BeSt-LeS: Benchmarking Stroke Lesion Segmentation using Deep Supervision

Prantik Deb[1], Lalith Bharadwaj Baru[1], Kamalaker Dadi[1], and Bapi Raju S[1]

[1]Brain Cognitive Computation Lab, Cognitive Science
IIIT, Hyderabad-32
prantik.deb@ihub-data.iiit.ac.in
lalith.baru@research.iiit.ac.in

Abstract. Brain stroke has become a significant burden on global health and thus we need remedies and prevention strategies to overcome this challenge. For this, the immediate identification of stroke and risk stratification is the primary task for clinicians. To aid expert clinicians, automated segmentation models are crucial. In this work, we consider the publicly available dataset ATLAS $v2.0$ to benchmark various end-to-end supervised U-Net style models. Specifically, we have benchmarked models on both 2D and 3D brain images and evaluated them using standard metrics. We have achieved the highest Dice score of 0.583 on the 2D transformer-based model and 0.504 on the 3D residual U-Net respectively. We have conducted the Wilcoxon test for 3D models to correlate the relationship between predicted and actual stroke volume. For reproducibility, the code and model weights are made publicly available.

Keywords: Stroke Lesion Segmentation · T1 Weighted MRI · Deep Supervision · ATLAS $v2.0$ · Deep Learning

1 Introduction

Brain stroke has become a significant burden on global health with increasing prevalence in low- and middle-income countries. Therefore there is an urgent need for targeted prevention strategies and improved healthcare infrastructure to address this growing public health challenge [20], [22]. Given brain Magnetic Resonance (MR) images on stroke populations, localizing and detecting the lesions is crucial for clinicians. However, the automation of the localization process has achieved significant reach [5] with novel machine-learning models that can aid clinicians. To fuel these models, we need datasets that could automatically segment to the level of expertise and doing so, it could ease the clinician's task.

In this context, ATLAS (Anatomical Tracings of Lesions After Stroke) v1.2 dataset made progress by creating 304 T1-weighted MRI samples collected from 11 cohorts. This ATLAS v1.2 [24] was released in 2018 and of which, 229 standardized subjects were available with T1-weighted MRI image and its corresponding lesion mask. Later in 2022, ATLAS v2.0 [25] was released and it has 1217 T1-weighted MRI samples collected from 44 cohorts of which 655 samples

Table 1. The table summarizes the glimpse of 2D and 3D U-Net variants whether they were implemented on the ATLAS v1.2 and v2.0 T1 MR images denoted as Yes or No. The implementation of 2D models on ATLAS v2.0 is sparse.

2D Models	**ATLAS v1.2**	**ATLAS v2.0**
U-Net	Yes [42] [30] [45][43] [33]	No
Residual U-Net	Yes [42] [30] [33]	No
Attention U-Net	Yes [33]	No
Transformer Based	Yes [40]	No
3D Models	**ATLAS v1.2**	**ATLAS v2.0**
U-Net	Yes [45] [43] [28]	Yes [36] [17]
Residual U-Net	Yes [34]	Yes [17]
Attention U-Net	No	No
Transformer Based	No	No

mask were availed to the public. In specific, there was an extra margin of 426 samples from version 1.2 to 2.0.

There were numerous impressive models that performed well on ATLAS v1.2. Whereas, ATLAS v2.0 was released recently and therefore, there is quite less progress. Therefore, we outpace and benchmark ATLAS v2.0 on various standard U-Net style architectures for both 2D and 3D brain images.

In Table 1, ATLAS v1.2 was applied using distinct U-Net architectures for 2D modality but, there isn't any model to date that has been implemented on ATLAS v2.0. Similarly, in the case of 3D modality, Table 1 illustrates there are few implementations for ATLAS v2.0 using nnU-Net as their underlying framework [18]. Thus, we contribute by analyzing the ATLAS v2.0 dataset for both 2D and 3D modality. A brief description of each of the models is detailed in the supplementary material.

2 Contributions of this work

1. To the best of our knowledge, ours is the first attempt to benchmark the standard segmentation models *i.e.* both convolution and transformers-based architectures on the ATLAS v2.0 dataset.
2. We have also conducted experiments for both 2D and 3D-based models. We report our highest dice score of 0.58 on the 2D transformer-based model. Also, we have achieved a 0.504 dice score on the 3D residual U-Net.
3. Finally, we conduct the Wilcoxon test on 3D models and compare the relationship between predicted and actual stroke volume.

3 Data and Models

This section briefly discusses the dataset and methods considered for analysis. The organization is as follows: First, we introduce the dataset and the training strategy implied. Next, we detail the significance of various U-Net style models.

3.1 Dataset

The ATLAS (Anatomical Tracings of Lesions After Stroke) data consists of T1-weighted MRI images of subjects having lesions due to stroke. This data has two versions, ATLAS v1.2 [24], and ATLAS v2.0 [25], respectively. For our analysis, we solely conduct our experiments on ATLAS v2.0 dataset [25], which is publicly available[1]. The samples in this data (ATLAS v2.0) were collected from 44 diverse cohorts with a total sample size of 1271. From these 1271 samples, only 655 samples consist of image-to-mask pairs dedicated to training the models. Another 300 samples are treated as hidden-set and do not reveal the masks of T1-weighted MRI images[2].

