UNCERTAINTY ESTIMATION IN THE SEGMENTATION OF BRAIN MAGNETIC RESONANCE ANGIOGRAMS

Master's Thesis

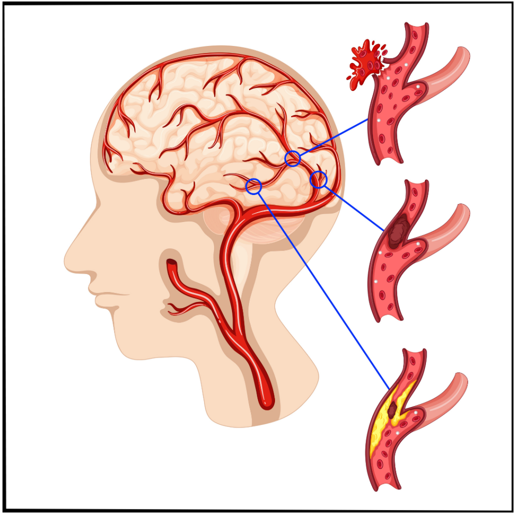
Omini Rathore
RWTH Aachen

January 15, 2025 | First Examiner: Prof. Dr. Abigail Morrison | Second Examiner: Prof. Dr. Michael Herty |
Supervisor: Elisabeth Pfaehler | Manager: Dr. Hanno Scharr | IAS-8

Part I: Motivation

CEREBROVASCULAR SEGMENTATION - IMPORTANCE

- Cerebrovascular diseases affect the brain's blood vessels
- Detection : analysing vasculature
- Cerebrovascular segmentation : segmenting brain vasculature from medical image scans
- Ususally done manually



Source: [1]

CEREBROVASCULAR SEGMENTATION - CHALLENGES

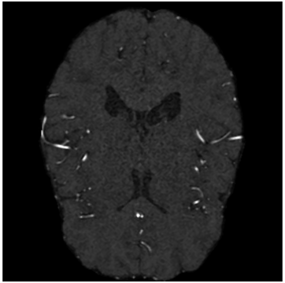
Manual segmentation

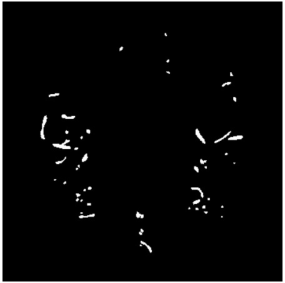
- prone to inter-observer variability
- labour-intensive
- subjective bias

Automatic segmentation

- traditional models perform subpar
- deep-learning based models are complex
- highly opaque, black-boxes

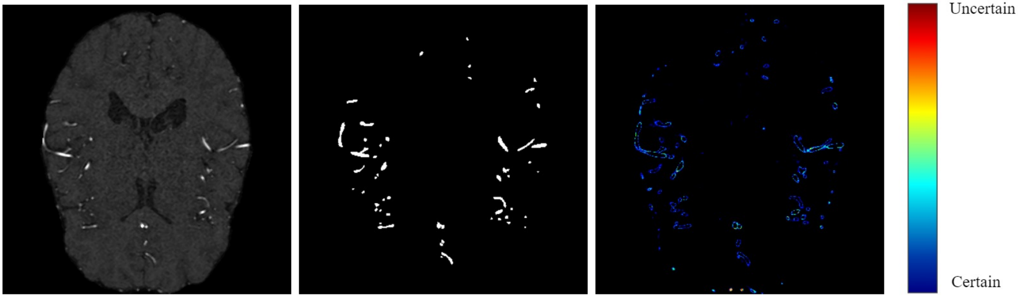
How can we make automatic segmentation more trustworthy and reliable?





UNCERTAINTY QUANTIFICATION

- Degree of confidence a model has in its predictions
- In case of image classification, pixel-wise uncertainty
- Makes models more reliable and trustworthy
- Research Gap: current deep-learning based cerebrovascular segmentation models do not incorporate uncertainty quantification -> Implement uncertainty estimation model for cerebrovascular segmentation



Part II: Background

BACKGROUND

Magnetic Resonance Angiogram (MRA)

- Process of acquiring MRAs
- Time-of-Flight MRAs (TOF-MRAs)
- Imaging parameters

Deep-learning Segmentation Model

- Medical Image Segmentation Model : U-Net
- U-Net hyperparameters and variations

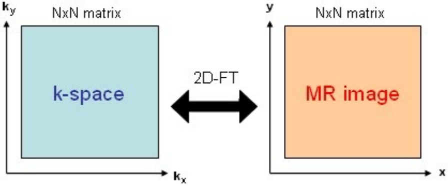
Uncertainty Estimation

- Types of uncertainties
- Uncertainty estimation approaches: Bayesian Approximation, Ensemble Methods
- Efficient Ensemble Model

MAGNETIC RESONANCE ANGIOGRAM (MRA)

