

PAPER

Deeply Supervised 3D U-Net with Auxiliary Classification for Cerebral Aneurysms Segmentation

Mario Mata-Castillo¹,
 Nelly Gordillo-Castillo¹,
 Boris Mederos¹, Dora-Luz
 Flores² , Vicente García¹,
 José Díaz-Román¹ 

¹Universidad Autónoma
 de Ciudad Juárez, Ciudad
 Juárez, Mexico

²Universidad Autónoma
 de Baja California,
 Mexicali, Mexico

david.roman@uacj.mx

ABSTRACT

Cerebral aneurysms are dilations of intracranial arteries whose early detection is essential for treatment planning and rupture prevention. Automated aneurysm segmentation remains challenging because of severe class imbalance and the small size of many lesions. This study presents a 3D U-Net incorporating deep supervision and an auxiliary classification branch, trained on the public ADAM dataset using skeleton-guided patch extraction. Fixed-size patches ($64 \times 64 \times 64$) were extracted from volumes resampled to 0.4 mm isotropic resolution, and performance was evaluated using stratified five-fold cross-validation. The proposed model achieved a Dice coefficient of 0.425 (95% CI: 0.362–0.488) and a false-positive (FP) patch rate of 0.026 (95% CI: 0.017–0.035) on the test set. For aneurysms larger than 2.5 mm, the Dice coefficient increased to 0.548 (95% CI: 0.485–0.612), while the FP patch rate was 0.030 (95% CI: 0.018–0.042), indicating improved performance on clinically relevant lesions. Statistical analysis confirmed that the proposed model was significantly superior to the compared models in both the Dice coefficient and FP patch rate. Although the framework produces patch-level rather than whole-volume segmentations, visual analysis demonstrates accurate localization of aneurysm regions. These results indicate that combining skeleton-guided patch extraction with deep supervision and auxiliary classification improves segmentation performance while providing reliable localization cues for automated aneurysm detection.

KEYWORDS

cerebral aneurysms, segmentation, deep supervision, magnetic resonance angiography (MRA)

1 INTRODUCTION

A cerebral aneurysm is a vascular disorder characterized by a localized dilation of a cerebral blood vessel, the rupture of which causes a subarachnoid hemorrhage (SAH) [1]. According to the World Health Organization (WHO), the incidence of SAH due to ruptured aneurysms ranges from 2 to 22.5 per 100,000 inhabitants per year [2]. Furthermore, 12% of affected individuals die before receiving medical

Mata-Castillo, M., Gordillo-Castillo, N., Mederos, B., Flores, D.-L., García, V., Díaz-Román, J. (2026). Deeply Supervised 3D U-Net with Auxiliary Classification for Cerebral Aneurysms Segmentation. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(10), pp. 140–160. <https://doi.org/10.3991/ijoe.v22i10.62047>

Article submitted 2026-04-24. Revision uploaded 2026-07-13. Final acceptance 2026-07-13.

© 2026 by the authors of this article. Published under CC-BY.

attention, and 40% of those who are hospitalized die within one month after the event. Survivors often suffer from chronic headaches, sleep disorders, and cognitive impairments that severely affect their quality of life [3].

Unruptured intracranial aneurysms are usually asymptomatic and are often diagnosed incidentally during imaging studies performed for other reasons. When symptomatic, unruptured aneurysms produce clinical manifestations due to the compression of adjacent brain structures as they enlarge, including severe head, face, or neck pain and oculomotor nerve palsy. Based on their morphology, cerebral aneurysms are commonly classified into three types: saccular, characterized by a berry-like outpouching; lateral or dissecting, in which the middle layer of the arterial wall separates, allowing blood flow between layers; and fusiform, defined by a circumferential dilation of the vessel segment. In clinical practice, aneurysm morphology is considered when evaluating rupture risk and choosing the most appropriate treatment approach.

Several imaging modalities are available for the detection of cerebral aneurysms. Among them, digital subtraction angiography (DSA) is commonly considered the reference standard because it provides detailed visualization of the cerebral vasculature after catheter-based contrast injection. Nevertheless, its invasive nature, cost, and limited availability have led to the widespread use of noninvasive techniques such as magnetic resonance angiography (MRA) and computed tomography angiography (CTA) [4].

MRA and CTA have shown high diagnostic accuracy for aneurysm detection, with sensitivity and specificity above 90% when interpreted by expert neuroradiologists, although DSA remains the reference standard. Even so, small aneurysms may still be missed on visual inspection, including by experienced physicians. For this reason, several studies have explored artificial intelligence (AI)-based computational tools for the automatic detection of cerebral aneurysms [5], [6], [7].

In recent years, significant progress has been made in the automatic segmentation of cerebral aneurysms. A prominent example is the Aneurysm Detection and Segmentation (ADAM) Challenge, held in 2020, where different research groups proposed innovative methodologies for both detection and segmentation tasks. The team that achieved first place obtained a sensitivity of 0.68 for detection and a Dice coefficient of 0.43 for segmentation using a 3D U-Net architecture trained on patches of $128 \times 128 \times 128$ voxels. In addition to these challenge-specific methods, general-purpose segmentation frameworks such as nnU-Net have established strong baselines across a wide range of 3D medical imaging tasks [8]. Recent hybrid architectures that integrate convolutional and transformer-based components, such as TransUNet, have further demonstrated the benefits of global context modeling in medical image segmentation [9]. Attention-based models such as Swin-UNet further highlight the importance of multi-scale contextual aggregation and long-range feature interactions in 3D and 2D medical image segmentation [10], particularly in volumetric tasks. However, despite these advances, the reported metrics remain insufficient for clinical implementation, particularly considering the high accuracy required in neurovascular applications. Furthermore, several studies have shown substantial variability when re-implementing ADAM Challenge methodologies on the official dataset partitions, highlighting the sensitivity of the task to data configuration and preprocessing choices [11].

A major difficulty in this task is the extreme imbalance between aneurysm voxels and background voxels. This is particularly problematic for small aneurysms, which may contain fewer than 50 voxels within a $64 \times 64 \times 64$ patch. As a result, their contribution to the loss function is very limited, which can make optimization unstable and reduce segmentation performance.

Performance differences also become evident when the results are analyzed by aneurysm size. In particular, segmentation is consistently poorer for small aneurysms

than for medium or large ones. Improving the delineation of small lesions therefore remains an unresolved issue with clear clinical relevance. Different strategies have been explored to address this limitation, including data augmentation, loss weighting, and auxiliary classifiers designed to increase sensitivity to small structures.

Ham et al. [12] reported a 3D U-Net with an auxiliary classification branch trained on skeleton-based patches, with competitive results on internal and external datasets related to the ADAM challenge. In our experience, however, performance decreases when that approach is reproduced using only the ADAM dataset and evaluated under cross-validation. This suggests that the reported results may depend, at least in part, on particular data settings or implementation details that were not fully specified in the original publication.

Here, we evaluate a deeply supervised 3D U-Net trained exclusively on the ADAM dataset. The network includes multiple decoder outputs, which are combined in the loss function to reinforce feature learning at different scales. We analyze model performance according to aneurysm size, with particular attention to its behavior in small lesions. We also consider the possible value of weak labels in clinical settings where detecting the presence and approximate location of an aneurysm may be more useful than obtaining exact voxel-level boundaries. Under this framework, we seek to improve sensitivity while preserving a reproducible experimental design.

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology, including the network architecture and training strategy. Section 3 presents the experimental setup, dataset, and evaluation metrics. Section 4 reports and discusses the experimental results. Section 5 describes the availability of the data used in this work. Finally, Section 6 concludes the paper and outlines directions for future work.

2 MATERIALS AND METHODS

2.1 Database

For the development of this system, the MRA image database from the ADAM Challenge [11] was used. The dataset consists of 93 patients with aneurysms and 20 controls. Images were acquired using 1T, 1.5T or 3T magnetic field strength scanners and a time-of-flight (TOF) pulse sequence. Magnetic field inhomogeneity artifacts were corrected in all images using the N4 bias-field correction algorithm. The database also provides the aneurysm masks and their locations.