Table 2. The below table describes the train validation and test proportions divided for training supervised 2D U-Net style architectures. The number of samples below represents the number of subjects. The slices are unevenly divided based on the volume of the T1-weighted MRI and the axial plane is considered while cropping each slice.

Split	%	Samples	Slices
Train	60	393	15394
Validation	20	131	4666
Test	20	131	5452

Table 3. In the below tabular data, the train validation and test proportions are divided for training supervised 3D U-Net style architectures. The number of samples below represents the number of subjects.

Split	%	Samples
Train	60	393
Validation	20	131
Test	20	131

While analyzing ATLAS v2.0 we conduct experiments both on 2D and 3D modality. As masks for the original test set are inaccessible we divide the training samples (655) into train validation and test proportions. For 3D modality, the data can be directly fed to the model. But for 2D modality, each subject's 3D T1-weighted MRI images are to be cropped into slices along the z-axis (axial) and then given as input to 2D U-Net style architectures. For both the modalities the Z-score Normalization are performed as a pre-processing step [2]. The information regarding train, test, and validation sets are elucidated in Table 2 and 3. For the 2D dataset, the discrepancy between the validation and the test is due to a rejection of 0.1% lesions in the given 2D slice. Specifically, wherever the slices and their respective mask pairs were rejected were not included in the study. Now, this data is to be processed using various segmentation architectures to segment the lesion in the T1-weighted MRI images in the protocol mentioned in Table 2 and 3.

[1] The dataset can be found at: link

[2] Additional information such as lesion numbers, cortical location, and severity of stroke for each subject can be found in the original paper [25]

3.2 U-Net Style Architectures

The delineation of a specific organ or certain tissue site or cell nuclei from given medical images is one of the crucial tasks in medical image analysis. Various deterministic algorithms were developed to automate this process, some of which are random walks [15] and SLIC [1]. But later with the aid of deep learning, more sophisticated and *learnable* methods were developed [12]. Later, Ronneberger *et al.* [32] proposed U-Net architecture from which the field of medical image segmentation caught its attention as U-Net was fast, modulable, and robust. Later there were many methods that were crafted using the U-Net style as the underlying framework. Thus, we study and benchmark some of the fundamental U-Net style architectures that achieved significant results in the field of medical image segmentation.

U-Net: This was the first deep learning architecture that superseded the existing models with large margins both computationally and performance-wise. This architecture, thus, was made as a baseline for many medical segmentation tasks [32]. Soon after the U-Net, the architecture was modified to learn volumetric data using 3D convolutions [11] with sparse annotations. The architecture is very similar to that of 2D U-Net except that, 3D can process volumetric information using 3D convolutions and it is depicted in the last row of Fig. 1.

The architecture is quite intuitive, as features are downsampled using pooling layers to a latent space or base and later upsampled to reconstruct the desired mask image. This latent space is perceived to preserve the crucial features, while the skip-connections (these are denoted as $(\cdots)$) aid the reconstruction by guiding to map of the structural information. In the Fig. 1, you can see each blue circle represents a series of convolution layers, and the red arrow ($\rightarrow$) indicates that an image of size $h \times w$ is downsampled (pooled) to $\frac{h}{2} \times \frac{w}{2}$. The downsampling operation is quadruply performed to get to the latent space. Now, from this latent space, the acquired features are upsampled quadruply (indicated with a green arrow ($\rightarrow$)) to produce the desired segmentation.

Residual U-Net: He *et al.*[16] proved that residual connections tend to provide refined representations for downstream tasks with less computation and better performance. In this regard, Zhang *et al.* [44] proposed a U-Net architecture with residual connections for extracting road patterns from aerial imagery. This later was implied in the domain of medical images by Alom *et al.*[3] with additional memory components. Later, the residual connections were implied between a series of 3D convolutions to produce volumetric segmentation [41], [7],[19]. The architecture is very similar to that of 2D residual U-Net except that, 3D can produce volumetric masks from 3D medical data, and the pictorial interpretations are illustrated in the last row of Figure 1.

The current Residual U-Net style was inspired by Zhang *et al.*[44]. In U-Net, before downsampling at each step, there are a series of convolutions with residual connections, which are represented by an orange circle. The rest, upsampling ($\rightarrow$) and downsampling ($\rightarrow$) operations, are similar to that of traditional U-Net.

