Evaluating Pipeline Design Choices for Chronic Stroke Lesion Segmentation in T1-Weighted MRI

Mina A. Nessiem[1,2][0000−0003−3738−678X] and Björn W. Schuller[1,2][0000−0002−6478−8699]

[1] Technical University of Munich - University Hospital, Chair for Health Informatics, Munich, Germany, Ismaninger Str. 22, 81675 Munich, Germany
[2] Munich Center for Machine Learning (MCML), Oettingenstraße 67, 80538 Munich, Germany
{m.nessiem, schuller}@tum.de

Abstract. Pipelines for chronic ischaemic stroke lesion segmentation often accumulate normalisation, sampling, architectural, and post-processing components whose individual contributions remain unclear. We examine these choices through staged development on multicentre T1-weighted magnetic resonance imaging (MRI), yielding a 3D DynUNet pipeline with raw intensities, 128^3-voxel patches, spatial augmentation, and Dice–focal loss, without atlas registration or threshold calibration. On the ATLAS v2.1 development split, leading DynUNet configurations attain Dice scores of 0.625–0.633, compared with 0.601 for an equivalently evaluated nnU-Net baseline. However, on ATLAS v3.0 three-fold cross-validation, nnU-Net attains a higher mean Dice than DynUNet (0.628 vs 0.610), reversing the development ranking; DynUNet records a slightly lower mean 95th-percentile Hausdorff distance. The three fold-specific DynUNets form the ISLES 2026 submission ensemble, whose hidden-test performance remains unavailable. Normalisation, aggressive lesion oversampling, and increased capacity provide no consistent development benefit. These findings show that promising pipeline choices may not transfer across dataset expansions, underscoring the importance of strong baselines and validation beyond a single development split.

Keywords: Ischaemic Stroke Lesions · 3D Medical Image Segmentation · T1-weighted MRI · Deep Learning · Medical Image Preprocessing.

1 Introduction

Chronic ischaemic stroke lesions vary in size, morphology, location, and appearance. These factors complicate their delineation in T1-weighted Magnetic Resonance Imaging (MRI) scans [2]. Large, multicentre resources and 3D convolutional networks promise segmentation at scale [8, 14], yet varying acquisition conditions and heterogeneous scanner characteristics make multicentre segmentation performance consistently difficult [4].

Through staged experiments spanning architecture, T1 representation, augmentation, training objective, spatial context, sampling, model capacity, and

Table 1. Dataset splits constructed as part of the experimental design

Split name	Full, n	**T**rain, n (%)	**V**alid, n (%)	Sites (T/V)
ATLAS v2.1 85/15	955	812 (85.03)	143 (14.97)	33/33
ATLAS v3.0 3-fold CV[†]	1 453	968–969 (66.62–66.69)	484–485 (33.31–33.38)	53–55/52–54

[†] Statistics per fold

thresholding, we investigate which factors lead to improved multicentre segmentation performance. We compare the performance of a baseline nnU-Net with that of our selected 3D DynUNet pipeline, identifying which forms of added complexity improve performance—and which do not.

2 Methods

Below, we describe the dataset and its preprocessing, the model architecture and training procedure, the experimental programme, and the evaluation protocol.

2.1 Dataset Contents and Splits

The ISLES'26 competition dataset comprises cases sampled across multiple acquisition centres. Each case consists of an unprocessed, single-modality, 3D T1-weighted MRI volume; its corresponding ischaemic stroke lesion segmentation mask; and an associated tabular metadata file.

The ISLES'26 competition dataset was released in two successive batches: ATLAS v2.1 (derived from ATLAS v2.0 [8]) and ATLAS v3.0 (which also included the 169-case Stroke Outcome Optimization Project (SOOP) dataset [1]). ATLAS v2.1 comprised 955 cases from 33 sites, with a median of 24 cases per site (interquartile range (IQR): 14–35), whereas ATLAS v3.0 comprised 1 453 cases from 55 sites, with a median of 16 cases per site (IQR: 9.5–35).

To construct the splits as shown in Table 1, we employ a site-aware, metadata-stratified candidate search that allocates cases proportionally among the subsets. We categorically bin the number of days post stroke to seven bins and combine these bins, along with the chronicity categorical variable and the acquisition centre to split the dataset into variable-stratified splits. For each candidate split, we sum the absolute differences between the proportions of the resulting strata in each validation subset and the corresponding training subset and select the candidate that minimises this sum. We apply this procedure to both the 85 %/15 % holdout split for ATLAS v2.1 and the three-fold cross-validation (CV) split for ATLAS v3.0.

