4_unit1_relu1
resnext50	62	246
seresnext50	254	1078
senet154	454	6884
densenet169	51	367
irv2	16	594
efficientnetb6	block2a_expand_activation	block4a_expand_activation

2.5.3. Loss functions

All the encoder–decoder networks were computed against dice loss (DLoss). It was noticed that DLoss (Eelbode et al., 2020) was more popular in the case of semantic segmentation. DLoss is designed to maximize the overlap between the predicted segmentation P and the ground truth G. To improve numerical stability, a small constant ϵ is added:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{2}$$

Here, p_i and g_i represent the predicted and ground truth values at pixel i, respectively.

3. Experimental design

3.1. Data augmentation

During the training phase, data augmentation technique was implemented using Tensorflows' ImageDataGenerator with the rotations ranging from 10° to 40° in the axial plane, zooming by 0.05, width and height shifts of 0.05, shearing by 0.05, and both horizontal and vertical flipping set to True. All augmentations were applied stochastically. For each batch, the generator randomly samples whether to apply each transformation and with what magnitude, resulting in a unique augmented realization of every training image. The same augmentation pipeline and training setup were applied uniformly across all backbone architectures and segmentation models to ensure fair comparison. Augmentation hyperparameters were selected during an initial tuning stage and subsequently fixed for all experiments. This consistent design avoids model-specific bias and stabilizes the evaluation of robustness and segmentation performance across architectures.

[1] integer indices correspond to layer positions in the segmentation_models implementation.

3.2. Performance evaluation metrics

The performance of the segmentation models across various backbone was evaluated both quantitatively and statistically to ensure comprehensive assessment. Statistical evaluation was performed through cross-validation-based stability analysis.

3.2.1. Quantitative evaluation

Quantitative evaluation employed IoU, DSC, as primary metric, and secondary metric involved MCC, Precision, and Recall. DSC, equivalent to the F1-score for binary segmentation, reflects the harmonic mean of Precision and Recall. Precision and Recall, measure the accuracy of positive predictions and the completeness of detected positives, respectively. IoU assesses the overlap between predicted and ground truth segmentations. MCC is crucial due to class imbalance, as it incorporates all confusion matrix elements and remains robust when background dominates. Metrics like pixel accuracy and specificity were excluded due to their sensitivity to class imbalance, as noted in Jun et al. (2020).

In this binary segmentation task, artery pixels were classified as the positive class and background pixels as the negative class. True positives (TP) were correctly predicted artery pixels, while false positives (FP) were incorrectly predicted ones. False negatives (FN) were background mistaken for artery, and true negatives (TN) were correctly predicted background pixels.

3.2.2. Stability analysis

While metrics such as IoU and DSC assess segmentation accuracy, they do not capture the consistency of model performance across folds. A model may achieve high average scores yet exhibit substantial variability, which can compromise reliability and reproducibility, key considerations for clinical deployment. Although standard deviation (S.D) quantifies variability, it does not account for differences in mean performance; two models with similar S.Ds may have very different relative stability depending on their average IoU or DSC.

To address this, the coefficient of variation (CV) (Everitt & Skrondal, 2010) was calculated for IoU and DSC across 10-fold cross-validation. CV normalizes the S.D by the mean, providing a relative measure of dispersion that allows fair comparison across models with differing average performance levels. To summarize overall consistency in a single metric, a composite *stability score* was defined as:

$$\text{Stability Score} = \frac{IoU_{CV} + DSC_{CV}}{2} \tag{3}$$

where IoU_{CV} and DSC_{CV} are the CV for the IoU and DSC, respectively. Lower stability scores indicate more consistent and reproducible segmentation, which is crucial for reliable clinical applications. While high mean IoU and DSC reflect segmentation accuracy, the stability score provides a normalized and comprehensive measure of model robustness across different data splits, ensuring confidence in deployment for critical medical tasks.

Table 5
Network complexity of the segmentation models with different backbones.

Backbones	Models									
	LinkNet		FPN		PSPNet		UNet		DeepLabV3+	
	Params (10^6)	Time (sec)	Params (10^6)	Time (sec)	Params (10^6)	Time (sec)	Params (10^6)	Time (sec)	Params (10^6)	Time (sec)
vgg16	20.33	1.16	17.58	1.31	10.01	0.92	23.76	1.22	18.39	1.131
resnet50	28.79	0.47	26.92	0.43	3.83	0.22	32.57	0.51	12.38	0.309
seresnet50	31.33	0.28	29.46	0.33	3.98	0.13	35.11	0.32	13.34	0.268
resnext50	28.29	0.99	26.42	0.99	4.07	0.42	32.07	0.99	13.34	0.751
seresnext50	30.82	1.07	28.95	1.08	3.96	0.41	34.6	1.06	13.24	0.744
senet154	118.54	6.32	116.67	6.27	7.57	1.37	122.33	6.51	77.86	5.81
densenet169	15.63	1.35	15.72	1.26	3.17	0.05	19.52	1.31	10.35	1.03
irv2	57.87	1.57	57.53	1.01	3.56	0.39	62.07	1.11	30.72	0.865
efficientnetb6	47.55	1.23	44.45	1.02	**2.88**	**0.41**	50.91	1.12	4.36	1.246

Table 6
Performance comparison on the in-house dataset. Results are reported as mean and S.D across cross-validation folds and the held-out test dataset. The best segmentation performance with the highest DSC and the most stable model with the lowest stability score are highlighted.