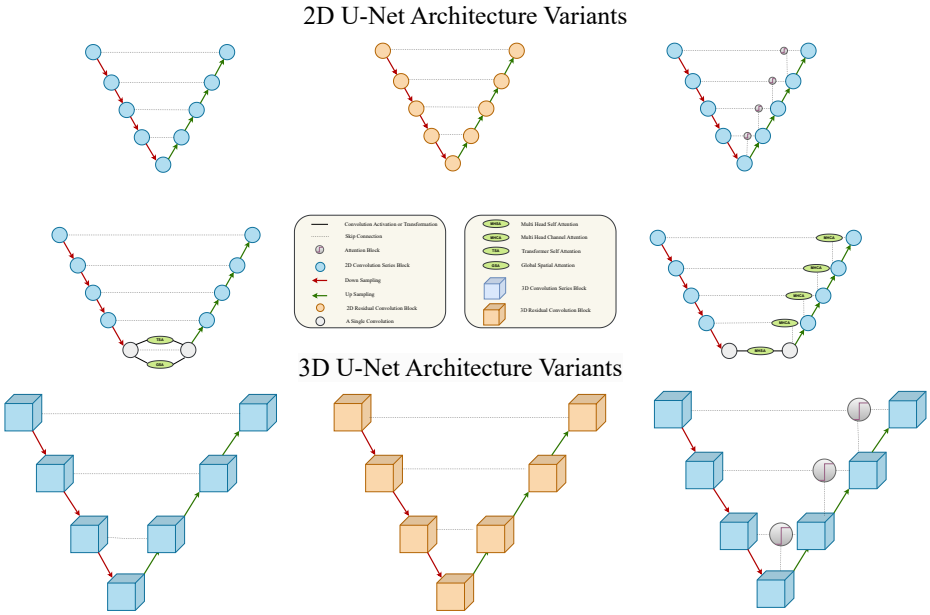

Fig. 1. The figure illustrates various U-Net style architectures. (First row) shows a diagrammatic view of the convolution-based Transformer models and (bottom row) shows two novel transformer-based U-Net architectures. We detailed all the symbols and signs used in the legend block.

Attention U-Net: The fundamental concept of *attention* was formulated by Bahdanau *et al.* [4]. Later Oktay *et al.* [27] applied this mechanism as '*Attention Gate*' (AG), which improved segmentation with detailed localization of multiple organs.

In this architecture, the core component is the attention gate which aids the U-Net in segmenting desired lesions. The architecture style in downsampling is similar to U-Net, i.e., the image is quadruply downsampled ($\rightarrow \times 4$) to obtain the latent space. In traditional U-Net, upsampling ($\rightarrow$) is achieved using transpose convolutions, and additional representations are aggregated from the skip connections. But in Attention U-Net, the representations from the previous layers and from skip connections are aggregated using AG, and now these features are concatenated with the upsampled ($\rightarrow$) features. Thus, they form cascaded connections resulting in better segmentation. Similarly, this attention gate is applied to 3D convolutions to get volumetric attention [26], [37].

TransAttn U-Net: This architecture was designed by Chen *et al.* [10] in which they propose SAA: Self-Aware Attention, which is an amalgamation of multi-level and multi-scale guided attention mechanisms. In specific, after downsampling features to the embedding space, they perform two attention mechanisms which are Transformer Self-Attention ($\mathcal{F}_{TSA}$) and Global Spatial Atten-

Table 4. The below table illustrates the performance of variants of 2D U-Net architectures. The first three models are pure convolution-based architectures and the remaining two are hybrid networks with convolutions and transformer components. The evaluation criteria implied is the same as Table 2; We report the performance of the model for the test set.

Method	**Performance Metrics (2D Data)**			
	Dice Score	IoU Score	Precision	Recall
U-Net [32]	0.417	0.337	0.580	0.360
Residual U-Net [44]	0.456	0.375	0.592	0.420
Attention U-Net [27]	0.487	0.396	0.636	0.439
TransAttn U-Net [10]	0.572	**0.477**	**0.660**	0.565
U-Net Transformer [29]	**0.583**	0.475	0.659	**0.591**

tion ($\mathcal{F}_{GSA}$). Now, these features are combined into a single convolution block, and with each step of upsampling, the previous layer features are attached as skip connections using *Bi-linear Upsampling* (refer Fig. 1). The significance of each attention mechanism is elucidated below.

Suppose our image is represented as $X \in \mathbb{R}^{t \times h \times w}$, where t, h, w are time-steps (channels), height, and width of the given image, respectively. The image is passed to the encoder, and the downsampled representation is denoted by $\mathcal{F}^t_{base} \in \mathbb{R}^{t \times (h \times w)}$. Now, to achieve GSA,

$$\mathcal{F}_{GSA}(M, N, W)_i = \sum_{k=1}^{h \times w} (W_k \mathcal{A}_{i,j}) \tag{1}$$

Where, $N \in \mathbb{R}^{t' \times (h \times w)}$ and $M \in \mathbb{R}^{(h \times w) \times t'}$. Also the $\mathcal{A}_{i,j} = \frac{e^{(M_i N_j)}}{\sum_{r=1}^{n} e^{(M_r N_j)}}$. This $\mathcal{A}_{i,j}$ measures the input of the i^{th} and j^{th} position. Similarly, the TSA attention is calculated as,

$$\mathcal{F}_{TSA}(K, Q, V) = soft\left(\frac{QK^T}{(d_k)^{1/2}}\right) V \tag{2}$$

Where K, Q, V are just the features of $\mathcal{F}^t_{base}$ added with positional encoding and d_k is the dimensionality of any key or Query or value sequence (i.e., $d_k = |V|or|Q|or|K|$ and $soft(.)$ is the softmax activation function [14].) These attentions are operated at the latent space or base, and now, in final step, all these features are amalgamated at the latent space as,

$$\mathcal{F}_{SAA} = \psi_1 \mathcal{F}_{TSA} + \psi_2 \mathcal{F}_{GSA} + \mathcal{F}_{base} \tag{3}$$

Where ψ_1 and ψ_2 are the scale parameters, respectively, and they control the importance assigned to each attention mechanism. Initially, they are assigned with null weights and gradually incremented to obtain a systematic consistency.