Image intensities were clipped to the [0.5, 0.95] percentiles and normalized to follow a standard distribution. To ensure a consistent voxel size across all samples, each volume was rescaled to $0.4 \text{ mm} \times 0.4 \text{ mm} \times 0.4 \text{ mm}$. A patch-based approach was employed for training, using subvolumes (patches) of $64 \times 64 \times 64$ voxels. Patch-based sampling has been widely used to improve the detectability of small lesions in volumetric medical imaging, particularly when full-resolution training is computationally prohibitive [13], [14].

2.2 Preprocessing and patch extraction

This section describes the preprocessing pipeline and patch extraction strategy used to address the severe class imbalance inherent to intracranial aneurysm segmentation (see Figure 1).

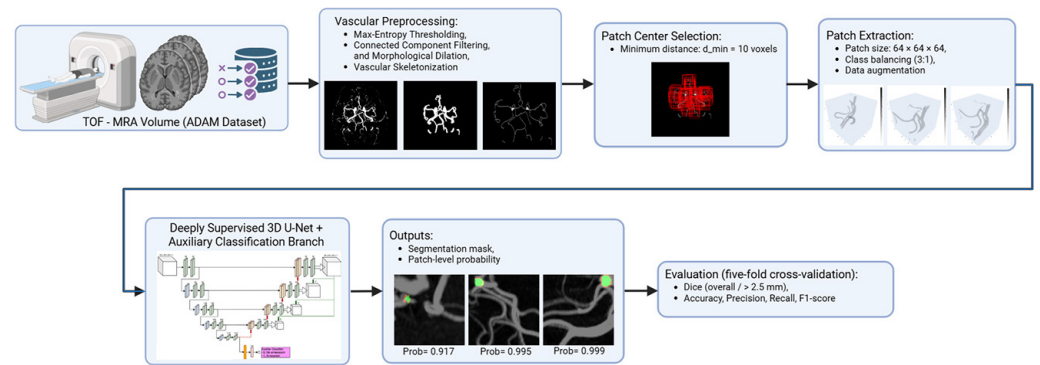


Fig. 1. Overview of the proposed methodological pipeline

Notes: The workflow starts from TOF-MRA volumes from the ADAM dataset, followed by vascular-focused preprocessing using maximum entropy thresholding, connected component filtering, and morphological dilation. A vascular skeleton is then extracted to guide patch center selection with a minimum spatial separation. Balanced and augmented 3D patches are fed into a deeply supervised 3D U-Net with an auxiliary classification branch, producing both voxel-wise segmentation masks and patch-level aneurysm presence probabilities. Model performance was assessed with the Dice similarity coefficient across aneurysm sizes and with standard classification metrics.

Because aneurysms are sparse and occupy only small regions within large volumetric images, the workflow first limits patch extraction to vascular areas. The vascular mask is then refined with morphological operations, and patch centers are selected to obtain the training samples. Patch labeling, negative subsampling, and data augmentation are subsequently applied to improve class balance and generalization. The main preprocessing steps are described below:

- Maximum entropy thresholding [15]: Let $I: \Omega \subset \mathbb{R}^3 \rightarrow \mathbb{R}$ denote the volumetric intensity image defined over the voxel domain Ω . The intensities values of I are quantized into L discrete bins $\{0, 1, \dots, L-1\}$. Accordingly, let n_i be the number of voxels in bin i and let

$$N = \sum_{i=0}^{L-1} n_i \quad (1)$$

be the total number of voxels. Consequently, the normalized histogram

$$h: \{0, 1, \dots, L-1\} \rightarrow [0, 1] \quad (2)$$

of voxel intensities is defined as

$$h(i) = \frac{n_i}{N} \quad (3)$$

For a given threshold t , the background and foreground probability distributions are computed as:

$$p_b(i) = \frac{h(i)}{\sum_{j=1}^t h(j)}, \quad \text{for } i \leq t \quad (4)$$

$$p_f(i) = \frac{h(i)}{\sum_{j=t+1}^L h(j)}, \quad \text{for } i > t \quad (5)$$

with corresponding entropies:

$$H_b(t) = - \sum_{i \leq t} p_b(i) \log p_b(i) \quad (6)$$

$$H_f(t) = - \sum_{i > t} p_f(i) \log p_f(i) \quad (7)$$

The optimal threshold is obtained by maximizing the total entropy:

$$t^* = \operatorname{argmax}_t [H_b(t) + H_f(t)] \quad (8)$$

From the threshold value t^* the binary segmentation is obtained as

$$B(x) = \begin{cases} 1, & \text{if } I(x) > t^* \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

- Connected component filtering: Although maximum entropy thresholding yields a consistent vasculature segmentation across cases, small, isolated components may persist due to noise or imaging artifacts. These are removed using connected component analysis. Let

$$\mathcal{S} = \{x \in \Omega: B(x) = 1\} \quad (10)$$

denote the set of foreground voxels of the binary mask B , defined over the image domain Ω . Let

$$\mathcal{P} = \{C_j\}_{j=1}^M \quad (11)$$

denote the set of connected components of $\mathcal{S}$, where C_j is the j -th connected component and M is the total number of components. Only components whose size exceeds a predefined threshold τ are retained, yielding the cleaned mask

$$\mathcal{S}_c = \bigcup_{C_j \in \mathcal{P}, |C_j| > \tau} C_j \quad (12)$$

where $|C_j|$ denotes the number of voxels in C_j .

- Morphological dilation: To improve vessel continuity and ensure robustness in subsequent processing, a morphological dilation is applied to the cleaned mask. Let $\mathcal{K} \subset \mathbb{R}^3$ denote a structuring element. The dilated volume is defined as

$$\mathcal{S}_d = \mathcal{S}_c \oplus \mathcal{K} \quad (13)$$

where $\oplus$ denotes the dilation operator [16].

- Vascular Skeletonization: From the dilated volume, a centerline representation of the vasculature is obtained via morphological skeletonization [17]. Let $\mathcal{K}$ denote the structuring element. The skeleton is defined as

$$\mathcal{S}_e = \operatorname{Skel}(\mathcal{S}_d) = \bigcup_{k=1}^{k_{\max}} (\mathcal{S}_d \ominus k\mathcal{K}) \setminus ((\mathcal{S}_d \ominus k\mathcal{K}) \circ \mathcal{K}) \quad (14)$$

where $\mathcal{S}_d$ is the dilated foreground set, $\mathcal{S}_e$ is the resulting vascular skeleton, and $k_{\max}$ is the largest integer for which $\mathcal{S}_d \ominus k\mathcal{K}$ is non-empty. Here, $k\mathcal{K}$ denotes the k -fold dilation of the structuring element $\mathcal{K}$. The operators $\ominus$, $\oplus$, and $\circ$ denote morphological erosion, dilation, and opening, respectively [16].

- Patch center selection: The skeleton is used as a reference for patch center selection. A subset of centers $\mathcal{C} \subset \mathcal{S}_e$ is constructed by enforcing a minimum Euclidean separation,

$$\|c_i - c_j\|_2 \geq d_{\min}, \quad \forall c_i, c_j \in \mathcal{C}, i \neq j \quad (15)$$

with $d_{\min} = 10$ voxels, ensuring that the extracted patches will not exhibit excessive overlap.

- Patch sampling: From each selected center $c \in \mathcal{C}$ it is defined as a cubic subvolume

$$\mathcal{V}_c = \{x \in \Omega : \|x - c\|_{\infty} \leq r\} \quad (16)$$

centered at c with patch half-size r , where $\|\cdot\|_{\infty}$ denotes the infinite norm on $\mathbb{R}^3$ defined on $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ as

$$\|x\|_{\infty} = \max_{i \in \{1,2,3\}} |x_i| \quad (17)$$

For each subvolume, $\mathcal{V}_c$, $c \in \mathcal{C}$, an intensity patch x is extracted from the image I and its corresponding mask y is extracted from the annotation volume Y . Each subvolume therefore provides one input-label pair.

Class imbalance at the patch level was addressed by randomly subsampling negative patches so that the training set contained three negative samples for every positive one. Finally, to enhance model generalization and reduce overfitting, data augmentation is applied independently to each extracted patch. Augmentations include random rotations in the range of -15° to 15° along all three axes, as well as random flips along each axis.

An illustrative overview of the main preprocessing steps, including vascular segmentation, skeletonization, and patch center selection, is shown in Figure 2.