Model	Best backbone	Cross-validation dataset					Test dataset				
		IoU	± S.D	DSC	±S.D	Stability	IoU	± S.D	DSC	± S.D	Stability
LinkNet	***efficientnetb6***	0.811	0.013	0.891	0.009	**1.24**	0.799	0.066	0.884	0.046	**6.68**
FPN	*seresnext50*	0.823	0.021	0.899	0.012	1.93	0.780	0.065	0.870	0.047	6.80
PSPNet	*efficientnetb6*	0.680	0.019	0.806	0.015	2.26	0.665	0.058	0.794	0.045	7.10
UNet	***seresnext50***	**0.874**	0.046	**0.930**	0.026	4.01	**0.832**	0.079	**0.901**	0.056	7.79
DeepLabV3+	*senet154*	0.764	0.015	0.769	0.016	1.97	0.709	0.066	0.713	0.050	8.05

4. Results and discussions

This section summarizes the segmentation model complexity in terms of parameter count and median test time, followed by the performance metrics given in Section 3.2 and visualization outputs are illustrated with individual dataset analysis.

This section presents a comprehensive evaluation of the IA segmentation models across multiple datasets. Performance was assessed using 10-fold cross-validation on the in-house TOF-MRA dataset, and 5-fold cross validation on CereVessMRA, IXI and ADAM datasets, followed by evaluation on a held-out test set. It is to be noted that, for the IXI and ADAM held-out dataset followed the test dataset division provided by COSTA. Segmentation quality was quantified using overlap-based metrics, while robustness was assessed through variability analysis.

4.1. Complexity analysis

The computational complexity of the segmentation models, including total parameter count and inference time per image, is an important consideration for deployment, particularly in resource-constrained environments such as portable devices, laptops, or emergency medical equipment. While shorter inference times facilitate faster clinical decision-making, maintaining robust segmentation accuracy remains the primary priority in medical applications.

Table 5 summarizes the model complexities, indicating that PSPNet exhibits the lowest parameter count and inference time (2.88 million parameters and 78 ms, respectively). Despite its efficiency, model selection in clinical contexts is ultimately guided by segmentation performance and reliability, which are assessed through quantitative metrics discussed in Sections 4.2–4.5.

4.2. Evaluation of segmentation models on in-house dataset

Following common practice in medical image segmentation studies, model performance was analyzed using mean ±S.D across cross-validation folds. In addition, robustness was assessed using descriptive statistical measures, including CV for stability analysis and 95% confidence intervals (CI) for DSC, rather than hypothesis testing. The segmentation performance of all models on both the cross-validation and held-out test datasets is summarized in Table 6.

Among the architectures evaluated, UNet with *seresnext50* achieved the highest segmentation accuracy, with a mean DSC of 0.930 (IoU = 0.874) and a 95% CI of 0.914–0.946. In contrast, LinkNet with *efficientnetb6* demonstrated the lowest variability across folds and test predictions, reflected by the lowest stability scores of 1.24 in Cross-validation and 6.68 in Test, indicating more consistent performance. For LinkNet, the DSC 95% CI was 0.885–0.896. The visualization result of these models are illustrated in Fig. 9.

4.3. Evaluation of segmentation models on CereVessMRA dataset

Similar to the in-house dataset, the performance of the segmentation models was evaluated using the same metrics under a five-fold cross-validation setup and is reported in Table 7. Robustness was assessed using the CV as a stability measure, and 95% CI were computed for DSC.

Among the evaluated architectures, FPN with *senet154* achieved the highest segmentation accuracy, with a mean DSC of 0.944, mean IoU of 0.895 and a 95% CI of 0.936–0.951 on DSC across cross-validation. On Test dataset, FPN with *senet154* achieved a mean DSC of 0.943 with a 95% CI of 0.925–0.958 and mean IoU of 0.894. In contrast, LinkNet with *senet154* demonstrated the lowest variability across folds and test predictions, with the lowest stability scores of 0.035 in Cross-validation and 0.177 in Test, indicating more consistent performance. The DSC 95% CI for LinkNet was 0.912–0.913 in Cross-validation and 0.899–0.903 in Test dataset respectively. It is also important to note that FPN with *senet154* maintains similar mean results across cross-validation and test datasets (0.944 vs 0.943) although its stability score is higher compared to LinkNet, reflecting a trade-off between peak segmentation accuracy and robustness. The visualization result of these models are illustrated in Fig. 10.

4.4. Evaluation of segmentation models on IXI dataset

The performance of the segmentation models was evaluated using the same metrics under a five-fold cross-validation setup and is reported

