A Cascaded 2D/3D Deep Learning Pipeline for Valve Artifact Segmentation, Catheter Segmentation, and Length Estimation in Brain MRI

Travis Tran

***Abstract*—Ventricular shunts are used to drain CSF from the brain, but their metallic valve component can produce strong MRI artifacts that obscure the surrounding anatomy and parts of the catheter. This project develops a cascaded pipeline for three related tasks: shunt valve artifact detection (Task A), catheter segmentation (Task B), and catheter length estimation (Task C). Task A utilizes a 2D U-Net with an ImageNet-pretrained ResNet34 encoder. Task B uses a 3D MONAI DynUNet that takes the MRI and Task A artifact mask output as inputs. Task C follows a classical pipeline of 3D skeletonization, cubic-spline arc-length integration along the ordered centerline, and a straight-line bridge term across the artifact gap. On a 30-subject validation split, a mean Dice of 0.9194 was achieved for Task A, a mean Dice of 0.7929 (median HD95 of 1.41 mm) was achieved for Task B, and a mean absolute length error of 5.40 mm was achieved for Task C.**

***Index Terms*—brain MRI, shunt valve artifact, catheter segmentation, catheter length estimation, cascaded segmentation, U-Net, DynUNet, centerline extraction**

I. Introduction

A ventricular shunt is a catheter-based device that is used to drain excess CSF from the brain, and is a standard treatment for hydrocephalus [1]. Typically, a catheter is inserted into a lateral ventricle and connected to a valve beneath the scalp which controls fluid flow. The catheter is used to route CSF away from the brain, usually to the abdominal cavity. On MRI, the valve can create a strong artifact since it has metallic components, causing local magnetic field inhomogeneity in T1-weighted sequences [2]. The artifact can obscure brain anatomy close to it and can also hide portions of the catheter which pass through.

Catheter trajectory is relevant because its placement affects whether the shunt can drain CSF effectively while minimizing the disruption to surrounding brain tissue. The problem itself is non-trivial because the two structures (the valve and catheter) have very different image characteristics. On one hand, the valve artifact is large, high-contrast, and compact, while the catheter is thin, low-contrast, and visible only across a small number of voxels.

This project is divided into three tasks: (A) binary valve artifact segmentation, (B) binary catheter segmentation (outside the artifact region), and (C) total catheter length estimation. A cascaded pipeline in which the artifact prediction is used for both downstream tasks was adopted. The predicted Task A mask is passed into the catheter segmentation network as an input for Task B, and is also used in Task C to estimate the hidden catheter segment.

II. Background and Theory

In this project, two structurally different objects require segmentation: the valve artifact (appearing as a locally compact signal void in axial slices) and the catheter (appearing as a thin structure extending continuously across slices). U-Net architectures are commonly used in biomedical segmentation because they combine an encoder path (which provides context) with a decoder path (for localization purposes). The encoder and decoder are then linked by skip connections which preserve spatial resolution [3].

With only 170 training subjects, a randomly initialized encoder runs the risk of poor generalization. Thus, a ResNet34 encoder pretrained on ImageNet is utilized for Task A, providing multi-scale feature representations from the beginning [4]. For Task B, a 3D DynUNet from the MONAI framework is used since it features dynamic-architecture 3D U-Net designs well suited for patch-based medical image segmentation with deep supervision [5].

Task C requires extracting a 1D centerline from the predicted catheter mask. The process of 3D skeletonization reduces a binary volume to a 1D voxel medial axis while preserving the topological structure of the object [6]. The resulting skeleton approximates the catheter centerline and is the input into the spline fitting and arc-length calculation described in Section III-D.

Segmentation is evaluated using the Dice similarity coefficient,

$$\mathrm{DSC}(Y,\hat{Y}) = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}, \tag{1}$$

where Y is the ground-truth mask and $\hat{Y}$ is the predicted mask. Additionally, Task B is evaluated using HD95. Let $d(p,S) = \min_{q\in S} \|p-q\|_2$ denote point-to-set distance. Then,

$$\mathrm{HD95}(Y,\hat{Y}) = \mathrm{P}_{95}\big(\{d(p,\partial\hat{Y}) : p \in \partial Y\} \cup \{d(q,\partial Y) : q \in \partial\hat{Y}\}\big), \tag{2}$$

where ∂Y and $\partial\hat{Y}$ denote the surface voxel sets of Y and $\hat{Y}$, and P_{95} denotes the 95th percentile [7]. The HD95 metric is less sensitive to outlier voxels than the maximum Hausdorff distance. This is important for thin structures because single-voxel shifts may dominate the metric. Task C is evaluated using relative absolute error,