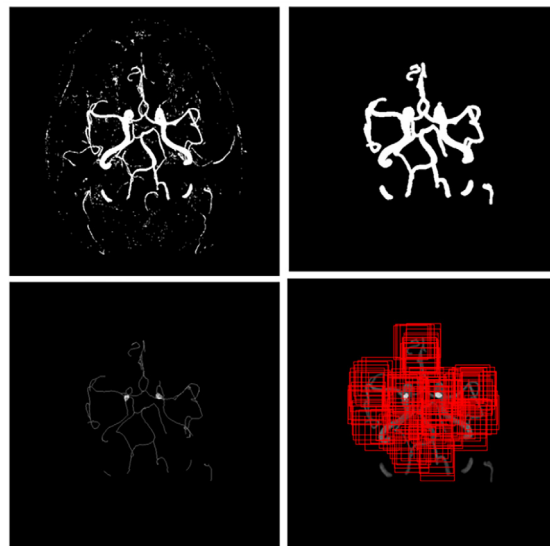


Fig. 2. Overview of the preprocessing and sampling stages for aneurysm segmentation

Notes: From left to right, top to bottom: vascular structures enhanced using maximum-entropy thresholding; refined binary mask after morphological cleaning and dilation to preserve thin vessels and connectivity; vessel centerlines extracted through skeletonization of the binary segmentation; and selected patch centers (red squares) placed along the extracted skeleton to guide localized aneurysm candidate generation.

2.3 Model

In this study, several strategies previously explored in other applications as well as in this domain were implemented. Initially, experiments were conducted using a 3D U-Net as a baseline. Subsequently, deep supervision was incorporated by introducing auxiliary decoder outputs and modifying the loss computation through weighted multi-scale predictions. Additionally, the dataset was divided into five folds to perform stratified cross-validation, taking aneurysm size into account.

Two 3D convolutional neural network architectures were evaluated. The baseline 3D U-Net was designed with an initial filter count of 16, structured over five stages with four downsampling steps, and used He Uniform initialization. Each block incorporated dropout, and ReLU was used as the activation function in all layers.

Additionally, deep supervision and an auxiliary classification branch were included (see Figure 3). This architecture extends the baseline 3D U-Net by generating multiple intermediate outputs from the decoder, which are weighted and combined in the loss function to reinforce feature learning at multiple scales. Deep supervision has been shown to enhance feature propagation across scales and stabilize optimization in highly imbalanced segmentation problems [18].

In parallel, an auxiliary classification branch is connected to the bottleneck of the network, providing an additional supervision signal that encourages the model to discriminate between patches containing aneurysms and background regions. Prior work has explored the use of auxiliary classification in 3D U-Net-based aneurysm segmentation to guide feature learning [12]. In this work, the combination of the auxiliary classifier with the deeply supervised approach aims to enhance sensitivity to small aneurysms while preserving robust segmentation performance for larger lesions by strengthening multi-scale feature propagation and enriching the supervisory signals during training.

More formally, given a volumetric patch $x \in \mathbb{R}^{64 \times 64 \times 64 \times 1}$ and its corresponding labels $y \in [0, 1]^{64 \times 64 \times 64 \times 2}$, the 3D U-Net encoder-decoder model with skip connections can be viewed as the estimation of a function

$$f_{\theta}: \mathbb{R}^{64 \times 64 \times 64 \times 1} \rightarrow [0, 1]^{64 \times 64 \times 64 \times 2} \quad (18)$$

Under the deep supervision framework, four output scales are introduced, each associated with a different level of the decoder. These outputs are denoted by $\{f_{\theta}^{(s)}\}_{s=1}^4$, where $s = 1$ corresponds to the coarsest resolution and $s = 4$ corresponds to the final output resolution of the 3D U-Net.

Each output $f_{\theta}^{(s)}$ produces a probability map

$$p^{(s)} = \sigma(f_{\theta}^{(s)}(x)) \quad (19)$$

where σ denotes a voxel-wise softmax function. Due to the different spatial resolutions of the decoder outputs, the reference label y is rescaled using a downsampling operator, yielding the set of labels $\{y^{(s)}\}_{s=1}^4$.

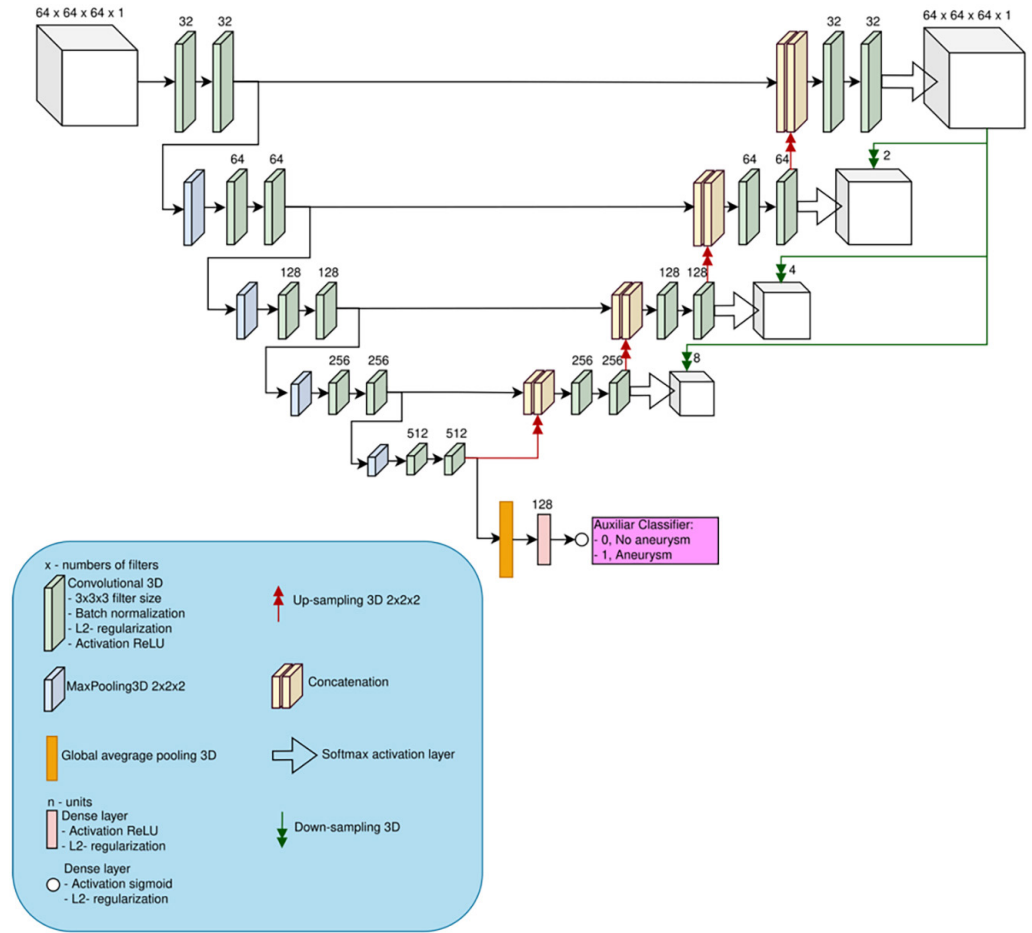


Fig. 3. Architecture of the proposed deeply supervised 3D U-Net with an auxiliary classification branch for aneurysm segmentation

Notes: The network follows an encoder–decoder design with 3D convolutions, multi-scale skip connections, and progressive down-sampling and up-sampling operations. Deep supervision is applied at multiple decoder levels to enhance gradient propagation and encourage discriminative feature learning across scales. In parallel, an auxiliary classification branch, driven by high-level features through global average pooling and fully connected layers, predicts aneurysm presence (0: no aneurysm, 1: aneurysm). The auxiliary output is used during both training and inference to provide complementary global context that supports robust segmentation.

For each scale s , a local loss $\mathcal{L}^{(s)}$ is defined as a weighted combination of Generalized Dice loss and Focal loss:

$$\mathcal{L}^{(s)}(p^{(s)}, y^{(s)}) = \lambda_{\text{dice}} \mathcal{L}_{\text{GDice}}(p^{(s)}, y^{(s)}) + \lambda_{\text{focal}} \mathcal{L}_{\text{Focal}}(p^{(s)}, y^{(s)}) \quad (20)$$

where λ_{dice} and λ_{focal} are constants.