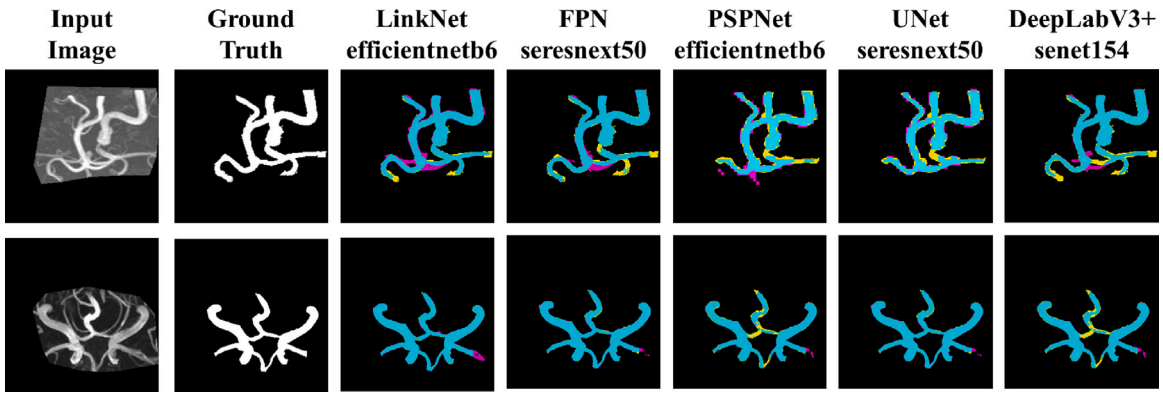


Fig. 9. Visual representation of the best segmentation model-backbone combination in overlapped form for in-house dataset. Here 'cyan' indicates TP, 'magenta' indicates FN, 'yellow' highlights the FP.

Table 7
Performance comparison on the CerevessMRA dataset. Results are reported as mean and S.D across cross-validation folds and the held-out test dataset. The best segmentation performance with the highest DSC and the most stable model with the lowest stability score are highlighted.

Model	Backbone	Cross-validation dataset					Test dataset				
		IoU	±S.D	DSC	±S.D	Stability	IoU	±S.D	DSC	±S.D	Stability
LinkNet	***senet154***	0.837	0.001	0.912	0.001	**0.035**	0.820	0.002	0.901	0.002	**0.177**
FPN	***senet154***	**0.895**	0.017	**0.944**	0.010	1.448	**0.894**	0.038	**0.943**	0.022	3.199
PSPNet	*vgg16*	0.675	0.003	0.804	0.002	0.368	0.630	0.002	0.770	0.001	0.179
UNet	*senet154*	0.835	0.010	0.910	0.006	0.854	0.834	0.010	0.910	0.006	0.849
DeepLabV3+	*efficientnetb6*	0.821	0.004	0.901	0.003	0.354	0.753	0.004	0.820	0.002	0.349

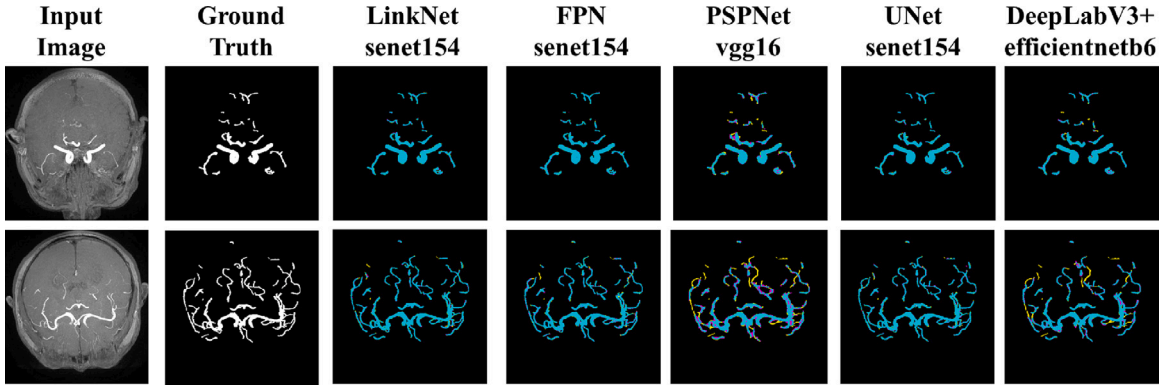


Fig. 10. Visual representation of the best segmentation model-backbone combination in overlapped form for CereVessMRA. Here 'cyan' indicates true positive, 'magenta' indicates false negative, 'yellow' highlights the false positives.

in Table 8. Robustness was assessed using the CV as a stability measure, and 95% CI were computed for DSC.

Cross-validation results show that FPN with *denenet169* achieved the highest segmentation accuracy, with a mean DSC of 0.968 (mean IoU = 0.938) and a 95% CI of 0.964–0.972. UNet and LinkNet achieved similar high DSCs of 0.960 and 0.959, respectively. On the held-out test dataset, LinkNet and UNet exhibited comparable DSCs of 0.907–0.908, mean IoU of 0.830–0.832, while FPN achieved 0.908, mean IoU of 0.832.

In terms of stability, LinkNet and UNet had the lowest stability score of 0.036 on the test set, indicating more consistent predictions despite the larger CV-to-test performance gap. FPN, while achieving the highest DSC in cross-validation, showed higher variability with stability score of 0.059 on the test dataset. PSPNet and DeepLabV3+ demonstrated both lower accuracy and moderate stability.

These results highlight that the trade-off between segmentation accuracy and stability is dataset-dependent. On IXI, high DSC in cross-validation does not necessarily translate to low variability on the test set, emphasizing the importance of stability analysis alongside accuracy metrics. The visualization result of these models are illustrated in Fig. 11.

4.5. Evaluation of segmentation models on ADAM dataset

The segmentation performance of the models on the ADAM dataset was evaluated using the same metrics under a five-fold cross-validation setup, with results reported in Table 9. Robustness was assessed using the CV as a stability measure, and 95% CI were computed for DSC.

Cross-validation results show that FPN with *densenet169* achieved the highest segmentation accuracy with DSC of 0.964, IoU = 0.931, and 95% CI of DSC as 0.953–0.954, followed closely by UNet with DSC of 0.956 and LinkNet with DSC of 0.953. On the held-out test dataset, both LinkNet and UNet achieved the same DSC of 0.909 and IoU of 0.834, with small reductions from cross-validation, indicating slight generalization gaps.