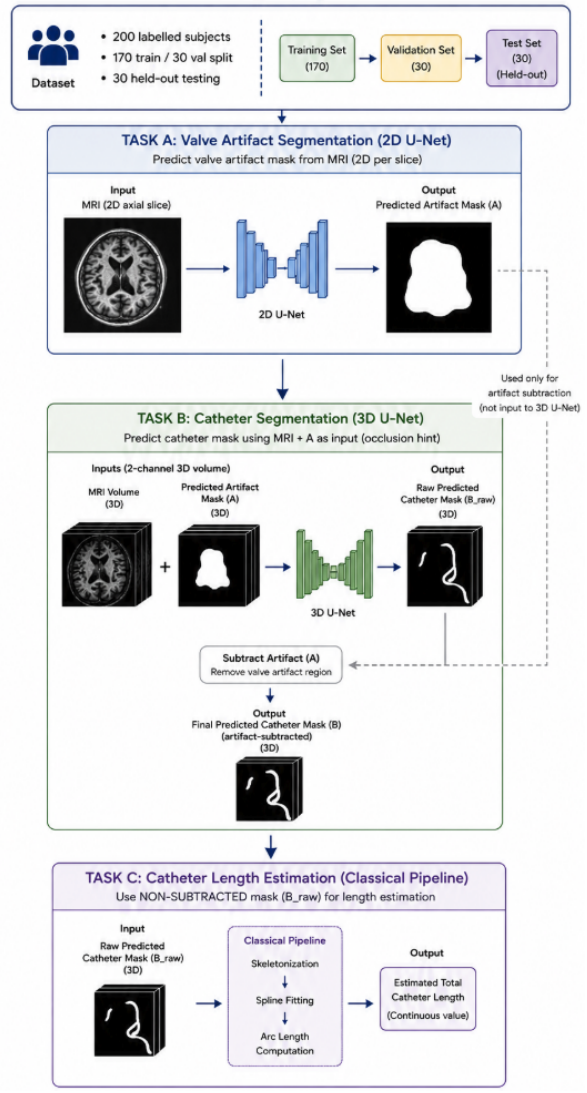


Fig. 1. An overview of the cascaded pipeline. The MRI volume is passed through Task A for artifact segmentation, then used as a second input for Task B catheter segmentation. Task C uses the catheter and artifact masks to estimate visible spline length and the artifact-gap bridge term.

$$\mathrm{RAE} = \frac{|\hat{L} - L|}{L}, \tag{3}$$

where $\hat{L}$ is the predicted length and L is the ground-truth.

III. Methods

The cascaded pipeline illustrated in Fig. 1 takes a 3D MRI volume in NIfTI (`.nii.gz`) format and produces a binary valve artifact mask, a binary catheter mask, and a catheter length estimate in mm. The dataset contains 200 labeled subjects while 30 are held out for testing. A fixed random split produced 170 training and 30 validation subjects. All volumes are $192 \times 224 \times 192$ voxels at 1 mm isotropic spacing.

A. Preprocessing

All MRI volumes are preprocessed using percentile clipping (0.5th and 99.5th percentiles), z-score normalization, and contrast-limited adaptive histogram equalization (CLAHE) with a clip limit of 0.03. N4 bias-field correction was considered as an additional step since it can reduce MRI intensity nonuniformity [8]. Ultimately, it was excluded from the final pipeline because of the high computational cost and negligible downstream impact.

B. Task A: Valve Artifact Segmentation

Task A utilizes a 2D U-Net from the `segmentation_models_pytorch` library with a ResNet34 encoder and ImageNet pretrained weights (24M params). The model operates on all axial MRI slices. Each single-channel slice is replicated into three channels to match the encoder's expected input format. The model was trained for 15 epochs using a combined loss of $0.5 \times$ Dice $+ 0.5 \times$ BCE, Adam optimizer (learning rate 3×10^{-4}), cosine annealing, automatic mixed precision, gradient clipping at 1.0, and 85% foreground oversampling. Augmentation includes random horizontal and vertical flips and intensity jitter.