- Magnetic Resonance Imaging (MRI) scanners use coils to detect electromagnetic energy in the form of a signal.
- Image characteristics are influenced by various parameters
- TR, TE, contrast agents
- MRA: Specialised application of MRI
- Used for imaging the body’s vascular system

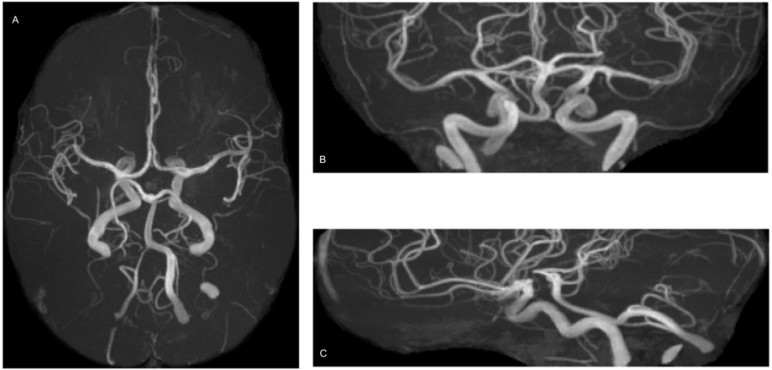
Source: [4]



Source: [5]

TOF-MRA

- Provides high-resolution images in less time compared to other imaging techniques
- Uses distinction between stationary tissues and non stationary flowing blood
- However, high variability in image quality



U-NET : MEDICAL IMAGE SEGMENTATION MODEL

Encoder

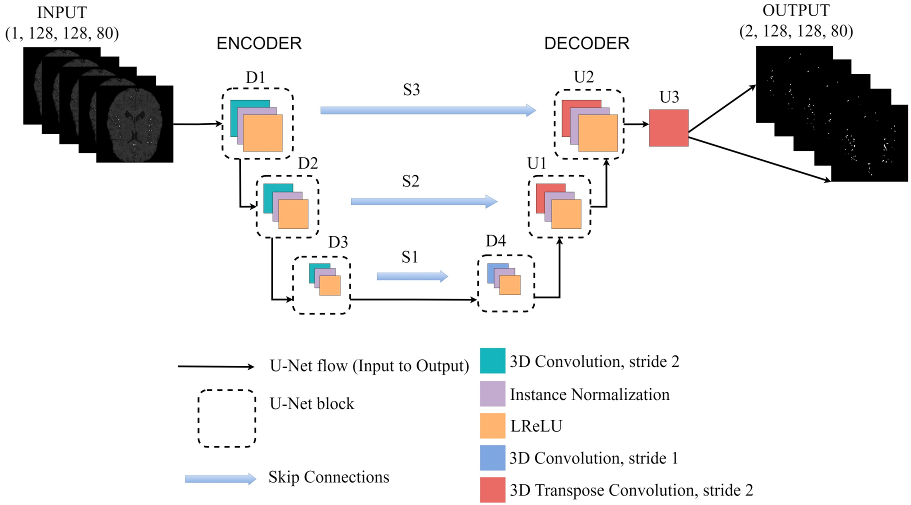
- down-samples the input image to extract features
- max-pooling or strided convolutions
- number of features increases with depth of the model to extract more complex high-level features

Decoder

- up-samples the down-sampled image
- reconstructs the image to original resolution using transposed convolution
- skip connections are concatenated with up-sampled image to ensure fine-details are not lost
- skip-connections bridge the encoder and decoder structure preserving spatial information

JÜLICH Forschungszentrum

U-NET ARCHITECTURE



U-NET ARCHITECTURE

Important hyperparameters

- **Number of feature maps**
- **Number of layers**

U-Net variations

- **Inception modules**
- **V-Net**
- **Attention gates**

Block	Operation	Output Shape	Kernel	Stride	Padding
D1	Conv3d	[32, 64, 64, 40]	3x3x3	2, 2, 2	1, 1, 1
	InstanceNorm3d	[32, 64, 64, 40]	–	–	–
	LeakyReLU	[32, 64, 64, 40]	–	–	–
D2	Conv3d	[64, 32, 32, 20]	3x3x3	2, 2, 2	1, 1, 1
	InstanceNorm3d	[64, 32, 32, 20]	–	–	–
	LeakyReLU	[64, 32, 32, 20]	–	–	–
D3	Conv3d	[128, 16, 16, 10]	3x3x3	2, 2, 2	1, 1, 1
	InstanceNorm3d	[128, 16, 16, 10]	–	–	–
	LeakyReLU	[128, 16, 16, 10]	–	–	–
D4	Conv3d	[256, 16, 16, 10]	3x3x3	1, 1, 1	1, 1, 1
	InstanceNorm3d	[256, 16, 16, 10]	–	–	–
	LeakyReLU	[256, 16, 16, 10]	–	–	–
S1	SkipConnection	[384, 16, 16, 10]	–	–	–
U1	ConvTranspose3d	[64, 32, 32, 20]	3x3x3	2, 2, 2	1, 1, 1
	InstanceNorm3d	[64, 32, 32, 20]	–	–	–
	LeakyReLU	[64, 32, 32, 20]	–	–	–
S2	SkipConnection	[128, 32, 32, 20]	–	–	–
U2	ConvTranspose3d	[32, 64, 64, 40]	3x3x3	2, 2, 2	1, 1, 1
	InstanceNorm3d	[32, 64, 64, 40]	–	–	–
	LeakyReLU	[32, 64, 64, 40]	–	–	–
S3	SkipConnection	[64, 64, 64, 40]	–	–	–
U3	ConvTranspose3d	[2, 128, 128, 80]	3x3x3	2, 2, 2	1, 1, 1