In terms of stability, UNet exhibited the lowest variability on the test dataset a stability score of 0.112, followed by LinkNet with stability

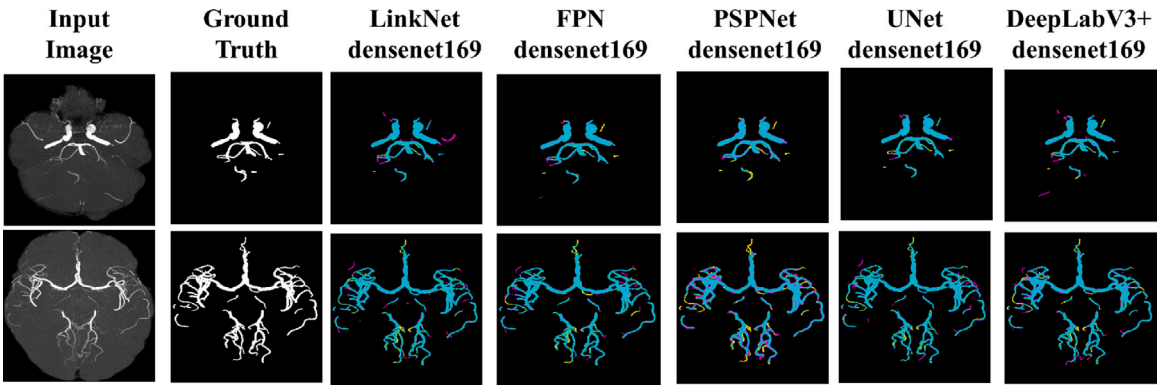


Fig. 11. Visual representation of the best segmentation model-backbone combination in overlapped form for IXI dataset. Here 'cyan' indicates TP, 'magenta' indicates FN, 'yellow' highlights the FP.

Table 8
Performance comparison on the IXI dataset. Results are reported as mean ±S.D across cross-validation folds and the held-out test dataset. The best segmentation performance with the highest DSC and the most stable model with the lowest stability score are highlighted.

Model	Backbone	Cross-validation folds					Test dataset				
		IoU	± S.D	DSC	±S.D	Stability	IoU	±S.D	DSC	±S.D	Stability
LinkNet	***densenet169***	0.921	0.002	0.959	0.001	0.108	0.830	0.001	0.907	0.001	0.036
FPN	***densenet169***	**0.938**	0.003	**0.968**	0.002	0.211	**0.832**	0.001	**0.908**	0.001	0.059
PSPNet	*densenet169*	0.714	0.002	0.832	0.002	0.172	0.679	0.001	0.808	0.001	0.083
UNet	***densenet169***	0.924	0.001	0.960	0.001	**0.057**	**0.832**	0.001	**0.908**	0.001	**0.036**
DeepLabV3+	*densenet169*	0.849	0.012	0.918	0.007	1.037	0.789	0.001	0.882	0.001	0.080

Table 9
Performance comparison on the ADAM dataset. Results are reported as mean ±S.D across cross-validation folds and the held-out test dataset. The best segmentation performance with the highest DSC and the most stable model with the lowest stability score are highlighted.

Model	Backbone	Cross-validation dataset					Test dataset				
		IoU	±S.D	DSC	±S.D	Stability	IoU	±S.D	DSC	±S.D	Stability
LinkNet	***densenet169***	0.911	0.002	0.953	0.001	**0.125**	0.834	0.003	0.909	0.002	0.209
FPN	***densenet169***	**0.931**	0.003	**0.964**	0.002	0.229	0.834	0.005	0.909	0.003	0.390
PSPNet	*vgg16*	0.801	0.003	0.888	0.002	0.265	0.762	0.005	0.864	0.004	0.508
UNet	***efficientnetb6***	0.917	0.006	0.956	0.003	0.425	0.832	0.002	0.907	0.001	**0.112**
DeepLabV3+	*senet154*	0.898	0.016	0.946	0.008	1.297	0.823	0.003	0.902	0.002	0.277

score of 0.209, suggesting more consistent predictions. In contrast, FPN achieved the highest cross-validation DSC but had a higher stability score of 0.390 on the test set, illustrating the trade-off between peak segmentation accuracy and robustness. PSPNet and DeepLabV3+ showed both lower accuracy and moderate stability. The visualization result of these models are illustrated in Fig. 12.

4.6. Discussion

This study presents a systematic and large-scale evaluation of five DL-based segmentation architectures, LinkNet, FPN, PSPNet, UNet, and DeepLabV3+, across four distinct TOF-MRA datasets, highlighting how backbone selection, model complexity, and dataset characteristics jointly influence segmentation performance and robustness. By benchmarking 45 model–backbone combinations across four heterogeneous datasets, the analysis extends beyond peak segmentation accuracy to examine performance stability and robustness, which are critical yet often underexplored aspects in clinical image segmentation.

4.6.1. Dataset-dependent performance

A consistent observation across all experiments is that segmentation performance is strongly dataset-dependent. The in-house dataset, derived from clinical MIP-based TOF-MRA, differs substantially from public whole-slice datasets in terms of contrast distribution, spatial context, and vessel continuity. These differences directly influenced how segmentation architectures leveraged contextual and multi-scale information.