The Generalized Dice loss [19] is defined as:

$$\mathcal{L}_{\text{GDice}}(p^{(s)}, y^{(s)}) = 1 - \frac{2 \sum_{k=1}^2 \alpha_k^{(s)} \sum_{v \in V^{(s)}} p_{k,v}^{(s)} y_{k,v}^{(s)}}{\sum_{k=1}^2 \alpha_k^{(s)} \sum_{v \in V^{(s)}} (p_v^{(s)} + y_v^{(s)})} \quad (21)$$

where v indexes the voxels over the set of voxels $V^{(s)}$ at scale s , $p_{k,v}^{(s)}$ denotes the predicted probability to belong to class k at voxel v and scale s , $y_{k,v}^{(s)} \in \{0, 1\}$ denotes

the indicator of class k , where $y_{k,v}^{(s)} = 1$ indicates that voxel v at scale s belongs to class k (0 background or 1 foreground), and the weight $\alpha_k^{(s)}$ per class is defined as

$$\alpha_k^{(s)} = \frac{1}{\left(\sum_{v \in V^{(s)}} y_{k,v}^{(s)}\right)^2} \quad (22)$$

The Focal loss [20] is defined as:

$$\mathcal{L}_{Focal}(p^{(s)}, y^{(s)}) = -\frac{1}{|V^{(s)}|} \sum_{v \in V^{(s)}} \beta_v^{(s)} (1 - p_v^{(s)})^{\gamma^{(s)}} \log p_v^{(s)} \quad (23)$$

where $p_v^{(s)}$ denotes the probability of voxel v to belong to the ground-truth class at scale s , the focal parameter $\gamma^{(s)} \geq 0$ controls the degree of emphasis placed on hard examples $p_v^{(s)} \approx 0$, and $\beta_v^{(s)}$ is a per-class weight that controls how much samples of the ground-truth class contribute to the loss at voxel v and scale s .

The auxiliary classifier is attached to the output of the bottleneck block of the 3D U-Net. This classifier contributes to the total loss through a weighted cross-entropy term, denoted as $w_{CE} \mathcal{L}_{CE}$, whose objective is to determine whether the volumetric patch x contains an aneurysm or corresponds to background tissue, is a classifier with two target classes: label 1 corresponds to aneurysm and label 0 corresponds to background tissue. Therefore, the total loss is defined as:

$$\mathcal{L}_{Total} = w_{CE} \mathcal{L}_{CE} + \sum_{s=1}^4 w_s \mathcal{L}^{(s)} \quad (24)$$

where w_s , $s = 1, \dots, 4$ are scale-dependent weights that ensure each decoder scale produces a semantically meaningful representation, and w_{CE} is a scalar weight controlling the relative contribution of the auxiliary classification loss. The auxiliary classification branch provides complementary supervision at the bottleneck level, promoting the learning of globally discriminative features related to aneurysm presence. In combination with deep supervision, this auxiliary signal reinforces both global semantic consistency and multi-scale spatial accuracy, resulting in improved segmentation performance compared with using either strategy independently.

In addition to the proposed deeply supervised 3D U-Net with an auxiliary classification branch, several comparison models were implemented to evaluate the contribution of each architectural component and to compare the proposed approach with transformer-based segmentation networks. First, a baseline 3D U-Net was trained using the same preprocessing, patch extraction, data augmentation, loss functions, optimizer, and five-fold cross-validation protocol. Two ablation variants were also evaluated: a model with the auxiliary classification branch but without deep supervision and a model with deep supervision but without the auxiliary classification branch. These ablation models were included to isolate the contribution of each component of the proposed architecture.

To provide a broader comparison with recent transformer-based medical image segmentation models, SwinUNETR and TransUNet3D were also implemented. SwinUNETR is a 3D encoder-decoder segmentation architecture that uses a hierarchical Swin Transformer encoder with shifted-window self-attention and connects multi-resolution transformer features to a convolutional decoder through skip connections [21]. This design allows the model to capture long-range contextual

information while preserving multi-scale spatial representations, which is useful for volumetric medical image segmentation.

TransUNet [22] was originally proposed as a hybrid CNN–Transformer architecture in which convolutional features are tokenized and processed by a Transformer encoder to model global context, while a U-Net-like decoder recovers spatial resolution through skip connections. In this work, a 3D adaptation of this principle, referred to as TransUNet3D, was used to process volumetric TOF-MRA patches. The input and output dimensions were adapted to the same 3D patch size used by the proposed model, and the final prediction layer was configured for binary aneurysm segmentation.

For a fair comparison, all models were trained and evaluated using the same patient-level five-fold cross-validation splits, preprocessing pipeline, patch size, intensity normalization, data augmentation strategy, and inference procedure. The same training and test partitions were used across models to enable paired statistical comparisons. All models were optimized using the same loss formulation, consisting of a combination of focal loss and generalized Dice loss, to ensure consistency in handling class imbalance and segmentation accuracy across architectures. No auxiliary classification output was added to SwinUNETR or TransUNet3D; these architectures were evaluated only as segmentation models.

2.4 Training

Training was performed on an NVIDIA H100 GPU, with a mean epoch duration of 92 seconds and a total training time of approximately 4 hours for 150 epochs. A five-fold cross-validation scheme was implemented. Folds were constructed to preserve a balanced distribution of small (<3 mm), medium (3–7 mm), and large (>7 mm) aneurysms. All aneurysms from a given patient were kept within the same subset to prevent information leakage across folds.

For the segmentation task, the loss function was defined as a weighted sum of generalized Dice loss and focal loss with $\alpha_k^{(s)} = 0.75$, $\gamma = 2$, $\lambda_{\text{dice}} = 0.55$, and $\lambda_{\text{focal}} = 0.45$. Deep supervision was implemented by attaching four intermediate decoder outputs, each contributing to the total loss with a decreasing weight w_s assigned from the top (highest-resolution) to the bottom (lowest-resolution) decoder outputs ($w_1 = 1.0$, $w_2 = 0.5$, $w_3 = 0.25$, and $w_4 = 0.1$, respectively), despite the decoder being constructed in a bottom-up manner. In addition, an auxiliary classification branch was incorporated, whose weighted cross-entropy loss was assigned a weight w_{CE} of 0.3. These weighted components were optimized jointly during training.

The model was trained using the AdamW optimizer [23] with an initial learning rate of 0.0005 and gradient clipping (clipnorm = 1.0). A step-decay learning rate schedule was applied, reducing the learning rate by a factor of 0.5 every 30 epochs to stabilize convergence during the later stages of training.

2.5 Inference strategy

During inference, the trained network was applied independently to every candidate patch extracted from the vascular skeleton. A probability threshold of 0.5 was used to obtain binary segmentations, and no additional post-processing was performed. Since the proposed framework operates at the patch level, predictions were

not reconstructed into complete image volumes. Instead, segmentation performance was evaluated by comparing each predicted patch with its corresponding ground-truth annotation. Patient-level metrics were subsequently obtained by aggregating the patch-level predictions belonging to each patient before computing the final performance measures.

2.6 Evaluation metrics and statistical analysis

Model performance was evaluated using complementary metrics at the overlap, boundary, lesion-detection, localization, and false-positive (FP) levels. This multi-metric evaluation was adopted because different metrics capture different aspects of segmentation quality and should be selected according to the target structure, clinical objective, and expected failure modes [24], [25].

Overlap-based performance was assessed using the Dice similarity coefficient and the intersection over union (IoU). Dice and IoU quantify the spatial agreement between the predicted segmentation mask and the reference annotation, with higher values indicating better segmentation performance. Precision and recall were computed from voxel-level true-positive, FP, and false-negative counts. Precision quantified the proportion of predicted aneurysm voxels that corresponded to the reference annotation, whereas recall quantified the proportion of reference aneurysm voxels detected by the model.

Boundary agreement was assessed using the 95th percentile Hausdorff distance (HD95) and the average surface distance (ASD), both expressed in millimeters. HD95 measures a robust upper-bound surface discrepancy between the predicted and reference masks, whereas ASD measures the mean distance between their surfaces. For HD95 and ASD, lower values indicate better boundary agreement [24].