JÜLICH Forschungszentrum

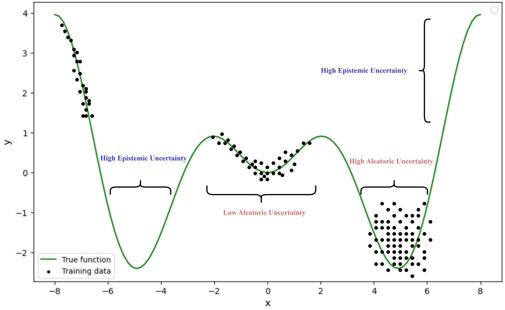
CATEGORIES OF UNCERTAINTY

Aleatoric uncertainty

- arises from inherent noise in the data
- can not be reduced with more data
- body movement during MR scanning
- short scan duration

Epistemic uncertainty

- arises from limitations in the model itself
- insufficient training data or uncertainty in the model parameters
- limited annotated data in biomedical tasks, out-of-distribution data



JÜLICH Forschungszentrum

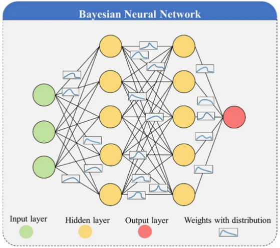
UNCERTAINTY ESTIMATION APPROACHES

Bayesian methods

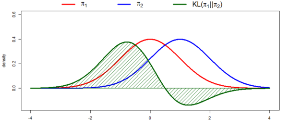
- **(i) Sampling-based methods :** based on Markov Chain Monte Carlo (MCMC) methods
- A Markov chain is constructed such that the samples are distributed following a desired distribution
- **(ii) Variational inference :** minimise the Kullback-Leibler (KL) divergence from the true posterior

$$P(\mathbf{w}|D) = \frac{P(D|\mathbf{w})P(\mathbf{w})}{P(D)}$$

$$\theta^* = \arg\min_{\theta} \mathrm{KL}\left[q(\mathbf{w}|\theta) \,||\, P(\mathbf{w}|D)\right]$$



Source: [2]

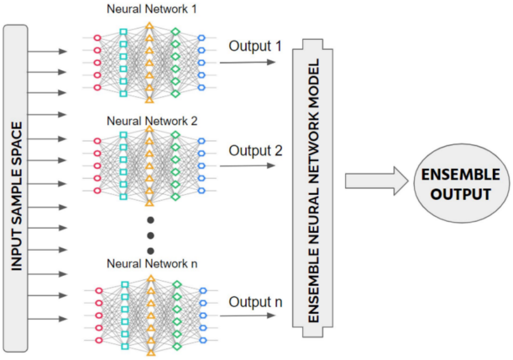


Source: [7]

UNCERTAINTY ESTIMATION APPROACHES

Ensemble methods

- **Deep Ensembles:** multiple independent models with varying hyper-parameter configurations are trained
- Provide uncertainty estimates by measuring the variance between the predictions of individual ensemble members
- Require substantial computational resources and memory



Source: [3]

EFFICIENT ENSEMBLE MODEL

Efficient Ensemble model by Lee et. al.

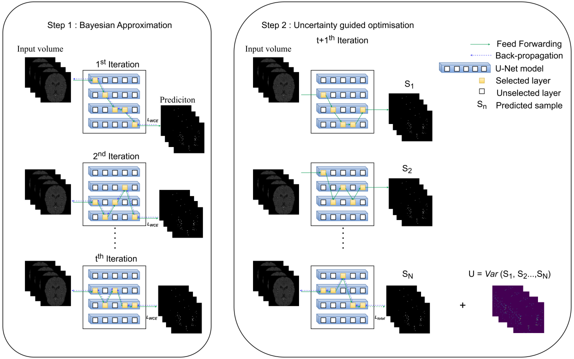
- Model set consists of M sub-models with L layers
- Step1:

$$L_{WCE} = -\alpha y_i \log(\hat{y}_i) - (1-\alpha)(1-y_i)\log(1-\hat{y}_i)$$

- Step 2:

$$L_U = -\sum_{i=1}^{N} ((U_i \otimes y_i)\log(U_i \otimes \hat{y}_i))$$

$$L_{\text{total}} = L_{WCE} + \lambda L_U$$

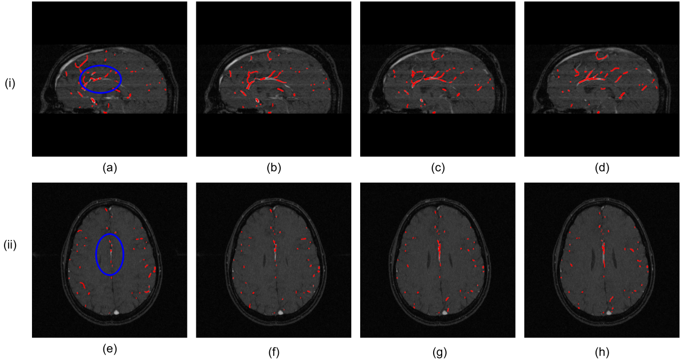


Part III: Experiments & Results

DATASET

TubeTK dataset

- 109 healthy subjects, 41 of these subjects with intracranial vasculature
- Discontinuities in vessel representations
- Other limitations - small dataset, only healthy subjects, all images from same scanner



DATASET

COSTA dataset

- Larger dataset
- More variability
- Precise annotations

Dataset	Subjects	MR Scanner	Image Dimensions	Voxel spacing (mm^3)
IXI-Guys [Lon]	60 - healthy	Philips Medical System Gyroscan Intera, 1.5T, TR/TE = 20/6.9 ms, flip angle = 25	512 x 512 x 100	0.47 x 0.47 x 0.80
IXI-HH [Lon]	60 - healthy	Philips Medical Systems Intera, 3.0T, TR/TE = 16.72/5.75 ms, flip angle = 16	512 x 512 x 100	0.47 x 0.47 x 0.80
IXI-IOP [Lon]	50 - healthy	General Electric (GE), 1.5T	1024 x 1024 x 92	0.26 x 0.26 x 0.80
ADAM [MIC20]	85 - diagnosed with aneurysms, 22 - healthy	3.0T	256 x 256 x 100 to 560 x 560 x 140	0.20 x 0.20 x 0.40 to 0.59 x 0.59 x 1.00
ICBM [icb]	25 - healthy females, 25 - healthy males, age 19 - 64 years	Siemens TRIO, 3.0T	288 x 320 x 200	0.62 x 0.62 x 0.62
LocH1 [iNIoMTE24]	2 - diagnosed with Moyamoya, 25 - healthy	Siemens HealthCare GmbH, 3.0T	406 x 512 x 186 to 500 x 640 x 121	0.34 x 0.34 x 0.80 to 0.39 x 0.39 x 0.90

BASELINE

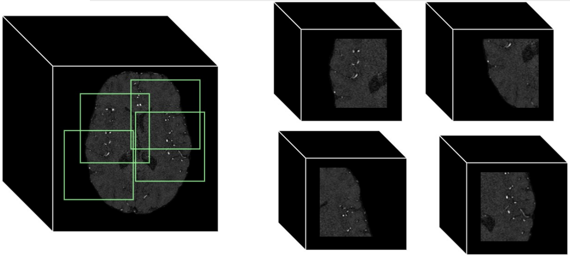
1. Pre-processing

- Images were already skull-stripped
- Bias correction
- Normalisation

2. Training

- Patch based training, patch size - 128 x 128 x 80
- Augmentations: random flipping, contrast adjustment and random motion

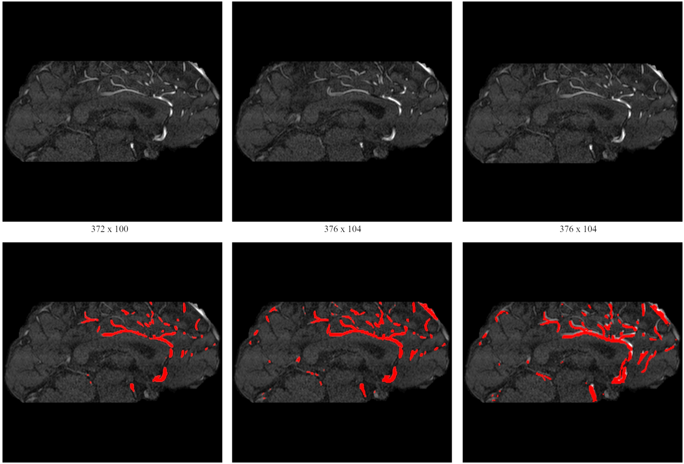
$$L_{WCE} = -\alpha y_i \log(\hat{y}_i) - (1-\alpha)(1-y_i)\log(1-\hat{y}_i)$$



BASELINE

3. Inference

- Full-sized images
- Padding
- Voxel-positions were preserved to avoid misalignment