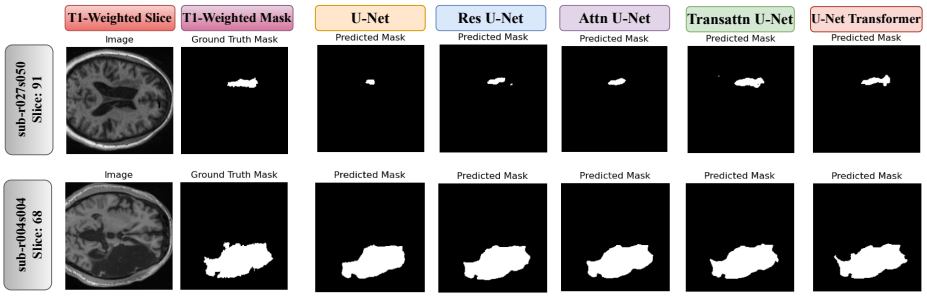

Fig. 2. 2D visualizations of the benchmarks between ground truth and predicted lesions for two subjects. As can be seen, we have two different sizes of stroke lesion subjects included for visualization (Subject ID: $sub-r027s050$ slice 91 and $sub-r004s004$ slice 68) of which one is small and the other being large. We display the predicted outputs of convolution and transformer-based 2D U-Net models for these subjects. All 2D models performed equally better but the U-Net transformer gave the finest boundaries as visible for both the subjects.

Table 5. The below table illustrates the performance of variants of 3D U-Net architectures. The models that are implied below are pure convolution-based architectures and the evaluation criteria implied are the same as Table 3; We report the performance of the model for the test set.

Method	**Performance Metrics (3D Data)**			
	Dice Score	IoU Score	Precision	Recall
U-Net [11]	0.450	0.350	0.584	0.444
Residual U-Net [44]	**0.504**	**0.393**	**0.585**	0.533
Attention U-Net [27]	0.469	0.369	0.498	**0.578**

U-Net Transformer: This method originated from the work by Petit *et al.* [29]. The authors impart self-attention and channel attention modules in this work to produce interpretative segmentation throughput. Fundamentally, the self-attention module uses Multi-head Self-Attention (MHSA) which is similar to Vaswani *et al.* [35], and this aims to acquire long-range structural information from the images that were downsampled to the latent-space. The underlying operation is quite similar to equation (2).

In the channel-attention module, the representations from skip connections (at each pooling step) are first applied with MHSA and then concatenate the features coming from latent space after each upsampling step. Thus this module is referred to as Multi-Head Channel-Attention (MHCA) as initially the input is transferred to MHSA and then concatenated with a cross-attention mechanism. The diagrammatic explanations are elucidated in Fig. 1.

4 Results and Discussion

In this section, we experiment with the aforementioned methods as summarized in Section 3. First, we evaluate the performance of 2D and 3D models. Later, we conducted the Wilcoxon test using the ground truth and predicted lesion volume for 3D U-Net models.

4.1 Results for 2D

Among the 2D U-Net convolution-based models the attention U-Net has proven to have a significant Dice score of 0.487 (with an extra verge of 7.0% dice score from baseline[3]). Whereas the hybrid transformer- and convolution-based models tend to provide a noteworthy performance of 0.583 and 0.572 dice scores. These models superseded 2D standard U-Net with an additional 15.5 and 16.6% of dice score respectively. All these results are illustrated in Table 4.

As there were no additional augmentations applied we can directly infer that adding self-attention components (latent space) such as TSA, MHSA, and GSA played a crucial role in providing significant performance with their cascaded attention [40].

In this regard, we have considered two subjects for visualizing the performance of each model in predicting lesions. In Figure 2 we have displayed the potential of each model to segment small lesions (first row) and large lesions (second row). All the models were near good in segmenting the large volume of lesions but the U-Net Transformer is able to accurately delineate the boundaries of tissues from T1-weighted MRI. But, most of the models struggle to extract the small lesions. Hybrid models such as transformer-based architecture did perform well (comparatively) to a certain extent but, still, there is a wide scope for developing novel models.

4.2 Results for 3D

Now, we consider 3D U-Net style architectures which include standard U-Net, Residual, and Attention U-Net. Due to the added temporal relationship among the features in 3D convolution, the standard 3D U-Net was able to achieve a 0.45 dice score. Previously, in 2D modality, the convolution-based model did not achieve more than a 0.47 dice score. However, 3D modalities, both standard and residual U-Net have a decent increment in performance without any data augmentation [36]. The shift of modality from 2D to 3D, for Residual U-Net and standard U-Net, had an increment of 7.7% and 4.3% of dice score respectively. But, attention U-Net did not succeed in incriminating its performance by shifting from 2D to 3D. The reason behind it might be due to a lack of augmentations

[3] In this paper we consider, standard U-Net to be our baseline for both 2D and 3D models respectively.