On the in-house dataset, UNet with *seresnext50* achieved the highest segmentation accuracy, whereas LinkNet with *efficientnetb6* demonstrated superior stability, indicating more predictable and reproducible behavior across folds and test data. In contrast, for whole-slice-derived datasets (CereVessMRA, IXI, and ADAM), FPN-based configurations frequently achieved the highest cross-validation DSC values; however, these gains were often accompanied by increased variability on held-out test sets. Conversely, LinkNet and UNet variants exhibited more consistent generalization behavior, despite occasionally lower peak accuracy.

These findings highlight that segmentation models optimized for one dataset or acquisition protocol may not generalize reliably to others, emphasizing the importance of cross-dataset benchmarking when assessing clinical suitability. Table 10 summarizes the backbone–architecture combinations that consistently achieved either the highest accuracy or the best stability across datasets.

The comparatively weaker performance of PSPNet across all datasets, and the reduced robustness observed for DeepLabV3+, can be attributed to architectural design choices in relation to the characteristics of TOF-MRA data. While global context aggregation and atrous convolution strategies are effective for natural scene understanding, they appear less effective at consistently preserving fine-scale boundary

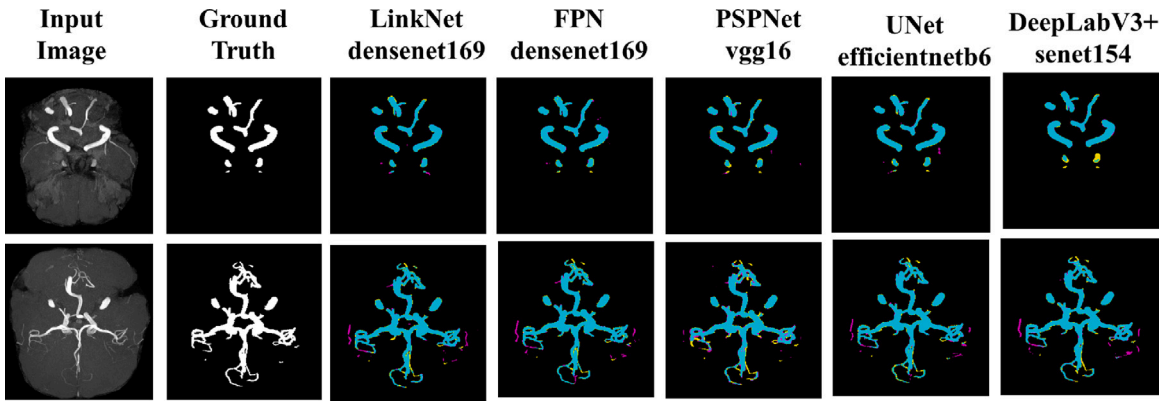


Fig. 12. Visual representation of the best segmentation model-backbone combination in overlapped form for ADAM dataset. Here 'cyan' indicates TP, 'magenta' indicates FN, 'yellow' highlights the FP.

Table 10
Summary of backbone–segmentation model combinations achieving the highest segmentation accuracy, best DSC, and the most stable performance across different datasets.

Backbone	In-house	CereVessMRA	IXI	ADAM
densenet169	–	–	FPN (Best DSC) LinkNet (Stable)UNet (Stable)	FPN (Best DSC) UNet (Stable)
senet154	–	FPN (Best DSC) LinkNet (Stable)	–	–
efficientnetb6	LinkNet (Stable)	–	–	UNet (Stable)
seresnext50	UNet (Best DSC)	–	–	–

information that is critical for accurate vascular delineation. In particular, PSPNet showed limited ability to segment thin and elongated vessel structures, resulting in lower overlap-based accuracy across datasets. DeepLabV3+, although capable of achieving moderate-to-high segmentation accuracy on several datasets, exhibited higher variability and reduced generalization to held-out test data compared to UNet- and LinkNet-based configurations. In addition, the relatively lightweight decoder designs of these architectures may constrain their capacity to reliably reconstruct small-caliber vessels from 2D MIP-based TOF-MRA images, where contrast between vascular and surrounding brain tissue is limited. Overall, these findings suggest that segmentation architectures originally designed for natural image understanding may require task-specific architectural adaptation to achieve both accurate and robust IA segmentation.

4.6.2. Stability-aware benchmarking and clinical relevance

By incorporating stability-aware evaluation alongside conventional accuracy metrics, this study provides a more nuanced framework for benchmarking segmentation models. In clinical applications, reproducibility and predictable behavior across patients and acquisition conditions are essential for building trust in automated systems. Models that exhibit high variability across cross-validation folds or test cohorts may yield inconsistent segmentations despite achieving high average performance, thereby limiting their translational potential.

While hyperparameter optimization and fine-tuning can improve peak segmentation accuracy, such improvements do not necessarily translate into enhanced robustness or reproducibility. The current findings indicate that model-backbone combinations exhibiting low intrinsic stability tend to retain higher variability even when accuracy is optimized, underscoring the importance of evaluating stability as a complementary criterion during model selection.

The stability analysis enables the identification of model-backbone configurations that achieve a favorable balance between segmentation accuracy and consistency, offering practical guidance for deployment in real-world and resource-constrained clinical settings. This stability-aware perspective aligns with emerging trends in medical image analysis that emphasize robustness, reliability, and uncertainty awareness alongside conventional performance metrics.