(i) Input and Ground truth image slice

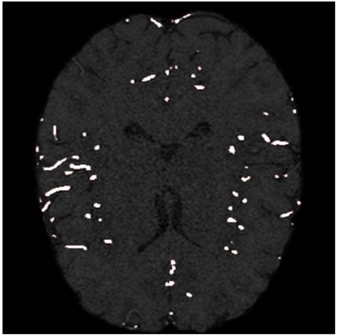
(ii) Output with voxel-position preserving padding & cropping
clDice = 0.79

(ii) Output with normal padding & cropping
clDice = 0.11

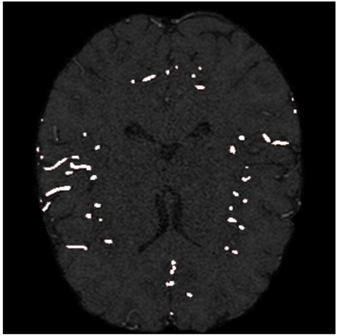
BASELINE

4. Post-processing

- Cropping
- Reduced over-segmentation



Model Output



Output after processing

BASELINE EVALUATION

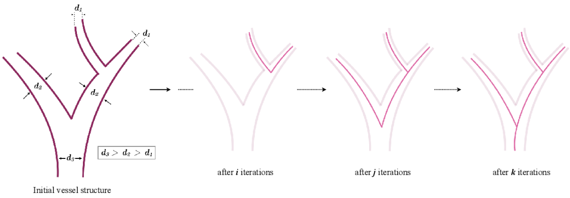
$$\textbf{Dice Score} = \frac{2 \cdot |\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction}| + |\text{Ground Truth}|}$$

$$T_{\text{prec}}(S_P, V_L) = \frac{|S_P \cap V_L|}{|S_P|}, \quad T_{\text{sens}}(S_L, V_P) = \frac{|S_L \cap V_P|}{|S_L|}$$

$$\textbf{clDice}(V_P, V_L) = 2 \times \frac{T_{\text{prec}}(S_P, V_L) \times T_{\text{sens}}(S_L, V_P)}{T_{\text{prec}}(S_P, V_L) + T_{\text{sens}}(S_L, V_P)}$$

$$\textbf{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\textbf{Sensitivity/ Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$



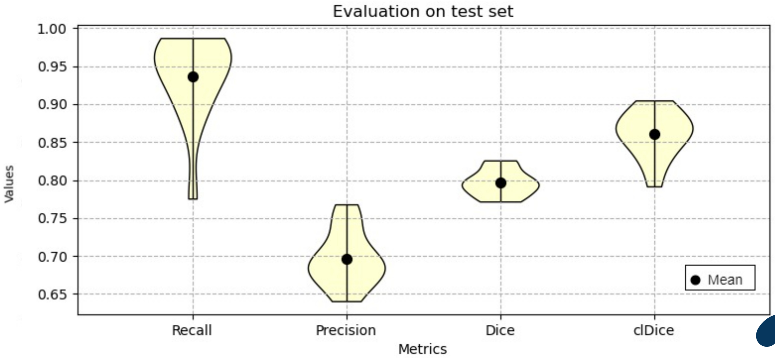
Source: [6]

- S_L, S_P : Skeletons
- V_L, V_P : Ground Truth, Prediction
- $T_{\text{prec}}(S_P, V_L)$: Topology Precision
- $T_{\text{sens}}(S_L, V_P)$: Topology Recall

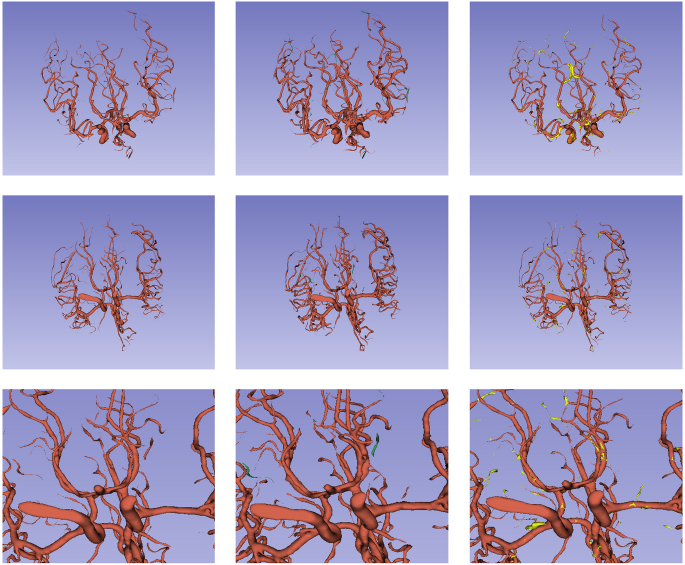
JÜLICH Forschungszentrum

BASELINE RESULTS

Sr. No.	Layer-wise feature maps	Recall	Precision	clDice	Dice Score % (Foreground)
1	(16, 32, 64, 128)	93.89	60.36	83.84	73.25
2	(32, 64, 128)	93.35	65.80	84.36	76.98
3	(32, 64, 128, 256)	93.61	**69.69**	**86.03**	**79.69**
4	(32, 64, 128, 256, 512)	**95.18**	55.67	80.66	70.09
5	(64, 128, 256, 512)	93.30	69.17	85.08	78.81