In addition to voxel-wise segmentation metrics, lesion-level and localization-oriented metrics were computed. Lesion sensitivity was defined as the proportion of aneurysm-containing cases or approximate lesions for which the model produced a positive prediction overlapping the reference aneurysm region. Also, FP burden was quantified using the FP patch rate. The FP patch rate was defined as the proportion of aneurysm-free patches for which the model produced a positive segmentation.

Because multiple patches may originate from the same patient, statistical comparisons were performed at the patient level rather than treating patches as independent observations. Patch-level predictions were aggregated by patient using the patient identifier extracted from each patch filename. The resulting case-level table was grouped by model, fold, and patient identifier.

For the size-stratified analysis, aneurysms larger than 2.5 mm were analyzed using an approximate lesion-level table grouped by model, fold, patient identifier, and ground-truth aneurysm voxel count. Lesions with more than 130 ground-truth voxels were considered equivalent to aneurysms larger than 2.5 mm according to the voxel-size criterion used in this study.

For each metric, results are reported as mean values with 95% confidence intervals. Pairwise comparisons between the proposed model and each baseline were performed using matched observations. The Wilcoxon signed-rank test was used for paired statistical comparisons [26], and p-values were adjusted using the Holm correction separately within each metric to account for multiple pairwise comparisons [27].

3 RESULTS

The aneurysm diameter distributions in the test sets were broadly comparable across folds, although some differences in variance were observed (see Figure 4). In particular, fold 1 included an outlier, whereas fold 2 showed greater variance than the other folds.

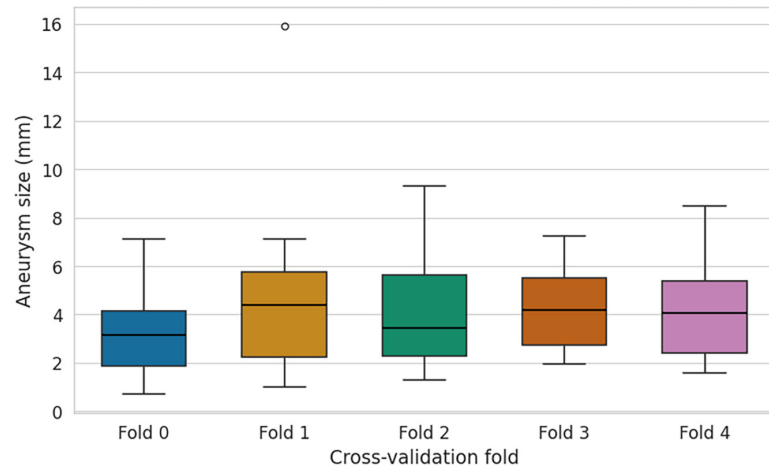


Fig. 4. Distribution of aneurysm diameters across folds

The proposed model was compared with the internal ablation variants, the baseline 3D U-Net, SwinUNETR, and TransUNet3D. Table 1 summarizes the case-level quantitative performance of all evaluated models, reporting the mean and 95% confidence interval for each metric. Table 2 presents the corresponding analysis restricted to aneurysms larger than 2.5 mm. Overall, the proposed model achieved the highest Dice coefficient, IoU, precision, and the lowest HD95, ASD, and FP patch rate, whereas SwinUNETR and TransUNet3D achieved higher recall and lesion sensitivity at the expense of substantially more FP detections.

Table 1. Case-level quantitative comparison of the evaluated models. Values are reported as mean [95% confidence interval]

Model	Dice	IoU	Precision	Recall	HD95 (mm)	ASD (mm)	Lesion Sensitivity	FP Patch Rate
Proposed model	0.425 [0.362, 0.488]	0.326 [0.274, 0.378]	0.704 [0.645, 0.763]	0.432 [0.364, 0.501]	2.525 [1.800, 3.249]	1.385 [0.721, 2.049]	0.680 [0.598, 0.762]	0.026 [0.017, 0.035]
Classifier without deep supervision	0.391 [0.335, 0.447]	0.297 [0.250, 0.343]	0.619 [0.561, 0.677]	0.447 [0.378, 0.517]	3.071 [2.147, 3.994]	1.806 [0.933, 2.679]	0.754 [0.679, 0.829]	0.036 [0.025, 0.047]
Deep supervision without classifier	0.370 [0.316, 0.424]	0.280 [0.235, 0.324]	0.567 [0.507, 0.628]	0.431 [0.362, 0.500]	3.240 [2.447, 4.033]	1.768 [1.068, 2.468]	0.744 [0.667, 0.822]	0.051 [0.038, 0.064]
3D U-Net	0.342 [0.290, 0.395]	0.255 [0.213, 0.297]	0.567 [0.507, 0.626]	0.406 [0.337, 0.476]	3.324 [2.553, 4.095]	1.899 [1.198, 2.599]	0.707 [0.625, 0.789]	0.050 [0.037, 0.063]
SwinUNETR	0.191 [0.168, 0.215]	0.142 [0.123, 0.161]	0.227 [0.197, 0.256]	0.546 [0.488, 0.605]	4.788 [3.982, 5.594]	2.097 [1.522, 2.672]	0.896 [0.845, 0.947]	0.546 [0.507, 0.585]
TransUNet3D	0.205 [0.175, 0.235]	0.153 [0.129, 0.178]	0.245 [0.212, 0.278]	0.527 [0.463, 0.592]	4.238 [3.466, 5.011]	1.977 [1.457, 2.498]	0.871 [0.817, 0.925]	0.529 [0.470, 0.588]

The case-level comparison showed that the proposed deeply supervised 3D U-Net with auxiliary classification achieved the best overall overlap and precision among the evaluated models. At the patient level, the proposed model obtained a Dice coefficient of 0.425 [95% CI: 0.362–0.488], IoU of 0.326 [0.274–0.378], precision of 0.704 [0.645–0.763], recall of 0.432 [0.364–0.501], HD95 of 2.525 mm [1.800–3.249], ASD of 1.385 mm [0.721–2.049], lesion sensitivity of 0.680 [0.598–0.762], localization accuracy of 0.688 [0.608–0.767], and FP patch rate of 0.026 [0.017–0.035].

Compared with SwinUNETR and TransUNet3D, the proposed model obtained substantially higher Dice, IoU, and precision. SwinUNETR achieved Dice = 0.191 and TransUNet3D achieved Dice = 0.205, whereas the proposed model achieved Dice = 0.425. The proposed model also produced fewer false positives, with a FP patch rate of 0.026 compared with 0.546 for SwinUNETR and 0.529 for TransUNet3D.

Table 2. Quantitative comparison for aneurysms larger than 2.5 mm

Model	Dice	IoU	Precision	Recall	HD95 (mm)	ASD (mm)	Lesion Sensitivity	FP Patch Rate
Proposed model	0.548 [0.485, 0.612]	0.427 [0.374, 0.481]	0.781 [0.734, 0.828]	0.536 [0.465, 0.607]	2.468 [1.663, 3.273]	1.070 [0.612, 1.528]	0.820 [0.747, 0.894]	0.030 [0.018, 0.042]
Classifier without deep supervision	0.467 [0.410, 0.523]	0.359 [0.311, 0.407]	0.663 [0.607, 0.719]	0.551 [0.479, 0.624]	2.625 [1.822, 3.427]	1.036 [0.655, 1.416]	0.891 [0.833, 0.949]	0.040 [0.026, 0.055]
Deep supervision without classifier	0.414 [0.357, 0.471]	0.319 [0.272, 0.366]	0.601 [0.541, 0.662]	0.509 [0.434, 0.584]	2.997 [2.107, 3.886]	1.480 [0.833, 2.128]	0.826 [0.751, 0.900]	0.054 [0.039, 0.069]
3D U-Net	0.378 [0.328, 0.428]	0.283 [0.243, 0.324]	0.589 [0.531, 0.647]	0.487 [0.416, 0.559]	2.631 [2.117, 3.146]	1.192 [0.831, 1.553]	0.835 [0.766, 0.903]	0.054 [0.039, 0.069]
SwinUNETR	0.178 [0.153, 0.204]	0.135 [0.115, 0.156]	0.221 [0.191, 0.251]	0.551 [0.484, 0.618]	3.849 [3.088, 4.611]	1.452 [1.018, 1.886]	0.910 [0.858, 0.963]	0.513 [0.472, 0.555]
TransUNet3D	0.197 [0.163, 0.231]	0.151 [0.124, 0.179]	0.229 [0.192, 0.265]	0.584 [0.513, 0.656]	3.330 [2.495, 4.165]	1.438 [0.915, 1.962]	0.906 [0.850, 0.963]	0.564 [0.502, 0.626]

Notes: Values are reported as mean [95% confidence interval]. Lesions larger than 2.5 mm were operationalized as those with more than 130 ground-truth voxels. For HD95, ASD, centroid distance, FP patch rate, and FP components, lower values indicate better performance.