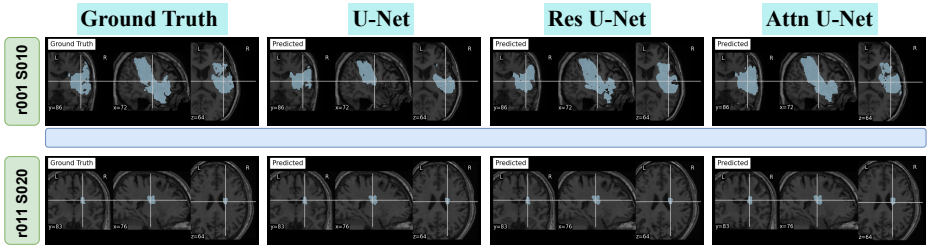

Fig. 3. The above visualization was considered from the test set (ID: $sub-r001s010$ and $sub-r011s020$) and compares three 3D U-Net models (standard, Residual, and Attention). For each model, the left part remains as ground truth and the right part is the model's predictions. In each image, the visualization elucidates the precise location of stroke in the brain using three axes (sagittal, coronal, and axial).

and an insufficient number of samples for training[4]. The results for these models under different performance metrics are illustrated in Table 5.

To specifically study the behaviour of 3D segmentation models we visualized certain test samples one of which is visualized in Figure 3. We have considered two test samples and visualized each plot in three different axes. In each image, the first image describes the Coronal plane and the middle one is the Sagittal plane. Finally, the terminal one is an axial plane. The combination of these three views gives us an estimate of the 3-dimensional pattern of stroke lesions in the brain. As can be seen from Figure 3, U-Net is unable to provide good segmentation outcomes. Though Residual and Attention U-Net were able to segment well but not on par with the mask.

Wilcoxon Signed Rank Test: Now, we also study the prominence of lesion volume using the ground truths [36]. This is achieved by conducting a statistical test, specifically, the Wilcoxon Signed Rank test, and studying whether the lesion volume distribution patterns for each test subject are similar or not [5]. Thus, we establish the results for all three 3D U-Nets (Refer Figure 4). For U-Net and Residual U-Net, the test rejects the null hypothesis and whereas, for attention U-Net it accepts the test with a p-value of 0.54. The detailed results of the Wilcoxon test are detailed in the supplementary material. Also, we visualize box plots to assess the original and predicted volume distribution for the test samples as in Figure 4.

The distribution of pixels changes after resizing them to a certain shape in the case of 2D. As 2D models are often shrunk and stretched, based on the model, and doing so can misguide the volume calculation. Thus, we cannot estimate the

[4] The authors have experimented with various optimizers, learning schedulers, and many more hyperparameters.

[5] This test could be understood as a non-parametric t-test. A detailed premier is provided in the supplementary material.

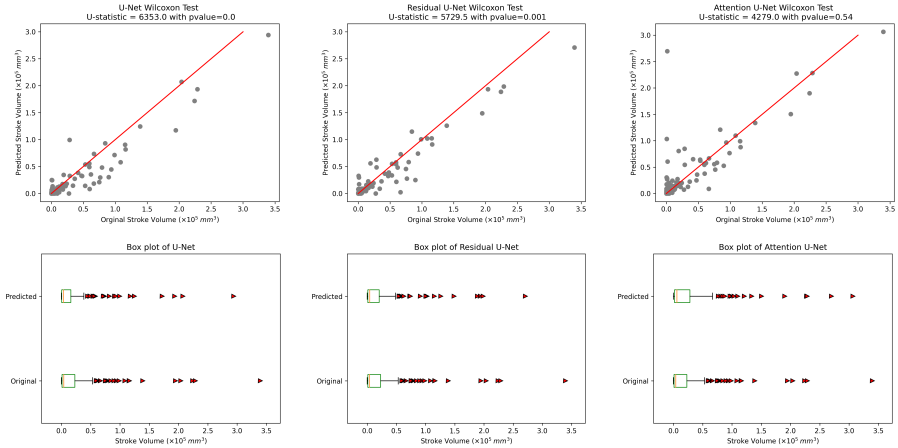

Fig. 4. The Wilcoxon test is carried out for all three 3D U-Net style architectures. The first row indicates a scatter plot and the second one indicates box plots of the predicted and actual stroke volume of the 3D models respectively. Specifically for the scatter plot. The ideal scenario must be the gray dots aligned with the red line.

true volumes in such cases and that is the reason we have only reported for 3D models.

5 Limitations and Future Directions

In the analysis, we have considered a set of state-of-the-art models. But, there are certain limitations in the current work as described below:

- Our motto was to provide standard U-Net style models that are trained on ATLAS v2.0 without any augmentations. Thus, in this current work, we have not used any frameworks such as Deep Medic [21], nnU-Net [18], and MONAI [9].
- The current analysis is only done using models that have masks and so we achieved the results with supervision. Later, this work can be extended with weak-supervision [39], [8] or self-supervision [17] approaches which can aid the learning of models in the absence of masks.
- This work does not focus on uncertainty or ambiguity in decision-making using generative models [6] [23] [31].
- Also, we did not address the issue of very small and disconnected lesions [17].