At inference, a 4-flip test-time augmentation (TTA) was applied for all combinations. The flipped probability maps were then unflipped and averaged, then thresholded at 0.55. Post-processing kept the largest connected component and applied a binary hole filling. A threshold sweep from 0.40 to 0.60 showed a DSC range of only 0.0007, meaning that threshold choice has minimal effect on performance.

C. Task B: Catheter Segmentation

Task B uses a 3D MONAI DynUNet with two input channels (the preprocessed MRI volume and predicted Task A artifact mask). The cascade design was chosen because the artifact mask provides an occlusion hint indicating where the catheter is likely hidden or distorted. A single-channel model (only MRI) was evaluated as a baseline. The two-channel design improved validation DSC by around 0.05. The network had five stages, $3 \times 3 \times 3$ kernels, residual blocks, and deep supervision with three output heads.

The model was trained for 15 epochs on 64^3 patches with batch size 4 and learning rate 1×10^{-4}. The loss function combined Dice with weighted BCE (positive class weight 20), averaged across the three deep supervision heads. A foreground oversampling rate of 0.95 was used to address the class imbalance between the catheter voxels and background. During training, deep supervision targets were downsampled via trilinear interpolation to match the lower-resolution output heads. This produced values outside $[0, 1]$ in some cases, causing unstable BCE values. Clamping the downsampled targets to $[0, 1]$ fixed the issue and stabilized training beyond epoch 5.

Inference used Gaussian-weighted sliding-window inference with a stride of 32 voxels and 8-flip 3D TTA. Probability maps were unflipped, averaged, and thresholded at 0.45. Post-processing kept the largest connected component and

applied skeleton endpoint extension to recover under-predicted catheter ends. The two following masks were maintained: a reporting mask retaining in-artifact predictions to match ground-truth labels for Task B evaluation, and a length mask zeroing voxels overlapping the predicted artifact (to prevent double-counting for Task C).

D. Task C: Catheter Length Estimation

Task C uses a classical geometry pipeline operating on the length mask from Task B. First, the catheter mask was dilated with a radius-1 ball structuring element to reduce small gaps, then skeletonized using 3D medial-axis thinning [6]. Branch points, voxels with more than two 26-connected skeleton neighbors, were removed, and the largest remaining connected component was kept.

Endpoints were found with a two-pass farthest-point strategy. Starting from a random seed, the algorithm finds the farthest skeleton voxel A by geodesic distance, then finds the farthest voxel B from A. Skeleton voxels were ordered from A to B using a greedy nearest-neighbor walk. Endpoint extension was then applied at both ends by walking outward along the tangent direction until the catheter mask boundary was reached.

A cubic spline was fit through the ordered centerline points and the visible catheter length was computed as the spline arc length,

$$L_{\text{spline}} = \int_0^1 \left\| \frac{d\mathbf{s}(t)}{dt} \right\|_2 dt, \tag{4}$$

where $\mathbf{s}(t)$ is the fitted 3D spline. A bridge term estimates the hidden catheter segment inside the artifact as the Euclidean distance from the mean position of catheter voxels at the catheter-artifact interface to the artifact centroid,

$$L_{\text{bridge}} = \|\bar{\mathbf{v}} - \bar{\mathbf{c}}\|_2 , \tag{5}$$

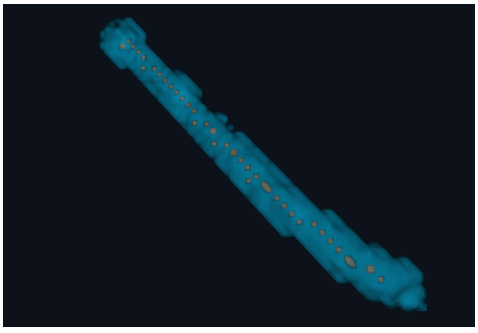
where $\bar{\mathbf{c}}$ is the mean position of catheter voxels overlapping the dilated artifact boundary and $\bar{\mathbf{v}}$ is the mean position of all artifact voxels. The total estimated length is $L_{\text{total}} = L_{\text{spline}} + L_{\text{bridge}}$.

IV. Experiments and Results

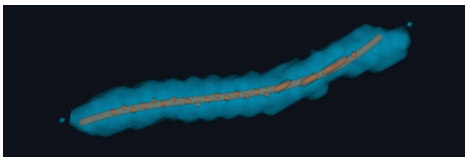
All experiments were performed on the provided labeled training data using the 170/30 training-validation split (the held-out competition test results are not included here). Table I summarizes the final validation performance across all three tasks.