BASELINE RESULTS



Ground truth | Model output with FN | Model output with FP

EFFICIENT ENSEMBLE MODEL

Training

Step 1

- Sub-models were initialised with different weights
- Ensemble members were generated using random layer selection
- Model was trained for 200 epochs

Step 2

- Fine-tuning for additional 100 epochs
- Variance was calculated for T=20 samples

EFFICIENT ENSEMBLE RESULTS

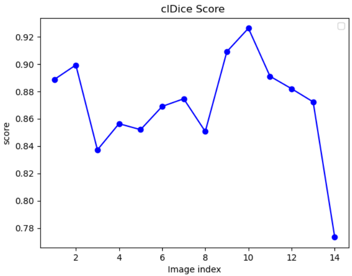
To identify optimal parameters, training was done for different combinations of sub-models and layers beginning with minimum number of sub-models

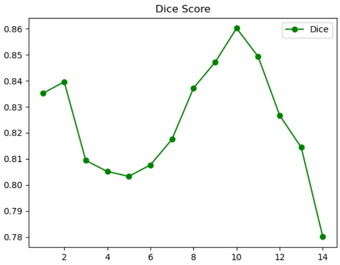
S No.	# of Sub-models (M)	# of Layers (L)	# of Ensemble members (M^L)	Layer-wise feature maps	# of Parameters	Training Time (hh:mm:ss)
1	2	7	128	(32,64, 128, 256)	3.8M	09:10:00
2	2	9	512	(32,64, 128, 256, 512)	15.8M	16:20:00
3	3	7	2187	(32,64, 128, 256)	5.8M	12:20:00
4	3	9	19683	(32,64, 128, 256, 512)	23.7M	19:40:07
5	4	7	16384	(32,64, 128, 256)	7.8M	15:30:00
6	4	9	262144	(32,64, 128, 256, 512)	31.6M	23:40:00

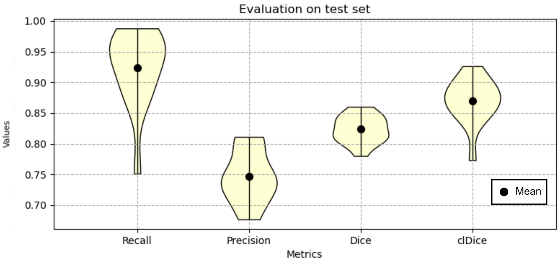
S No.	# of Sub-models (M)	# of Layers (L)	Recall	Precision	clDice	Dice Score % (Foreground)
1	2	7	92.71	67.53	85.41	77.87
2	2	9	91.67	73.98	85.53	81.64
3	3	7	92.40	**74.74**	**87.02**	**82.38**
4	3	9	92.52	71.83	85.36	80.63
5	4	7	92.12	74.07	86.92	81.85
6	4	9	91.76	73.00	85.88	81.05
7	1 (Baseline)	7	**93.61**	69.69	86.03	79.69

JÜLICH Forschungszentrum

EFFICIENT ENSEMBLE RESULTS

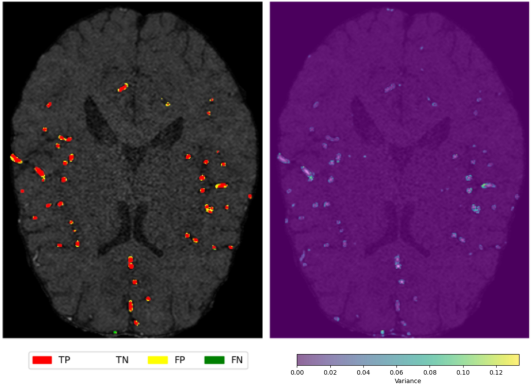




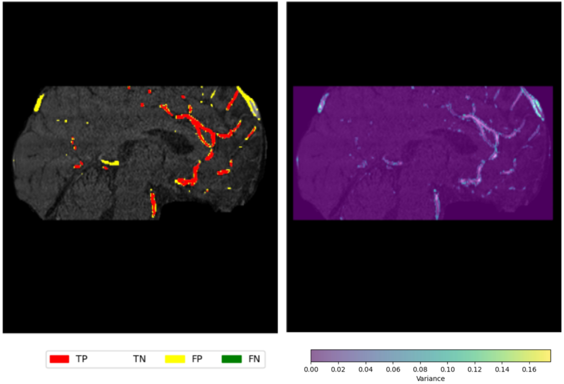


JÜLICH
Forschungszentrum

UNCERTAINTY ANALYSIS



(i) Axial view slice



(i) Sagittal view slice

Figure: Axial and Sagittal view of segmented output (left) with associated uncertainty map (right). The uncertainty map illustrates variance distributed across the spectrum, with purple indicating near-zero variance. As the variance increases, it transitions to green and yellow, representing the highest variance levels exceeding 0.1.