The limitations described above can be seen as future directions and they could contribute to the progressing field of Neuroimaging for Stroke prediction.

6 Conclusion

This paper fundamentally benchmarks variants of U-Net models on the ATLAS v2.0 dataset and deploys standard stroke lesion segmentation models which could be reproducible both for 2D and 3D brain images. We infer that current 2D or 3D brain imaging prediction requires much more attention towards developing hybrid models with the aid of *self-attention* mechanisms to improve the performance of the models. In the future, we tend to develop fine-grained segmentation models with data augmentation, multi-modality (Diffusion Weighted Imaging and T2-FLAIR), and cascaded attention mechanisms. We hope this research could progress and contribute to Stroke prediction.

Acknowledgement

The authors would like to acknowledge Manasa Kondamadugu for her invaluable coordination efforts throughout the project. Additionally, we extend our gratitude to IHub-Data, International Institute of Information and Technology, Hyderabad for their generous funding and support.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012)
2. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J.: Deep learning for brain mri segmentation: state of the art and future directions. Journal of digital imaging **30**, 449–459 (2017)
3. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual u-net for medical image segmentation. Journal of Medical Imaging **6**(1), 014006–014006 (2019)
4. Bahdanau, Dzmitry, K.C., Bengio, Y.: Neural machine translation by jointly learning to align and translate. ICLR (2014)
5. Baird, A.E., Warach, S.: Magnetic resonance imaging of acute stroke. Journal of Cerebral Blood Flow & Metabolism **18**(6), 583–609 (1998)
6. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 119–127. Springer (2019)
7. Bhalerao, M., Thakur, S.: Brain tumor segmentation based on 3d residual u-net. In: International MICCAI Brainlesion Workshop. pp. 218–225. Springer (2019)
8. Cao, C., Liu, Z., Liu, G., Jin, S., Xia, S.: Ability of weakly supervised learning to detect acute ischemic stroke and hemorrhagic infarction lesions with diffusion-weighted imaging. Quantitative Imaging in Medicine and Surgery **12**(1), 321 (2022)

9. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
10. Chen, B., Liu, Y., Zhang, Z., Lu, G., Kong, A.W.K.: Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274 (2021)
11. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 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, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
12. Ciresan, D., Giusti, A., Gambardella, L., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. Advances in neural information processing systems **25** (2012)
13. Du, X., Ma, K., Song, Y.: Agmr-net: Attention-guided multiscale recovery framework for stroke segmentation. Computerized Medical Imaging and Graphics **101**, 102120 (2022)
14. Elfadel, I.M., Wyatt Jr, J.L.: The" softmax" nonlinearity: Derivation using statistical mechanics and useful properties as a multiterminal analog circuit element. Advances in neural information processing systems **6** (1993)
15. Grady, L.: Random walks for image segmentation. IEEE transactions on pattern analysis and machine intelligence **28**(11), 1768–1783 (2006)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
17. Huo, J., Chen, L., Liu, Y., Boels, M., Granados, A., Ourselin, S., Sparks, R.: Mapping: Model average with post-processing for stroke lesion segmentation. arXiv preprint arXiv:2211.15486 (2022)
18. Isensee, F., Jaeger, P.F., Kohl, S.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)
19. Isensee, F., Maier-Hein, K.H.: An attempt at beating the 3d u-net. arXiv preprint arXiv:1908.02182 (2019)
20. Johnson, W., Onuma, O., Owolabi, M., Sachdev, S.: Stroke: a global response is needed. Bulletin of the World Health Organization **94**(9), 634 (2016)
21. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis **36**, 61–78 (2017)
22. Kim, J., Thayabaranathan, T., Donnan, G.A., Howard, G., Howard, V.J., Rothwell, P.M., Feigin, V., Norrving, B., Owolabi, M., Pandian, J., et al.: Global stroke statistics 2019. International Journal of Stroke **15**(8), 819–838 (2020)
23. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
24. Liew, S.L., Anglin, J.M., Banks, N.W., Sondag, M., Ito, K.L., Kim, H., Chan, J., Ito, J., Jung, C., Khoshab, N., et al.: A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Scientific data **5**(1), 1–11 (2018)

25. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
26. Nodirov, J., Abdusalomov, A.B., Whangbo, T.K.: Attention 3d u-net with multiple skip connections for segmentation of brain tumor images. Sensors **22**(17), 6501 (2022)
27. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
28. Paing, M.P., Tungjitkusolmun, S., Bui, T.H., Visitsattapongse, S., Pintavirooj, C.: Automated segmentation of infarct lesions in t1-weighted mri scans using variational mode decomposition and deep learning. Sensors **21**(6), 1952 (2021)
29. Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., Soler, L.: U-net transformer: Self and cross attention for medical image segmentation. In: Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12. pp. 267–276. Springer (2021)
30. Qi, K., Yang, H., Li, C., Liu, Z., Wang, M., Liu, Q., Wang, S.: X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 247–255. Springer (2019)
31. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11536–11546 (2023)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
33. Sheng, M., Xu, W., Yang, J., Chen, Z.: Cross-attention and deep supervision unet for lesion segmentation of chronic stroke. Frontiers in Neuroscience **16**, 836412 (2022)
34. 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)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
36. Verma, K., Kumar, S., Paydarfar, D.: Automatic segmentation and quantitative assessment of stroke lesions on mr images. Diagnostics **12**(9), 2055 (2022)
37. Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N.: Volumetric attention for 3d medical image segmentation and detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 175–184. Springer (2019)
38. Wilcoxon, F., Katti, S., Wilcox, R.A., et al.: Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. Selected tables in mathematical statistics **1**, 171–259 (1970)

39. Wu, K., Du, B., Luo, M., Wen, H., Shen, Y., Feng, J.: Weakly supervised brain lesion segmentation via attentional representation learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 211–219. Springer (2019)
40. Wu, Z., Zhang, X., Li, F., Wang, S., Huang, L., Li, J.: W-net: A boundary-enhanced segmentation network for stroke lesions. Expert Systems with Applications p. 120637 (2023)
41. Yu, W., Fang, B., Liu, Y., Gao, M., Zheng, S., Wang, Y.: Liver vessels segmentation based on 3d residual u-net. In: 2019 IEEE international conference on image processing (ICIP). pp. 250–254. IEEE (2019)
42. Yu, W., Huang, Z., Zhang, J., Shan, H.: San-net: Learning generalization to unseen sites for stroke lesion segmentation with self-adaptive normalization. Computers in Biology and Medicine **156**, 106717 (2023)
43. Zhang, Y., Wu, J., Liu, Y., Chen, Y., Wu, E.X., Tang, X.: Mi-unet: multi-inputs unet incorporating brain parcellation for stroke lesion segmentation from t1-weighted magnetic resonance images. IEEE Journal of Biomedical and Health Informatics **25**(2), 526–535 (2020)
44. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018)
45. Zhou, Y., Huang, W., Dong, P., Xia, Y., Wang, S.: D-unet: a dimension-fusion u shape network for chronic stroke lesion segmentation. IEEE/ACM transactions on computational biology and bioinformatics **18**(3), 940–950 (2019)

Supplementary Material

Related Works: This section explains each model that was detailed in Table 1 that elucidates their underlying novelty.

Huo *et al.* [17] presented a stroke lesion segmentation based on the nnU-Net framework and applied it to the ATLAS v2.0 dataset. They also introduced an effective post-processing strategy to enhance segmentation metrics. Their method achieved first place in the 2022 MICCAI ATLAS Challenge with impressive scores, including an average Dice score of 0.6667 and a Lesion-wise F1 score of 0.5643. Verma *et al.* [36] reported for the first time an automated lesion segmentation model on the ATLAS v2.0 dataset which was a 3D U-Net architecture similar to the nnU-Net based framework. The model achieved a Dice similarity coefficient of 0.65. Yu *et al.* [42] proposed SAN-Net, a self-adaptive normalization network and it also incorporates symmetry-inspired data augmentation (SIDA) for improved generalization for unseen sites. Experimental results show better performance compared to recent methods on the ATLAS v1.2 dataset.

Wu *et al.* [40] present a novel two-stage network called W-Net for lesion segmentation in ischemic stroke using multi-modal MRI data. W-Net combines CNN-transformer architecture, introduces a boundary deformation module (BDM) to approximate the target boundary, and achieves the best performance in terms of DSC, HD, and F2 metrics, with scores of 61.76%, 32.47, and 64.60%, respectively, on the ATLAS v1.2 and ISLES2022 datasets. Du *et al.* [13] proposed AGMR-Net, a novel method for stroke lesion segmentation, addressing intra-class inconsistency and interclass indistinction challenges. AGMR-Net utilizes a coarse-grained patch attention module, a cross-dimensional feature fusion module, and a multiscale deconvolution upsampling module. AGMR-Net achieved impressive results with a high Dice similarity coefficient of 0.594, Hausdorff distance of 27.005 mm, and average symmetry surface distance of 7.137 mm, on the ATLAS v1.2 dataset.

Sheng *et al.* [33] proposes CADS-UNet, a novel cross-attention and deep supervision UNet, for segmenting chronic stroke lesions from T1-weighted MR images. The model incorporates a cross-spatial attention module to enhance spatial focus, a channel attention mechanism to highlight channel characteristics and deep supervision with mixed loss. The model was evaluated on the ATLAS v1.2 with a Dice Similarity Coefficient (DSC) of 0.5564. Qi *et al.* [30] introduced X-Net, a depthwise separable convolution-based model with a Feature Similarity Module (FSM). They evaluated X-Net on the ATLAS v1.2, achieving encouraging results with a Dice Similarity Coefficient (DSC) of 0.4867 and an Intersection over Union (IOU) of 0.3723. Zhuo *et al.* [45] introduces D-UNet, a novel architecture combining 2D and 3D convolution in the encoding stage with a new loss function called Enhance Mixing Loss (EML). The proposed method was tested on the ATLAS v1.2 dataset, and it achieved a Dice Similarity Coefficient (DSC) of 0.5349 ± 0.2763 and a precision of 0.6331 ± 0.295. Zhang *et al.* [43] developed a novel stroke lesion segmentation approach called multi-inputs UNet (MI-UNet), which integrates brain parcellation information (GM, WM, LV) with original MR image and achieves notable segmentation performance

with a Dice score of 56.72%, Hausdorff distance of 23.94mm, average symmetric surface distance of 7.00mm, and precision of 65.45%.