When aneurysms larger than 2.5 mm were considered (refer to Table 2), the proposed model showed improved performance, reaching Dice = 0.548 [0.485–0.612], IoU = 0.427 [0.374–0.481], precision = 0.781 [0.734–0.828], recall = 0.536 [0.465–0.607], HD95 = 2.468 mm [1.663–3.273], and ASD = 1.070 mm [0.612–1.528]. In this subset, the proposed model significantly outperformed SwinUNETR and TransUNet3D in Dice, IoU, precision, HD95, ASD, and FP metrics after Holm correction. Recall was numerically lower than that of SwinUNETR and TransUNet3D, but the difference was not statistically significant after Holm correction. Lesion sensitivity remained higher for the transformer-based models, suggesting that they are more sensitive but less specific under the evaluated setting.

To determine whether the observed differences were statistically significant, paired Wilcoxon signed-rank tests with Holm correction were performed using matched observations. The statistical comparisons for all aneurysms and for aneurysms larger than 2.5 mm are presented in Tables 3 and 4, respectively.

Table 3. Pairwise statistical comparison between the proposed model and the evaluated models at the case level

Comparator	Dice	IoU	Precision	Recall	HD95	ASD	Lesion Sensitivity	FP Patch Rate
Classifier no deep supervision	↑ 0.002	↑ 0.004	↑ <0.001	↓ ns 0.636	↑ ns 0.054	↑ ns 0.067	↓ ns 0.059	↑ ns 0.068
Deep supervision no classifier	↑ <0.001	↑ <0.001	↑ <0.001	↑ ns 0.721	↑ 0.038	↑ 0.042	↓ ns 0.099	↑ <0.001
3D U-Net	↑ <0.001	↑ <0.001	↑ <0.001	↑ ns 0.636	↑ 0.021	↑ 0.042	↓ 0.023	↑ 0.001
SwinUNETR	↑ <0.001	↑ <0.001	↑ <0.001	↓ <0.001	↑ <0.001	↑ <0.001	↓ <0.001	↑ <0.001
TransUNet3D	↑ <0.001	↑ <0.001	↑ <0.001	↓ <0.001	↑ <0.001	↑ 0.002	↓ <0.001	↑ <0.001

Notes: Holm-adjusted p-values from paired Wilcoxon signed-rank tests. ↑ and ↓ indicate that the proposed model performed numerically better or worse than the comparator, respectively. “ns” denotes non-significant differences. For HD95, ASD, and FP patch rate, lower values indicate better performance.

Table 4. Pairwise statistical comparison for aneurysms larger than 2.5 mm using matched approximate lesion-level observations

Comparator	Dice	IoU	Precision	Recall	HD95	ASD	Lesion Sensitivity	FP Patch Rate
Classifier no deep supervision	↑ <0.001	↑ <0.001	↑ <0.001	↓ ns 0.248	↑ ns 0.415	↓ ns 0.409	↓ 0.050	↑ ns 0.106
Deep supervision no classifier	↑ <0.001	↑ <0.001	↑ <0.001	↑ ns 0.267	↑ ns 0.086	↑ 0.004	↓ ns 0.867	↑ <0.001
3D U-Net	↑ <0.001	↑ <0.001	↑ <0.001	↑ ns 0.699	↓ ns 0.086	↓ ns 0.160	↓ 0.050	↑ <0.001
SwinUNETR	↑ <0.001	↑ <0.001	↑ <0.001	↓ ns 0.699	↑ <0.001	↑ 0.005	↓ 0.004	↑ <0.001
TransUNet3D	↑ <0.001	↑ <0.001	↑ <0.001	↓ ns 0.114	↑ 0.024	↑ 0.018	↓ 0.050	↑ <0.001

Notes: Lesions larger than 2.5 mm were operationalized as those with more than 130 ground-truth voxels. Holm-adjusted p-values were obtained from paired Wilcoxon signed-rank tests. ↑ and ↓ indicate that the proposed model performed numerically better or worse than the comparator, respectively. “ns” denotes non-significant differences. For HD95, ASD, and FP patch rate, lower values indicate better performance.

The statistical analysis confirmed that the proposed model significantly outperformed the baseline 3D U-Net and the transformer-based architectures in Dice, IoU, precision, and FP metrics. In contrast, SwinUNETR and TransUNet3D achieved significantly higher recall and lesion sensitivity, reflecting a different operating point that favors sensitivity at the expense of substantially more FP detections.

Figure 5 summarizes the Dice scores obtained by the evaluated 3D U-Net configurations for the training, validation, and test sets across the five cross-validation folds. The proposed model consistently achieved the highest validation and test Dice scores, while all configurations showed improved performance for aneurysms larger than 2.5 mm.

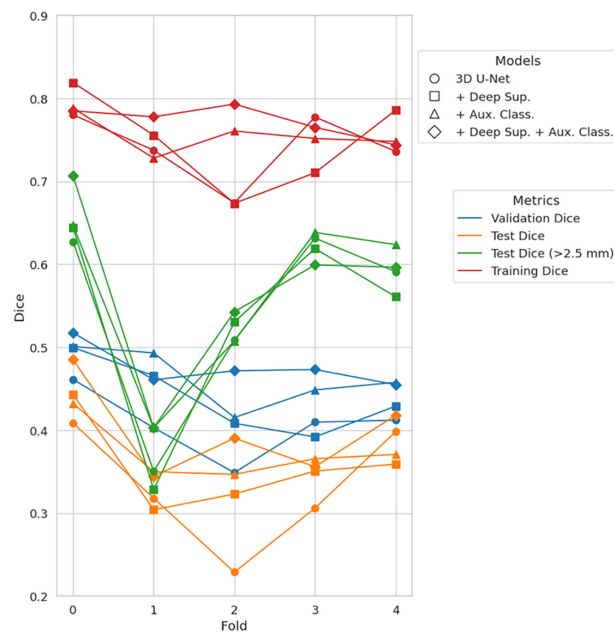


Fig. 5. Fold-wise Dice scores for the 3D U-Net configurations evaluated in this study

Note: Different marker shapes indicate the model configuration, and colors distinguish the reported Dice metrics.

Representative qualitative examples are shown in Figure 6 to illustrate typical successful and challenging segmentation cases. The first and third patches correspond to the same aneurysm but produced markedly different Dice scores (0.823 and 0.000, respectively). The associated auxiliary classifier probabilities (0.54 and 0.49) were close to the decision threshold, resulting in different segmentation outcomes.