UNCERTAINTY ANALYSIS

Uncertainty in vessel boundaries

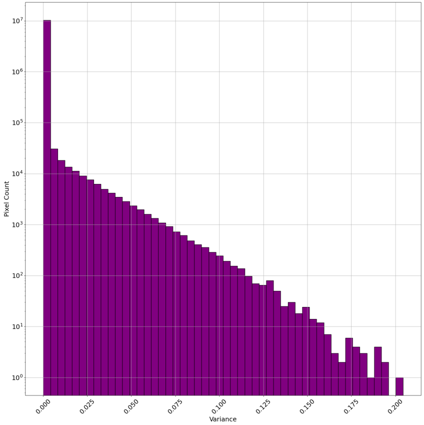
- Arises from the model's limitation in accurately segmenting edges
- Vasculature consists of vessels of varying thickness
- Spatial details may not be perfectly restored by the U-Net

Uncertainty in non-vessel structures with vessel-like intensities

- Higher uncertainty in the false predictions close to the skull
- Dataset consists of some slices with low contrast in these regions, confusing the model to interpret them as vessels
- Lack of training data containing similar regions

UNCERTAINTY-BASED FILTERING

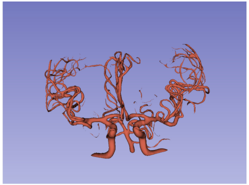
- Ideally, voxels with high uncertainty are more prone to errors
- Labels with uncertainty higher than a certain threshold were altered
- Multiple thresholds were used based on uncertainty distribution

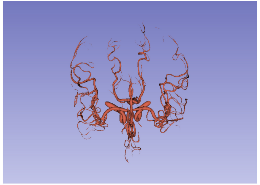


JÜLICH Forschungszentrum

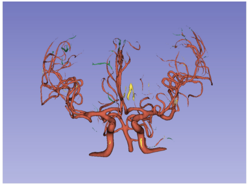
UNCERTAINTY-BASED FILTERING

Ground truth

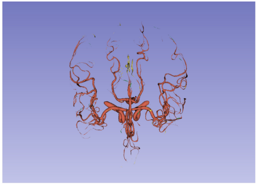




Threshold = 0.005

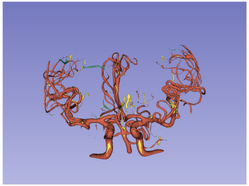


clDice = 92.45, Dice = 88.0

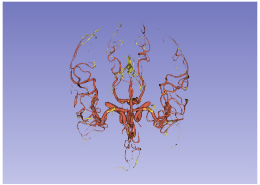


clDice = 89.55, Dice = 88.64

No threshold

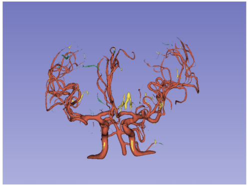


clDice = 90.91, Dice = 84.7

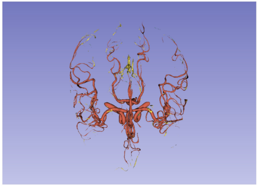


clDice = 85.20, Dice = 80.33

Threshold = 0.025



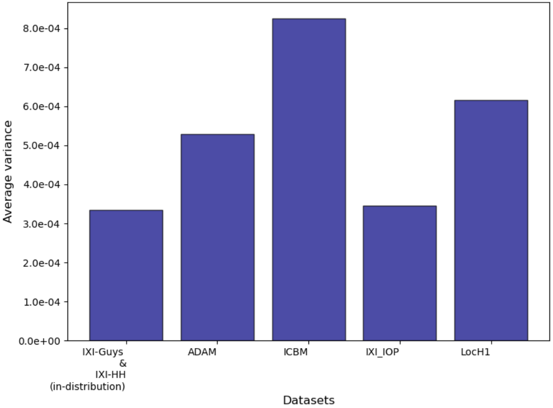
clDice = 91.57, Dice = 88.0



clDice = 86.43, Dice = 84.71

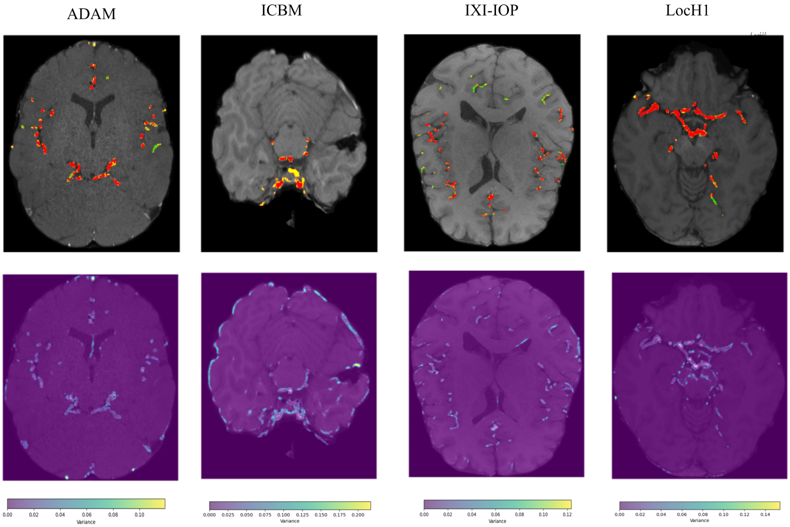
UNCERTAINTY IN OUT-OF-DISTRIBUTION DATA

- Model was tested on four OOD datasets
- Images from different scanners, healthy and unhealthy individuals
- Average uncertainty was higher in these datasets
- More false negatives as compared to in-distribution data