4.6.3. Limitations and future directions

Several limitations of this study should be acknowledged. First, all experiments were conducted on 2D MIP projections, which simplify inherently three-dimensional vascular anatomy and may obscure depth-related information. This choice may disadvantage architectures optimized for volumetric learning, such as nnUNet, and limits direct comparison with 3D segmentation approaches. Second, the in-house dataset originates from a single institution and scanner, which may introduce residual domain-specific bias despite the inclusion of three public datasets. However, the individual results on public comparatively larger dataset provide valuable guidance on segmentation-backbone selection on whole-slice TOF-MRA. Third, fixed cropping strategies, although empirically validated for the in-house cohort, may not generalize to atypical vascular anatomy or alternative acquisition protocols. Fourth, the in-house dataset primarily focuses on main trunk arteries, where TOF-MRA is more reliable, but future work will extend segmentation to include small-flow arteries alongside the main trunks. However, the three public dataset are derived from whole-slice TOF-MRA, and thus they capture a more complete representation of the cerebrovasculature, including both large and small vessels. Finally, although multiple datasets were included, domain adaptation and cross-institutional generalization were not explicitly addressed.

Future research will focus on extending this benchmarking framework to 3D volumetric TOF-MRA, integrating adaptive region-of-interest extraction, and incorporating domain adaptation strategies to improve cross-institutional robustness. Additional, task-specific contrast metrics such as Contrast-to-Noise Ratio (CNR) will be explored in future work. Therefore, exploring hybrid architectures that jointly optimize accuracy, stability, and computational efficiency will further enhance the clinical applicability of IA segmentation models.

5. Conclusions

This study presents a large-scale, systematic benchmarking of DL encoder–decoder architectures for IA segmentation from TOF-MRA across heterogeneous datasets. Evaluation of 45 architecture-backbone

combinations on both in-house and public datasets (CereVessMRA, IXI, ADAM) demonstrates that segmentation performance is governed by a complex interplay between architectural design, backbone capacity, and dataset characteristics. Importantly, high overlap-based metrics alone are shown to be insufficient for identifying clinically reliable models. Architectures achieving the highest Dice or IoU scores frequently exhibited substantial variability across cross-validation and independent test cohorts, limiting reproducibility. In contrast, LinkNet configurations with backbones such as *efficientnetb6*, *densenet169*, and *senet154* consistently demonstrated greater stability, whereas UNet with *seresnext50*, *densenet169* and FPN with *senet154* and *densenet169* often achieved higher peak DSC values at the cost of increased variability. PSPNet variants underperformed consistently, and DeepLabV3+ exhibited reduced robustness despite competitive accuracy, highlighting stability and generalization as key challenges in this task.

However, the analysis was restricted to 2D MIP representations, which simplify the inherently three-dimensional vascular anatomy and may obscure depth-related information, particularly in complex or overlapping vessel structures. The in-house dataset was acquired from a single institution and scanner, potentially introducing domain-specific bias. In addition, fixed cropping strategies, while effective for the studied in-house cohort, may not generalize to atypical anatomy or alternative acquisition protocols. Future research should extend this benchmarking framework to fully 3D volumetric TOF-MRA, incorporate adaptive ROI extraction, and integrate domain-adaptive and uncertainty-aware learning strategies to improve cross-institutional robustness. The development of hybrid architectures that jointly optimize segmentation accuracy, stability, and computational efficiency is expected to play a critical role in enabling reliable clinical translation of automated IA segmentation.

CRediT authorship contribution statement

Mekhla Sarkar: Conceptualization, Data curation, Methodology, Formal analysis, Writing – original draft, Visualization. **Yen-Chu Huang:** Conceptualization, Data curation, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Tsong-Hai Lee:** Conceptualization, Investigation, Data verification, Writing – review & editing, Visualization. **Jiann-Der Lee:** Investigation, Data verification, Writing – review & editing, Visualization. **Prasan Kumar Sahoo:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Project administration, Funding acquisition.

Sources of funding

This work is supported in part by the National Science and Technology Council (NSTC), Taiwan grant number 114-2221-E-182-022-MY3 and in part by the Chang Gung Memorial Hospital, Taiwan research grants number CORPG6L0151.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table A.11
Simulation environment configuration and training setup.

Hyperparameter	Value
Input Image Size	(384, 384)
Batch Size	2
Training Epochs	50
Dropout Rate	0.05
Learning Rate	0.000125
Early Stopping Patience	6 epochs

Appendix

A.1. Implementation details

To evaluate the robustness and generalizability of different model–backbone configurations, the datasets were processed as follows:

- **In-house dataset:** The dataset was split into 90% for training (889 images from 51 patients) and 10% for testing (114 images from 7 patients), with the latter reserved exclusively for external performance assessment. Within the training subset, ten-fold cross-validation was performed, with approximately 800 images for training and 96 images from 6 patients for validation in each fold, ensuring no data overlap across folds.
- **CereVessMRA:** This dataset consisted of 5420 2D MIP projection slices from 271 patients. For training, 90% of patients (4880 slices from 244 patients) were used, while 10% (540 slices from 27 patients) were reserved for testing. Five-fold cross-validation was performed on the training subset, with approximately 3900 slices from 195 patients for training and 980 slices from 49 patients for validation per fold, ensuring no overlap across folds.
- **ADAM:** The dataset included 1900 2D MIP slices from 95 patients for training and 240 slices from 12 patients for testing. Five-fold cross-validation was applied to the training set, with each fold containing approximately 1720 slices from 86 patients for training and 180 slices from 9 patients for validation, ensuring no repeated data across folds.
- **IXI:** This dataset contained 2720 2D MIP slices from 136 patients and was used for five-fold cross-validation. In each fold, approximately 2440 slices from 122 patients were used for training, and 280 slices from 14 patients for validation. The held-out test set contained 680 slices from 34 patients, following the COSTA dataset distribution.