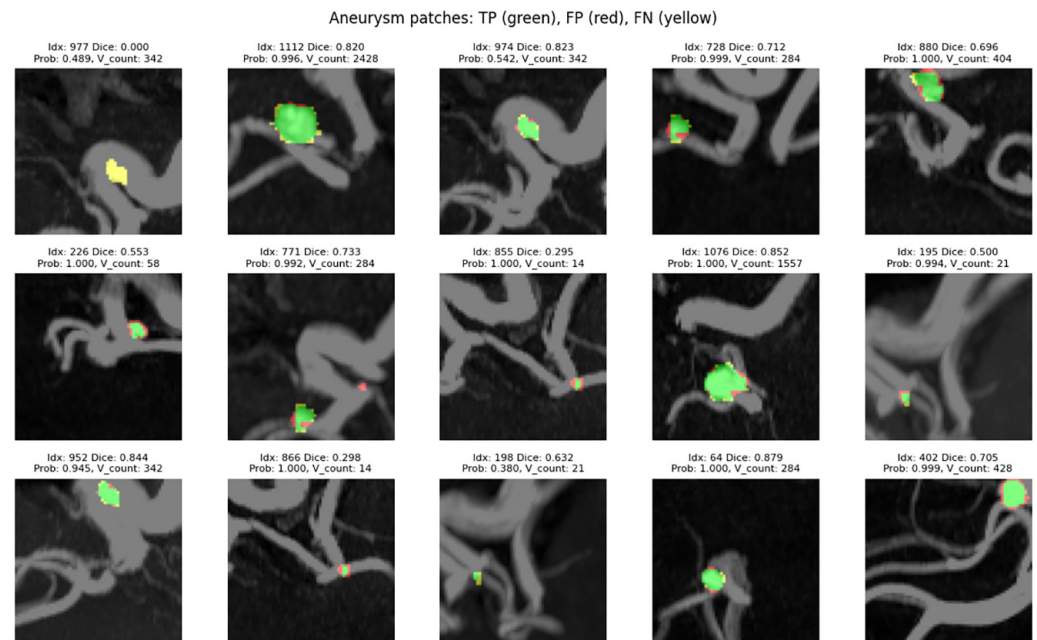


Fig. 6. Representative test patches containing aneurysms and their segmentation results

Notes: Each panel displays a maximum-intensity projection of a 3D TOF-MRA patch along one axis, together with the corresponding voxel-level segmentation. True positives (TP) are shown in green, false positives (FP) in red, and false negatives (FN) in yellow. The Dice score, auxiliary classifier probability (Prob), and aneurysm voxel count (V_count) are also indicated for each patch.

4 DISCUSSION

The main contribution of this work is not the introduction of a new network architecture but the systematic integration of established techniques (including 3D U-Net, deep supervision, auxiliary classification, and combined loss functions) into a unified framework tailored for intracranial aneurysm segmentation. This design specifically addresses key challenges of the ADAM dataset, including small lesion size, severe class imbalance, and heterogeneous image quality. Furthermore, the study provides a reproducible evaluation protocol based on five-fold cross-validation together with a comprehensive size-stratified performance analysis. These contributions offer a clearer understanding of model behavior across aneurysm sizes and facilitate more consistent comparisons with future studies using the same benchmark dataset.

In general, models that incorporate both deep supervision and the auxiliary classifier show better performance and lower variability than the baseline model. All configurations also improved when segmenting aneurysms larger than 2.5 mm, suggesting that the proposed framework is particularly effective for larger lesions.

The observed improvement obtained by combining deep supervision with auxiliary classification is consistent with the complementary roles of these two components. Deep supervision facilitates optimization by providing stronger gradient signals at multiple decoder levels, encouraging the learning of discriminative multiscale features. Meanwhile, the auxiliary classification branch guides the network to distinguish patches containing aneurysms from background regions, promoting more informative feature representations and reducing FP activations. Together, these mechanisms likely contribute to the improved segmentation stability and higher precision observed across the cross-validation folds.

However, fold 1 shows a marked decrease in Dice score, which may be related to the presence of a 16-mm aneurysm in the test set (see Figure 4). Because this case was not included in the corresponding training partition, it may have posed an additional generalization challenge for the models.

It is also important to consider the characteristics of the evaluation protocol when interpreting these results. Each test fold contains only 18 cases, making it difficult to achieve perfectly balanced data distributions across folds. In addition, some patients present multiple aneurysms, further increasing dataset heterogeneity and the complexity of the segmentation task. Although a leave-one-out strategy could reduce partition-related variability, it would require substantially greater computational resources and longer training times.

The comparison with the evaluated transformer-based architectures further highlights the behavior of the proposed framework. The proposed framework consistently favored precision and boundary agreement over sensitivity. While SwinUNETR and TransUNet3D detected a larger proportion of aneurysm-containing patches, they also generated substantially more false positives. This trade-off suggests that the proposed CNN-based framework adopts a more conservative segmentation strategy, which may be advantageous in clinical workflows where reducing FP findings is important for minimizing unnecessary image review.

The statistical analysis further demonstrated that the improvements in Dice, IoU, precision, and FP rate over the baseline 3D U-Net were significant after Holm correction, supporting that the observed gains were consistent across the evaluated cases rather than arising from fold-to-fold variability.

To facilitate more consistent comparisons across future studies, it would be beneficial for researchers using the ADAM dataset to report or publicly share the exact

cross-validation splits employed in their experiments. Standardized evaluation protocols would improve reproducibility, reduce variability introduced by different data partitions, and enable fairer benchmarking of segmentation methods.

The qualitative examples in Figure 6 demonstrate that accurate aneurysm localization can still be achieved even when voxel-level segmentation is imperfect. As expected, patches containing fewer aneurysm voxels remain more difficult to segment accurately. From a clinical perspective, however, detecting the presence and approximate location of an aneurysm may be more valuable than achieving a perfectly accurate boundary, particularly during the initial screening stage. This observation supports the potential usefulness of weak-label strategies, especially in applications where reliable localization is more critical than precise voxel-level delineation.

Although the proposed framework demonstrates consistent improvements over the baseline and provides meaningful localization capability, the overall segmentation performance remains limited, with an average test Dice score of approximately 0.40, which is still insufficient for reliable clinical-grade aneurysm segmentation. These findings indicate that, despite the observed performance gains, intracranial aneurysm segmentation remains a highly challenging task because of the small size, anatomical variability, and heterogeneity of aneurysms in the ADAM dataset.

Despite these limitations, the experimental results suggest that the proposed framework achieves a more conservative segmentation behavior, favoring precision, boundary agreement, and FP reduction over aggressive lesion delineation. This characteristic is particularly relevant for intracranial aneurysm segmentation, where excessive false positives may reduce clinical confidence and increase the burden of image review. Nevertheless, the relatively low Dice score obtained on the complete test set indicates that the proposed model is not yet suitable for standalone clinical deployment. Instead, its current value lies in assisting clinicians by localizing aneurysm candidates and providing an initial segmentation, particularly for aneurysms larger than 2.5 mm.

A limitation of this study is the relatively small size of the available dataset. However, this limitation reflects the current scarcity of publicly available datasets with voxel-level annotations for intracranial aneurysm segmentation. The ADAM dataset remains one of the few publicly accessible benchmarks of this type and is particularly challenging because of heterogeneous acquisition protocols, varying spatial resolutions, and the large proportion of small aneurysms. These characteristics increase the difficulty of the segmentation task while also better representing the variability encountered in clinical practice. Future work should evaluate the proposed framework on larger multicenter datasets with expert annotations to further improve robustness and generalization.

5 DATA AVAILABILITY

The preprocessed dataset used in this study is publicly available at the following GitHub repository: https://github.com/josedaviddr/Aneurysm_segmentation_DataSet_folds.

The repository contains the five data folds used in the experiments, organized into training, validation, and test subsets. Each fold includes 3D image patches and their corresponding binary segmentation masks stored in NumPy format (.npy). The patches have a size of $64 \times 64 \times 64$ voxels and a spatial resolution of 0.4 mm per voxel.

These patches were generated through the preprocessing procedure described in the methodology section of this paper. The original images were obtained from the ADAM challenge dataset, introduced by Timmins et al. in “Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAs: The ADAM challenge” [11]. The shared repository does not replace the original ADAM dataset; rather, it provides the preprocessed patch-level folds used in this work to facilitate reproducibility and comparison under the same experimental conditions.

6 CONCLUSION

A deeply supervised 3D U-Net with an auxiliary classifier was evaluated for cerebral aneurysm segmentation using the ADAM challenge dataset. The proposed framework achieved a patient-level Dice coefficient of 0.425 (95% CI: 0.362–0.488) and an FP patch rate of 0.026 (95% CI: 0.017–0.035). For aneurysms larger than 2.5 mm, the Dice coefficient increased to 0.548 (95% CI: 0.485–0.612), while the FP patch rate remained low at 0.030 (95% CI: 0.018–0.042). Compared with the baseline 3D U-Net, SwinUNETR, and TransUNet3D, the proposed model showed significant improvements in Dice, IoU, precision, boundary-based metrics, and FP reduction. However, these gains were accompanied by lower recall and lesion sensitivity than some competing models, particularly the transformer-based architectures. This suggests that the proposed model tends to produce more selective segmentations, reducing false positives at the cost of missed detections in some cases.