2.2 Preprocessing and Input Configurations

We implement our data-processing pipeline using the MONAI toolkit [3]. We first reorient all input volumes to the right–anterior–superior (RAS) coordinate

convention and resample them to an isotropic voxel spacing of 1 mm, using linear interpolation for the image data and nearest-neighbour interpolation for the segmentation masks. During training, we extract 128^3-voxel patches using a 1:1 foreground-to-background centre-sampling ratio. We apply synchronised random flips across all three spatial axes with a probability of 0.5 and rotations sampled from $[-360°, 360°]$ about each spatial axis with a probability of 0.3. We apply no intensity normalisation, atlas registration, brain cropping, or skull stripping, thereby retaining subject-specific anatomical variation. We do not employ the accompanying tabular metadata during training.

2.3 Segmentation Architecture and Training

We select the MONAI 3D DynUNet architecture [3] as the segmentation backbone. It comprises four resolution stages with feature widths $(32, 64, 128, 256)$ and $3 \times 3 \times 3$ convolutional kernels throughout. The encoder follows the isotropic stride schedule $(1, 2, 2, 2)$; three decoder stages perform upsampling through $2 \times 2 \times 2$ transposed convolutions. Residual convolutional blocks span the network, with skip connections linking corresponding encoder and decoder blocks. Given a batch size of three, instance normalisation avoids reliance on batch statistics. LeakyReLU retains gradients for negative activations, while spatial augmentation provides regularisation without dropout.

Model optimisation minimises the `DiceFocalLoss` objective [12] applied to the binary output logits. For predicted logits z_i and binary reference labels y_i, let $p_i = \sigma(z_i)$ denote the predicted foreground probability at voxel i. The Dice component [13] is

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \epsilon}, \qquad \epsilon = 10^{-5}, \tag{1}$$

where N denotes the number of voxels in one training patch. For the focal component [9], let $p_{t,i} = y_i p_i + (1 - y_i)(1 - p_i)$ denote the probability assigned to the reference class at voxel i. The focal loss is

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N} \sum_{i=1}^{N} (1 - p_{t,i})^{\gamma} \log(p_{t,i}), \qquad \gamma = 2. \tag{2}$$

The combined objective is

$$\mathcal{L}_{\text{DF}} = \lambda_{\text{D}} \mathcal{L}_{\text{Dice}} + \lambda_{\text{F}} \mathcal{L}_{\text{focal}}, \qquad \lambda_{\text{D}} = 1.0, \quad \lambda_{\text{F}} = 0.5. \tag{3}$$

Each experiment runs on a single NVIDIA A100 GPU with 40 GB of memory in an NVIDIA DGX system. Training for the selected configuration proceeds with a batch size of three under `bfloat16` mixed precision. Each model trains for 100 000 optimisation steps with the AdamW optimiser [11], a peak learning rate of 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and zero weight decay. The learning rate increases linearly from zero to 2×10^{-4} during the first 10 % of training and subsequently follows cosine annealing to zero [10]. We clip the global gradient norm to a maximum of 1, evaluate the complete validation set every 5 000 steps, and retain the checkpoint with the highest mean subject-wise 3D Dice score.

nnU-Net baseline We train an auto-configuring nnU-Net v2 model as the comparison baseline [6]. We provide nnU-Net with the same single-channel, raw-intensity T1-weighted images and dataset split as the DynUNet pipeline; nnU-Net then performs its automatically planned preprocessing. The resulting 3D full-resolution plan selects a six-stage PlainConvUNet with feature widths $(32, 64, 128, 256, 320, 320)$ and an isotropic target spacing of 1 mm. All convolutions have $3 \times 3 \times 3$ kernels. The encoder follows the isotropic stride schedule $(1, 2, 2, 2, 2, 2)$, with two convolutional blocks per encoder and decoder stage. We evaluate the resulting binary segmentations with the same subject-wise 3D metrics as the DynUNet models.

2.4 Experiments

We develop the final preprocessing and training pipeline through a staged experimental programme on the ATLAS v2.1 85/15 split described in Section 2.1. Section 3.1 and Table 2 report the corresponding results.

Each experimental stage addresses a distinct design question; its best-performing setting advances to subsequent stages. Unless stated otherwise, we train each configuration once for 100 000 optimisation steps with random seed 42 and characterise it by its highest validation-set mean subject-wise 3D Dice score. These comparisons therefore do not estimate variability across random seeds.

We first compare DynUNet and Swin UNETR [5] as segmentation backbones across T1 intensity representations and augmentation policies. The DynUNet configuration with raw T1 intensities and spatial augmentation alone advances to subsequent stages. These stages compare training objectives, binarisation-threshold calibration, the balance between spatial context and lesion-aware crop oversampling, and DynUNet capacity. From these experiments, we adopt the `DiceFocalLoss` described in Section 2.3, retain a binarisation threshold of 0.5, and select 128^3-voxel patches with balanced centre sampling and a four-stage DynUNet with feature widths $(32, 64, 128, 256)$. We fix this configuration across the ATLAS v3.0 three-fold CV models without split-specific tuning. The ATLAS v2.1 validation subset therefore serves as a development set rather than an independent test set.