This strategy enabled a comprehensive evaluation of model stability, with the average performance across folds reflecting overall consistency. The mean ±S.D of test performance across folds was additionally reported to assess generalization and variability across experimental runs.

Key hyperparameters used during training are summarized in Table A.11. Optimal configurations were determined using the Keras Tuner framework (O'Malley et al., 2019), which automates the search for the best-performing settings. Regularization techniques such as *early stopping* and cross-validation were applied throughout to prevent overfitting and ensure reliable convergence.

All experiments were conducted using GPU-enabled TensorFlow 2.4.0 on a system with an Intel Core i7-8700K CPU (3.7 GHz), 32 GB RAM, and a GeForce GTX 1070 Ti GPU (driver 418), running Ubuntu 18.04.3. Segmentation architectures (UNet, LinkNet, FPN, PSPNet, DeepLabV3+) with pretrained backbones were implemented using the open-source library (Yakubovskiy, 2019).

Table A.12
Summary of architectural and training configurations for all segmentation models.

Model	Skip-connections	Decoder block	Upsampling method	Batch norm	Input handling	Decoder filters
UNet	Concatenative skips from all encoder stages	Two 3 × 3 conv + ReLU per stage	Bilinear upsampling	Trainable BN	1-channel → 3-channel replication; backbone-specific preprocessing	[256,128,64,32,16]
LinkNet	Additive residual encoder-to-decoder skips	1 × 1 channel reduction + bilinear upsampling + residual refinement block	Bilinear upsampling	Trainable BN	1→3 replication; backbone-specific preprocessing	[512,512,256,128,16]
FPN	Lateral 1 × 1 merges from encoder stages	Top-down fusion + 3 × 3 smoothing conv per pyramid level	Bilinear upsampling	Trainable BN	1→3 replication; backbone-specific preprocessing	[128,128,128,128,128]
PSPNet	Global contextual skip via pooling module	Pyramid pooling (1,2,3,6 bins) + 3 × 3 refinement conv	Bilinear upsampling	Trainable BN	1→3 replication; backbone-specific preprocessing	[512,512]
DeepLabV3+	Single low-level skip (128 × 128 stage)	ASPP + two 3 × 3 convs + fusion with low-level features	4×upsampling + final bilinear upsampling	Trainable BN	1→3 replication; backbone-specific preprocessing	[256,256]

A.2. Decoder architecture details

This section provides a detailed description of the decoder configurations employed in each segmentation model to facilitate experimental reproducibility. The decoder plays a critical role in restoring spatial resolution and refining feature representations derived from the encoder, and the architectural choices in this stage directly influence segmentation accuracy, boundary precision, and computational complexity.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2026.100843.

Data availability

The authors do not have permission to share data.

References

Abdelrahman, A., & Viriri, S. (2023a). EfficientNet family U-net models for deep learning semantic segmentation of kidney tumors on CT images. *Frontiers in Computer Science*, *5*, Article 1235622.

Abdelrahman, A., & Viriri, S. (2023b). FPN-SE-ResNet model for accurate diagnosis of kidney tumors using ct images. *Applied Sciences*, *13*(17), 9802.

Bhatti, U. A., Liu, J., Huang, M., & Zhang, Y. (2025). FF-UNet: Feature fusion based deep learning-powered enhanced framework for accurate brain tumor segmentation in MRI images. *Image and Vision Computing*, Article 105635.

Chaurasia, A., & Culurciello, E. (2017). Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing* (pp. 1–4). IEEE.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chiang, K.-Y. (2020). Polygon-crop 0.0.3. https://pypi.org/project/polygon-crop/.

Deshpande, A., Jamilpour, N., Jiang, B., Michel, P., Eskandari, A., Kidwell, C., Wintermark, M., & Laksari, K. (2021). Automatic segmentation, feature extraction and comparison of healthy and stroke cerebral vasculature. *NeuroImage: Clinical*, *30*, Article 102573.

Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. B. (2020). Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE Transactions on Medical Imaging*, *39*(11), 3679–3690.

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics: vol. 4*, Cambridge university press Cambridge, UK.

Fu, Q., Liu, D.-X., Zhang, X.-Y., Deng, X.-B., & Zheng, C.-S. (2020). Pointwise encoding time reduction with radial acquisition in subtraction-based magnetic resonance angiography to assess saccular unruptured intracranial aneurysms at 3 tesla. *Neuroradiology*, 1–11.

Ge, X., Zhao, H., Zhou, Z., Li, X., Sun, B., Wu, H., et al. (2019). Association of fractional flow on 3D-TOF-mra with cerebral perfusion in patients with MCA stenosis. *American Journal of Neuroradiology*, *40*(7), 1124–1131.

Guo, B., Chen, Y., Lin, J., Huang, B., Bai, X., Guo, C., et al. (2024). Self-supervised learning for accurately modelling hierarchical evolutionary patterns of cerebrovasculature. *Nature Communications*, *15*(1), 9235.