A. Task A Results

Qualitatively, the predicted artifact masks closely match the main void region in all visualized subjects. As shown in Fig. 3, the remaining errors are concentrated at the artifact boundary rather than the interior. The largest connected component and hole-filling post-processing prevents stray false-positive blobs distant from the true artifact.



(a) Skeletonized catheter with farthest-point endpoints.



(b) Cubic spline fitted through the ordered centerline.

Fig. 2. Task C centerline extraction. The skeletonized mask is pruned and ordered between its two farthest endpoints (a), then a cubic spline is fit through the ordered centerline and arc length is integrated numerically (b).

TABLE I
Validation Performance Before Demo Day.

Task	Metric	Result
A	Mean Dice	0.9194
A	Dice std.	0.013
B	Mean Dice	0.7929
B	Median HD95	1.41 mm
C	MAE	5.40 mm

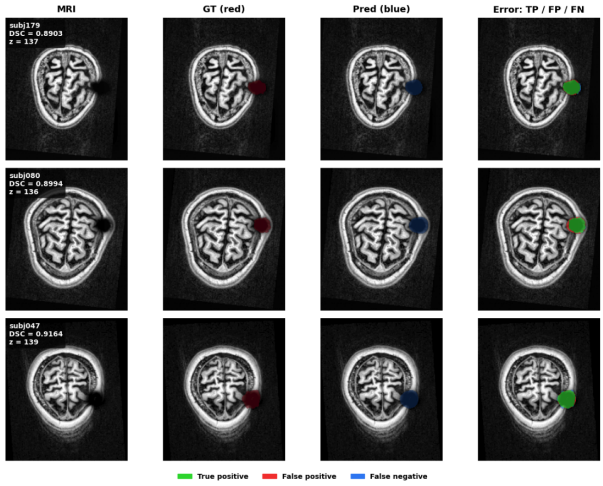


Fig. 3. Task A validation results for three subjects. Columns show axial MRI slice, ground-truth artifact mask (red), predicted artifact mask (blue), and error map with true positive, false positive, and false negative regions.

B. Task B Results

The Task B catheter segmentation model achieved a mean Dice of 0.7929 and a median Dice of 0.8023. The median HD95 was 1.41 mm, and 27/30 validation subjects had HD95 below 10 mm. Table II shows the incremental DSC improve-

TABLE II
TASK B CATHETER SEGMENTATION ABLATION ON VALIDATION SET.

Configuration	Mean Dice
3D U-Net baseline	0.704
DynUNet v2 (deep supervision, two-channel)	0.755
+ strict post-processing	0.767
+ 8-flip TTA, stride 32, endpoint extension	0.793

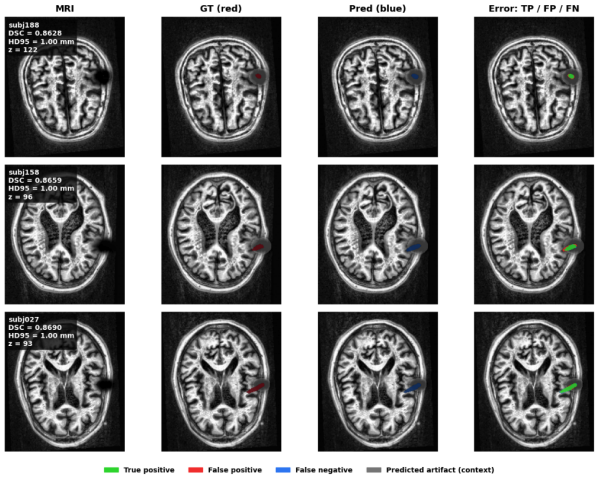


Fig. 4. Task B validation results for three subjects. Columns show axial MRI slice, ground-truth catheter mask (red), predicted catheter mask (blue), and error map with true positive, false positive, false negative, and predicted artifact context.

ment across the development stages.

The largest single gain came from replacing the 3D U-Net baseline with a DynUNet, and the final combination of test-time augmentation and endpoint extension achieved a Dice of 0.793. In several subjects, the predicted catheter followed the correct trajectory but was offset by one or two voxels. This produces a large DSC penalty since the catheter mask occupies few total voxels. Fig. 4 shows validation examples with true positive, false positive, and false negative regions.