Although voxel-level segmentation remains challenging, particularly for small aneurysms, the qualitative results show that the proposed framework can often localize aneurysm regions even when precise boundary delineation is not achieved. This behavior may be useful in a computer-assisted detection setting, especially for clinically relevant aneurysms larger than 2.5 mm. Future work should therefore focus on improving sensitivity to small or low-contrast aneurysms while preserving the observed reduction in FP predictions. Further validation on larger multicenter datasets will also be necessary to assess the generalization of the proposed approach under more diverse imaging conditions.

7 REFERENCES

- [1] M. T. Lawton and G. E. Vates, “Subarachnoid hemorrhage,” *New England Journal of Medicine*, vol. 377, no. 3, pp. 257–266, 2017. <https://doi.org/10.1056/NEJMcp1605827>
- [2] T. Ingall, K. Asplund, M. Mähönen, and R. Bonita, “A multinational comparison of subarachnoid hemorrhage epidemiology in the WHO MONICA Stroke study,” *Stroke*, vol. 31, no. 5, pp. 1054–1061, 2000. <https://doi.org/10.1161/01.STR.31.5.1054>
- [3] N. Etminan and G. J. Rinkel, “Unruptured intracranial aneurysms: Development, rupture and preventive management,” *Nat. Rev. Neurol.*, vol. 12, pp. 699–713, 2016. <https://doi.org/10.1038/nrneurol.2016.150>
- [4] B. M. Howard, R. Hu, J. W. Barrow, and D. L. Barrow, “Comprehensive review of imaging of intracranial aneurysms and angiographically negative subarachnoid hemorrhage,” *Neurosurg. Focus*, vol. 47, no. 6, pp. 1–13, 2019. <https://doi.org/10.3171/2019.9.FOCUS19653>
- [5] Z. Shi et al., “Artificial intelligence in the management of intracranial aneurysms: Current status and future perspectives,” *American Journal of Neuroradiology*, vol. 41, no. 3, pp. 373–379, 2020. <https://doi.org/10.3174/ajnr.A6468>

- [6] M. Din, S. Agarwal, M. Grzeda, D. A. Wood, M. Modat, and T. C. Booth, "Detection of cerebral aneurysms using artificial intelligence: A systematic review and meta-analysis," *Journal of Neurointerventional Surgery*, vol. 15, no. 3, pp. 262–271, 2023. <https://doi.org/10.1136/jnis-2022-019456>
- [7] Ž. Bizjak and Ž. Špiclin, "A systematic review of deep-learning methods for intracranial aneurysm detection in CT angiography," *Biomedicines*, vol. 11, no. 11, Art. no. 2921, 2023. <https://doi.org/10.3390/biomedicines11112921>
- [8] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021. <https://doi.org/10.1038/s41592-020-01008-z>
- [9] J. Chen, "TransUNet: Transformers make strong encoders for medical image segmentation," Github, 2026. [Online]. Available: <https://github.com/Beckschen/>
- [10] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021. <https://doi.org/10.48550/arXiv.2105.05537>
- [11] K. M. Timmins et al., "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge," *Neuroimage*, vol. 238, Art. no. 118216, 2021. <https://doi.org/10.1016/j.neuroimage.2021.118216>
- [12] S. Ham 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.*, vol. 13, no. 1, Art. no. 12018, 2023. <https://doi.org/10.1038/s41598-023-38586-9>
- [13] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference*, Athens, Greece, Berlin, Heidelberg: Springer-Verlag, 2016, pp. 424–432. https://doi.org/10.1007/978-3-319-46723-8_49
- [14] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Berlin, Heidelberg: Springer, 2013, pp. 411–418. https://doi.org/10.1007/978-3-642-40763-5_51
- [15] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Comput. Vis. Graph. Image Process.*, vol. 29, no. 3, pp. 273–285, 1985. [https://doi.org/10.1016/0734-189X\(85\)90125-2](https://doi.org/10.1016/0734-189X(85)90125-2)
- [16] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 4, pp. 532–550, 1987. <https://doi.org/10.1109/TPAMI.1987.4767941>
- [17] P. Maragos and R. Schafer, "Morphological skeleton representation and coding of binary images," *IEEE Trans. Acoust.*, vol. 34, no. 5, pp. 1228–1244, 1986. <https://doi.org/10.1109/TASSP.1986.1164959>
- [18] J. Yang, B. Wu, L. Li, P. Cao, and O. Zaiane, "MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT," *Computerized Medical Imaging and Graphics*, vol. 92, Art. no. 101957, 2021. <https://doi.org/10.1016/j.compmedimag.2021.101957>
- [19] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, in Lecture Notes in Computer Science, vol. 10553, Cham, Switzerland: Springer, 2017, pp. 240–248. https://doi.org/10.1007/978-3-319-67558-9_28
- [20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007. <https://doi.org/10.1109/ICCV.2017.324>

- [21] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021*, in Lecture Notes in Computer Science, A. Crimi and S. Bakas, Eds., vol. 12962, Springer, Cham, 2022. https://doi.org/10.1007/978-3-031-08999-2_22
- [22] J. Chen *et al.*, "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Med. Image Anal.*, vol. 97, Art. no. 103280, 2024. <https://doi.org/10.1016/j.media.2024.103280>
- [23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [24] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imaging*, vol. 15, no. 1, Art. no. 29, 2015. <https://doi.org/10.1186/s12880-015-0068-x>
- [25] L. Maier-Hein *et al.*, "Metrics reloaded: Recommendations for image analysis validation," *Nat. Methods*, vol. 21, no. 2, pp. 195–212, 2024. <https://doi.org/10.1038/s41592-023-02151-z>
- [26] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. <https://doi.org/10.2307/3001968>
- [27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

8 AUTHORS

Mario Mata-Castillo is a doctoral candidate at the Autonomous University of Ciudad Juárez (UACJ), Chihuahua, México, and a professor at the Autonomous University of Chihuahua. His research focuses on artificial intelligence in medical imaging, with an emphasis on deep learning-based methods for biomedical image segmentation. His interests span machine learning and the application of AI techniques to improve diagnostic processes and healthcare outcomes (E-mail: al221148@alumnos.uacj.mx).

Nelly Gordillo-Castillo is a full-time Professor at the Autonomous University of Ciudad Juárez (UACJ), México, and Head of the Department of Electrical and Computer Engineering. Her research focuses on medical image processing and artificial intelligence for the detection of neurological pathologies. She is a member of the National System of Researchers (SNII Level 2) (E-mail: nellygordillo@uacj.mx).

Boris Mederos is a full-time Professor at the Autonomous University of Ciudad Juárez. His work is in the area of applied mathematics and artificial intelligence, with interests in machine learning, image processing, pattern recognition, and mathematical modeling. He earned both his Ph.D. and M.Sc. in Mathematics from IMPA, Brazil, and completed postdoctoral appointments at the University of California, Davis, and at IMPA itself. His academic work combines theoretical foundations, computational development, and applications to scientific and technological problems (E-mail: boris.mederos@uacj.mx).

Dora-Luz Flores is a Full Professor and currently serves as the General Coordinator of Research and Graduate Studies at the Autonomous University of Baja California (UABC), México. She has a trajectory in biomedical engineering and research management. Her work focuses on artificial intelligence applied to biological systems, computational modeling, and technology transfer. She has led multiple national and international collaborations, particularly in cross-border initiatives. Her research interests include AI in health, innovation policy, graduate education, and scientific development strategies (E-mail: dflores@uabc.edu.mx).

Vicente García is a full-time Professor at the Department of Electrical and Computer Engineering at the Autonomous University of Ciudad Juárez (UACJ), México, and Coordinator of the Bachelor's Program in Data Science and Artificial Intelligence. He holds a Ph.D. degree. His research focuses on Machine Learning and Data Mining, particularly on classification, data preprocessing, data complexity, and predictive model evaluation, with emphasis on class imbalance and bias and fairness in artificial intelligence (E-mail: vicente.jimenez@uacj.mx).

José Díaz-Román is a full-time Professor at the Autonomous University of Ciudad Juárez (UACJ), Chihuahua, México, and Coordinator of the Doctorate in Advanced Engineering Sciences. His research interests include digital signal and image processing, artificial intelligence, machine learning, deep learning, biomedical instrumentation, and wireless sensor networks (E-mail: david.roman@uacj.mx).