Han, Y., Qiao, H., Chen, S., Jing, J., Pan, Y., Li, D., et al. (2018). Intracranial artery stenosis magnetic resonance imaging aetiology and progression study: Rationale and design. *Brain and Behavior*, *8*(12), Article e01154.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hou, L., Zhang, J., Zhao, L., Meng, K., & Feng, X. (2025). CTA image segmentation method for intracranial aneurysms based on mglia net. *Scientific Reports*, *15*(1), 10593.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Hussain, T., Shouno, H., Mohammed, M. A., Marhoon, H. A., & Alam, T. (2025). DCSSGA-UNet: Biomedical image segmentation with DenseNet channel spatial and semantic guidance attention. *Knowledge-Based Systems*, *314*, Article 113233.

I. X. I. Project (2025). IXI dataset. https://brain-development.org/ixi-dataset/. (Accessed 06 December 2025).

Jun, T. J., Kweon, J., Kim, Y.-H., & Kim, D. (2020). T-net: Nested encoder–decoder architecture for the main vessel segmentation in coronary angiography. *Neural Networks*, *128*, 216–233.

Li, S., & Xie, B. (2025). Boosting unet performance via VGG-based encoder for medical image segmentation. In *2025 6th international conference on computer engineering and application* (pp. 1964–1968). IEEE.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Martinsson, J., & Mogren, O. (2019). Semantic segmentation of fashion images using feature pyramid networks. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Min, Y., Li, J., Jia, S., Li, Y., & Nie, S. (2024). Automated cerebrovascular segmentation and visualization of intracranial time-of-flight magnetic resonance angiography based on deep learning. *Journal of Imaging Informatics in Medicine*, 1–14.

Mou, L., Lin, J., Zhao, Y., Liu, Y., Ma, S., Zhang, J., et al. (2024). COSTA: A multi-center TOF-mra dataset and a style self-consistency network for cerebrovascular segmentation. *IEEE Transactions on Medical Imaging*, *43*(12), 4442–4456.

Mu, N., Lyu, Z., Rezaeitaleshmahalleh, M., Tang, J., & Jiang, J. (2023). An attention residual u-net with differential preprocessing and geometric postprocessing: Learning how to segment vasculature including intracranial aneurysms. *Medical Image Analysis*, *84*, Article 102697.

Nageler, G., Gergel, I., Fangerau, M., Breckwoldt, M., Seker, F., Bendszus, M., Möhlenbruch, M., & Neuberger, U. (2023). Deep learning-based assessment of internal carotid artery anatomy to predict difficult intracranial access in endovascular recanalization of acute ischemic stroke. *Clinical Neuroradiology*, 1–10.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., & Team, K.-T. (2019). KerasTuner (version 1.0). URL https://github.com/keras-team/keras-tuner.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Varoquaux, A., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rahman, S., Rahman, M. M., Abdullah-Al-Wadud, M., Al-Quaderi, G. D., & Shoyaib, M. (2016). An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, *2016*(1), 1–13.

Rayed, M. E., Islam, S. S., Niha, S. I., Jim, J. R., Kabir, M. M., & Mridha, M. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked*, *47*, Article 101504.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, *7*(3), 8–18.

Sartoretti, T., Sartoretti, E., Schwenk, Á., van Smoorenburg, L., Mannil, M., Euler, A., et al. (2020). Clinical feasibility of ultrafast intracranial vessel imaging with non-cartesian spiral 3D time-of-flight MR angiography at 1.5 T: An intra-individual comparison study. *PLoS One*, *15*(4), Article e0232372.

Sharma, N., Gupta, S., Rajab, A., Elmagzoub, M. A., Rajab, K., & Shaikh, A. (2023). Semantic segmentation of gastrointestinal tract in mri scans using pspnet model with resnet34 feature encoding network. *IEEE Access*, *11*, 132532–132543.

Shi, Z., Li, J., Zhao, M., Peng, W., Meddings, Z., Jiang, T., et al. (2020). Quantitative histogram analysis on intracranial atherosclerotic plaques: A high-resolution magnetic resonance imaging study. *Stroke*, *51*(7), 2161–2169.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singh, M., Ansari, M. S. A., & Govil, M. C. (2025). Detection of fractional difference in inter vertebral disk MRI images for recognition of low back pain. *Image and Vision Computing*, *153*, Article 105333.

Sulaiman, A., Anand, V., Gupta, S., Al Reshan, M. S., Alshahrani, H., Shaikh, A., & Elmagzoub, M. (2024). An intelligent LinkNet-34 model with EfficientNetB7 encoder for semantic segmentation of brain tumor. *Scientific Reports*, *14*(1), 1345.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Tian, X., Tian, B., Shi, Z., Wu, X., Peng, W., Zhang, X., et al. (2020). Assessment of intracranial atherosclerotic plaques using 3D black-blood MRI: Comparison with 3D time-of-flight MRA. *Journal of Magnetic Resonance Imaging*.

Timmins, K., Bennink, E., van der Schaaf, I., Velthuis, B., Ruigrok, Y., & Kuijf, H. (2020). Intracranial aneurysm detection and segmentation challenge. In *Proc. MICCAI* (pp. 30–31).

Yakubovskiy, P. (2019). Segmentation models [computer software]. URL https://github.com/qubvel/segmentation_models.

Yao, L., Chen, D., Zhao, X., Fei, M., Song, Z., Xue, Z., et al. (2024). AASeg: Artery-aware global-to-local framework for aneurysm segmentation in head and neck CTA images. *IEEE Transactions on Medical Imaging*.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890).