C. Task C Results

Task C length estimation achieved a mean absolute error of 5.40 mm on the 30 validation subjects using predicted masks from the full cascade. Ground-truth catheter lengths ranged from around 40 to 77 mm. A sanity check with the ground-truth catheter masks reproduced lengths within 1-2 mm, meaning that the residual 5.40 mm MAE was caused by upstream segmentation errors in Task B.

V. DISCUSSION

The ImageNet-pretrained ResNet34 encoder was the key design decision for Task A. With only 170 training subjects, the pretrained weights gave feature representations that a randomly initialized encoder cannot reliably learn from scratch.

Task B was the most challenging task but still successful. The catheter occupies a small fraction of the volume and is sometimes only one or two voxels wide, making it difficult to segment with high Dice overlap. The predicted artifact mask gave the network a spatial prior about where the catheter was likely occluded, and the approximate 0.05 DSC improvement over the single-channel baseline confirms that this information was useful. The deep-supervision target clamp was a critical fix for Task B. Without it, interpolated targets would have fallen outside $[0, 1]$ and corrupted loss values and checkpoint selection. Dice is a harsh metric for thin structures. A one-voxel shift can substantially drop the score even when the trajectory is clinically correct, which is why the 1.41 mm median HD95 is more informative.

Future improvements to Task B, such as reducing false positive branches, improving catheter continuity near the artifact, or training with centerline-aware loss functions, would directly reduce Task C error. The two-mask design (reporting mask for Task B DSC, length mask for Task C) was also important for consistency. Removing the in-artifact voxels from the length mask prevented double-counting between the visible spline and the bridge term.

Demo-day performance: The full cascaded pipeline was run on the 30 held-out test subjects during the project-day session. The submitted predictions achieved a Task A Dice of 0.9156, a Task B Dice of 0.8368, and a Task C relative absolute error of 0.0913, placing the group first of nine teams. The Task A test performance was very close to the validation mean (0.9194), confirming good generalization of the model. Task B test performance (0.8368) was noticeably above the validation mean (0.7929), suggesting that the two-channel cascade design and the 8-flip TTA were particularly effective on the held-out distribution. The Task C relative absolute error of 0.0913 was the lowest of all groups and corresponds to roughly 3.5-7 mm of absolute length error given the 40-77 mm ground-truth range. This is consistent with the 5.40 mm MAE observed on the validation set. Many other groups used similar 2D/3D U-Nets, CC-filtering, and z-score normalization. Notable differences included nnU-Net for Task B, morphological opening, and snake-based or learned refinement for Task C. With more time, two high-priority changes would be adding patch-based training for Task A to increase Dice and applying morphological opening as a post-processing step for Task B to reduce false-positive catheter branches.

VI. TEAM MEMBER CONTRIBUTIONS

Shreeram, Tyler, and I developed the project two methodology, while the entirety of project one was worked on by Amy and Jazmine. Shreeram and I worked on the segmentation side of the pipeline, including the U-Net and DynUNet models, inference, post-processing, validation evaluation, and result visualization. I worked on Task A while Shreeram worked on Task B. Tyler worked on the catheter length-estimation method, including skeletonization, point ordering, spline fitting, and length calculation for Task C. All three of us contributed to project discussions and method refinement.

REFERENCES

[1] J. J. Kahle, A. V. Kulkarni, D. D. Limbrick, and B. C. Warf, "Hydrocephalus in Children," *The Lancet*, vol. 387, no. 10020, pp. 788–799, 2016.

[2] B. A. Hargreaves, P. W. Worters, K. B. Pauly, J. M. Pauly, K. M. Koch, and G. E. Gold, "Metal-Induced Artifacts in MRI," *American Journal of Roentgenology*, vol. 197, no. 3, pp. 547–555, 2011.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[5] M. J. Cardoso et al., "MONAI: An Open-Source Framework for Deep Learning in Healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[6] T.-C. Lee, R. L. Kashyap, and C.-N. Chu, "Building Skeleton Models via 3-D Medial Surface/Axis Thinning Algorithms," *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 6, pp. 462–478, 1994.

[7] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 29, 2015.

[8] N. J. Tustison et al., "N4ITK: Improved N3 Bias Correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.