2.5 Inference and Evaluation

For development-time validation, we transform each image and reference mask to model space with RAS orientation and 1 mm isotropic voxel spacing. MONAI sliding-window inference produces whole-volume predictions from 128^3-voxel windows with 50 % overlap and Gaussian-weighted merging. A sigmoid activation converts the merged logits to foreground probabilities, which we binarise at a threshold of 0.5. We calculate the 3D Dice score, surface Dice at a 1 mm tolerance, and the 95th-percentile Hausdorff distance (HD95) for each subject in model space.

For ATLAS v3.0, we train three fold-specific models, each on two folds while reserving the remaining fold for validation. For the competition submission, we

Table 2. ATLAS v2.1 staged development on the 15 % validation subset. Values are means across subjects from the best checkpoint of one training run per configuration. Surface Dice is calculated with a 1.0 mm tolerance, and HD95 denotes the 95th-percentile Hausdorff distance in millimetres. Bold marks the best value per metric within each stage; inherited settings may differ between stages.

Design factor	Configuration	Dice ↑	Surface Dice ↑	HD95 (mm) ↓
Baseline	nnU-Net (3D full resolution)	**0.6012**	**0.5432**	**26.58**
Architecture	SwinUNETR	0.5612	0.5122	62.83
	DynUNet	**0.6075**	**0.5742**	**36.25**
T1 processing	Nonzero z-score T1	0.3795	0.3058	120.89
	Percentile-normalised T1	0.5183	0.4611	72.05
	Percentile-clipped z-score T1	0.3567	0.2587	110.63
	Raw-intensity T1	**0.6050**	**0.5686**	**47.56**
Augmentation	Spatial + intensity scaling	0.6134	0.5722	48.14
	Spatial + Gaussian blur	0.6190	0.5836	35.95
	Spatial only	**0.6234**	**0.5969**	**20.68**
Objective	Dice + BCE	0.6132	0.5727	40.91
	Tversky + BCE	0.6073	0.5708	30.50
	Dice + separate focal term	0.6070	0.5686	36.63
	Generalized Dice + BCE	0.5696	0.5228	36.40
	Dice + BCE + HD-DT	0.5188	0.4519	61.63
	Dice–focal	**0.6253**	**0.5941**	**20.93**
Crop size	32^3	0.4942	0.4096	60.56
	64^3	0.5814	0.5271	49.38
	96^3	0.5507	0.4979	73.47
	128^3	**0.6330**	**0.5949**	**23.31**
Positive sampling	1:1	**0.6330**	**0.5949**	**23.31**
	3:1	0.6241	0.5902	26.55
	5:1	0.6242	0.5931	24.18
DynUNet topology	5 stages; filters 24–320; 3^3 kernels	0.6060	0.5626	30.00
	5 stages; filters 32–320; 3^3 kernels	0.6001	0.5556	38.76
	4 stages; filters 24–192; 3^3 kernels	**0.6253**	0.5880	**25.70**
	4 stages; filters 32–256; mixed kernels	0.6246	0.5884	27.38
	4 stages; filters 32–256; 3^3 kernels	0.6212	0.5875	26.80
	4 stages; filters 48–384; 3^3 kernels	0.6205	**0.5951**	30.56

average the three foreground-probability maps and binarise the mean at a threshold of 0.5.

Competition masks must match the original image geometry. Before preprocessing, we record each image's dimensions, voxel spacing, orientation, and affine matrix. After inference and binarisation, we apply the inverse spatial transforms with nearest-neighbour interpolation to restore each mask to the input geometry.

3 Results

3.1 ATLAS v2.1 Staged Development Experiments

Table 2 shows the largest Dice differences across T1 representation and crop size: raw intensities record higher scores than the tested normalisations, while 128^3-voxel patches lead the patch-size sweep.

Spatial augmentation alone and `DiceFocalLoss` lead their respective stages, although the augmentation margins remain small under single-run evaluation. Stronger lesion-centred sampling shows no improvement, while smaller DynUNets

Table 3. ATLAS v3.0 three-fold cross-validation results across all 1 453 cases. Each row reports mean subject-wise scores from the best checkpoint of one training run on the indicated held-out fold. The final rows give unweighted means across folds. Surface Dice is calculated with a 1 mm tolerance, and HD95 denotes the 95th-percentile Hausdorff distance in millimetres. Bold marks the better value within each fold or mean pair.