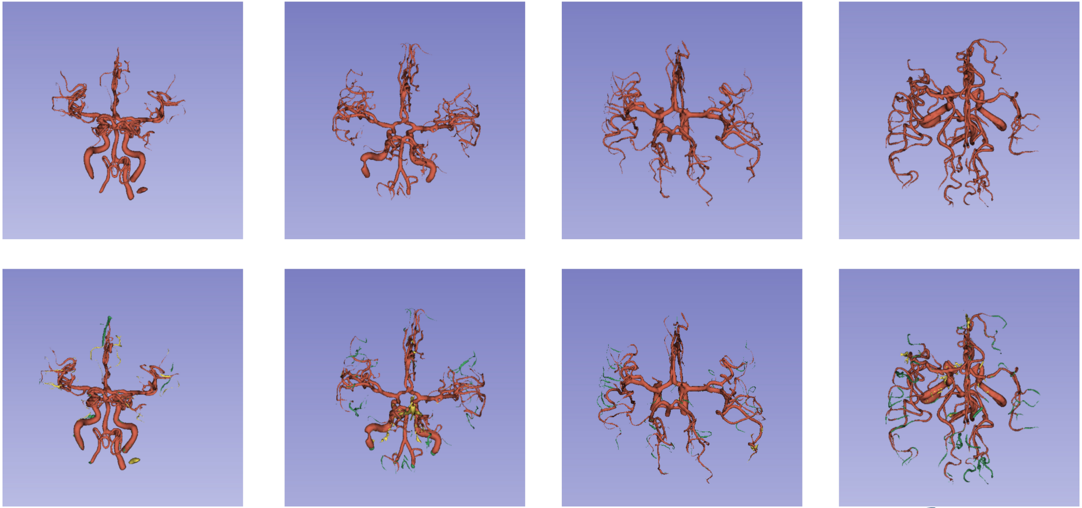
JÜLICH Forschungszentrum

UNCERTAINTY IN OUT-OF-DISTRIBUTION DATA



UNCERTAINTY IN OUT-OF-DISTRIBUTION DATA

ADAM | ICBM | IXI-IOP | LocH1



Part IV: Future Work

Incorporating Aleatoric Uncertainty

- Noise introduced by shorter scan durations, patient movement or scanner settings
- Aleatoric uncertainty can help explain these data-driven errors

Exploration of Uncertainty Quantification Techniques

- Alternative techniques such as Bayesian neural networks can be applied
- Combines both uncertainties in a single model

Enhancing Sub-Models for Improved Segmentation

- U-Net variations such as nnU-Net for training data with different voxel spacing
- More generalized sub-model

Optimizing Thresholding

- Image-Specific Threshold Optimization
- Dynamic Threshold

REFERENCES I

[1] Cerebrovascular Disease.
https://continentalhospitals.com/diseases/cerebrovascular-disease/.

[2] Jacek Fleszar. Bayesian Neural Networks - Capturing The Uncertainty Of The Real World.
https://medium.com/@fleszarjacek/bayesian-neural-networks-capturing-the-uncertainty-of-the-real-world-de8ce2dcf660.

[3] Juned Inamdar. Deep Learning Based Ensembled Models.
https://medium.com/@Juned_Inamdar/deep-learning-based-ensembled-models-1da08a8d0d99.

[4] Courtney Misher. MRI (Magnetic Resonance Imaging).
https://www.oncolink.org/cancer-treatment/procedures-diagnostic-tests/radiology-tests/mri-magnetic-resonance-imaging.

[5] Nuclear Magnetic Resonance at the Institute for Molecular Cardiology.
https://nmr.hhu.de/main/mtheorie_2_e.html. Accessed: 2025-01-12.

REFERENCES II

[6] Suprosanna Shit et al. “clDice - a Novel Topology-Preserving Loss Function for Tubular Structure Segmentation”. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), pp. 16555–16564. URL: https://api.semanticscholar.org/CorpusID:232428560.

[7] Duco Veen et al. “Using the Data Agreement Criterion to Rank Experts’ Beliefs”. In: Entropy 20 (Sept. 2017). DOI: 10.3390/e20080592.

JÜLICH Forschungszentrum

THANK YOU FOR YOUR ATTENTION.