Fold	Cases	Model	Dice ↑	Surface Dice ↑	HD95 (mm) ↓
1	485	nnU-Net	**0.635**	**0.578**	**32.4**
1	485	DynUNet	0.608	0.548	37.6
2	484	nnU-Net	**0.632**	**0.579**	32.9
2	484	DynUNet	0.607	0.563	**32.0**
3	484	nnU-Net	**0.617**	0.565	34.9
3	484	DynUNet	0.614	**0.573**	**29.4**
Mean	–	nnU-Net	**0.628**	**0.574**	33.4
Mean	–	DynUNet	0.610	0.561	**33.0**

provide negligible improvement. Because batch size varies with patch size, the patch comparison does not isolate spatial context from batch-size effects.

3.2 ATLAS v3.0 Three-Fold Cross-Validation

Table 3 reverses the ATLAS v2.1 ranking: nnU-Net records higher Dice in every fold and higher mean surface Dice (0.628 vs 0.610 Dice; 0.574 vs 0.561 surface Dice).

DynUNet records a 0.4 mm lower mean HD95 and leads that metric in two folds. The selected DynUNet therefore does not retain its development-set Dice advantage on the expanded cohort, illustrating the limitations of selection on a single holdout.

The three fold-specific DynUNets form the competition ensemble; hidden-test reference labels remain unavailable.

4 Discussion and Conclusion

This study examines how practical pipeline choices affect chronic ischaemic stroke lesion segmentation in heterogeneous, multicentre T1-weighted MRI. Across the staged ATLAS v2.1 experiments, the largest observed Dice ranges occur across T1 representations, patch sizes, and training objectives. Raw-intensity T1 inputs, spatial augmentation alone, `DiceFocalLoss`, 128^3-voxel patches, balanced centre sampling, and a four-stage DynUNet define the configuration carried to ATLAS v3.0. Added intensity augmentation, stronger lesion oversampling, and greater DynUNet capacity do not improve the observed Dice. Because each stage inherits earlier decisions, these findings describe a practical development path rather than isolated causal effects.

The ATLAS v3.0 results temper the development findings: nnU-Net records higher Dice in every held-out fold and higher mean surface Dice, whereas DynUNet records slightly lower mean HD95. The model ranking observed on ATLAS v2.1 therefore does not transfer to the expanded cohort. This reversal supports recent calls for rigorous comparisons against fully configured U-Net base-

lines [6, 7] and cautions against treating performance on one multicentre split as evidence of transfer to a related dataset with different composition [4].

Limitations Single runs leave seed variability unknown. Sequential selection prevents causal attribution, and patch size is confounded with batch size. Validation data guide checkpoint selection, while centres span training and validation subsets; thus, the results do not establish unseen-centre performance. Without competition test labels, ensemble performance also remains unknown.

Model rankings from one development split may not survive dataset expansion. Future work should quantify uncertainty across seeds and subjects, test centre-held-out cohorts, and evaluate the ensemble on hidden competition data. Robust validation and strong baselines remain as important as architectural choice.

References

1. Absher, J., Goncher, S., Newman-Norlund, R., et al.: The stroke outcome optimization project: Acute ischemic strokes from a comprehensive stroke center. Scientific Data **11**(1), 839 (2024). https://doi.org/10.1038/s41597-024-03667-5
2. Ahmed, R., Al Shehhi, A., Hassan, B., Werghi, N., Seghier, M.L.: An appraisal of the performance of AI tools for chronic stroke lesion segmentation. Computers in Biology and Medicine **164**, 107302 (2023). https://doi.org/10.1016/j.compbiomed.2023.107302
3. Cardoso, M.J., Li, W., Brown, R., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022). https://doi.org/10.48550/arXiv.2211.02701
4. Guan, H., Liu, M.: Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2022). https://doi.org/10.1109/TBME.2021.3117407
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Lecture Notes in Computer Science, vol. 12962, pp. 272–284. Springer, Virtual Event (2022). https://doi.org/10.1007/978-3-031-08999-2_22
6. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
7. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15009, pp. 488–498. Springer Nature Switzerland, Marrakesh, Morocco (2024). https://doi.org/10.1007/978-3-031-72114-4_47
8. Liew, S.L., Lo, B.P., Donnelly, M.R., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific Data **9**(1), 320 (2022). https://doi.org/10.1038/s41597-022-01401-7

9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007. IEEE (2017). https://doi.org/10.1109/ICCV.2017.324
10. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Skq89Scxx
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
12. Ma, J., Chen, J., Ng, M., et al.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021). https://doi.org/10.1016/j.media.2021.102035
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016). https://doi.org/10.1109/3DV.2016.79
14. Tomita, N., Jiang, S., Maeder, M.E., Hassanpour, S.: Automatic post-stroke lesion segmentation on MR images using 3D residual convolutional neural network. NeuroImage: Clinical **27**, 102276 (2020). https://doi.org/10.1016/j.nicl.2020.102276