Evaluation Metrics: For segmenting brain stroke lesions, we employed the mean Dice Similarity Coefficient (DSC), Intersection over Union (IOU), Precision, and Recall as our evaluation metrics. These metrics were utilized to assess the performance of our segmentation models.

Dice Similarity Coefficient (DSC): For the segmentation problem Dice Similarity Coefficient (DSC) is one of the important parameters used to assess the similarity between predicted and ground truth pixel or voxels values by measuring the overlap between two sets. Where X is the predicted set of pixels and Y is the ground truth.

$$DSC = \frac{2(X \cap Y)}{X + Y}$$

Intersection Over Union (IOU): IOU is an evaluation metric commonly used in image segmentation to assess the performance of the model.

$$IOU = \frac{(X \cap Y)}{(X \cup Y)}$$

RECALL: Also known as sensitivity or True Positive Rate (TPR), is an important evaluation metric used for segmentation. It measures the ability of a model to correctly identify positive instances. Recall qualifies how well the model captures the relevant pixel or voxels that belong to the target class.

$$Recall = \frac{TP}{(TP + FN)}$$

PRECISION: Precision measures how well the model performs in correctly identifying the relevant pixels or voxels that belong to the target class out of all the pixels or voxels predicted as positive.

$$Precision = \frac{TP}{(TP + FP)}$$

Wilcoxon Signed Rank Test Primer: The Wilcoxon Signed Rank Test is a non-parametric statistical test used to compare two related or paired samples. It assesses whether there is a significant difference between the paired observations. It does this by ranking the absolute differences between paired values, summing positive and negative ranks separately, and comparing these sums using a specific statistical distribution. If the test statistic is significant, it indicates a significant difference between the two groups. It's often used when data doesn't meet the assumptions of parametric tests like the t-test. The null hypothesis of the Wilcoxon test is that there is no difference between the distributions of the two groups, while the alternative hypothesis suggests that there is a significant

difference [38]. The Python library SciPy is used to generate the results for the Wilcoxon test: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html`.

Loss Function In medical image analysis, we often use a combination of Dice Loss (DICE) and Binary Cross-Entropy Loss (BCE) for segmentation tasks. Dice Loss quantifies the overlap between predicted and actual regions of interest, while Binary Cross-Entropy Loss measures the dissimilarity between predicted probabilities and ground truth labels. By combining these two losses, we can effectively train deep learning models to segment medical images, helping in stroke lesion segmentation accurately. However, we have provided proportionate weightage for each of the loss functions, and the combined loss is described in the equation (6).

$$\mathcal{L}_{DICE} = 1 - \frac{2 \cdot \sum_{i=1}^{N} p_i \cdot g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{4}$$

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [g_i \cdot \log(p_i) + (1 - g_i) \cdot \log(1 - p_i)] \tag{5}$$

$$\mathcal{L}_{OURS} = \gamma \mathcal{L}_{DICE} + (1 - \gamma) \mathcal{L}_{BCE} \tag{6}$$

Where p_i is the predicted probability of a pixel/voxel being in the ROI, g_i is the corresponding ground truth label (1 for ROI, 0 for background), and N is the number of pixels/voxels. Here, γ is the weightage parameter, and we have obtained the best results for $\gamma = 0.9$, and thus, all the models (both 2D and 3D) have used this value as default for a fair evaluation.

2D Parameter description: We implemented five U-Net style architectures for 2D image segmentation, namely, 2D U-Net, ResU-Net, Attention U-Net, TransAttn U-Net, and U-Net Transformer. These models possess approximately 31 M, 32 M, 34 M, 25 M, and 11 M parameters, respectively, and have 4 levels of up and down samplings, except for the U-Net Transformer, which utilizes 3 levels of ups and downsamplings.

A preprocessing step was implemented by cropping the images (coordinates: (10, 40) and (190, 220)) to focus on relevant regions and resizing them to 192x192 pixels via bilinear interpolation, we standardized the input dimensions and emphasized salient features associated with stroke lesions. During the training process, we employed the Adam optimizer with an initial learning rate of 0.001. To optimize learning adaptability, we incorporated the "ReduceLROnPlateau" learning rate schedule, which dynamically adjusted the learning. Dropout regularization with a probability of 0.2 was employed to mitigate overfitting during training and these models were trained for 50 epochs, utilizing a batch size of 32 for all models except for the U-Net Transformer, which had a batch size of 16. The 2D architecture of each model consists of multiple convolutional layers

Table 6. Wilcoxon and Pearson correlation test for actual and predicted stroke volumes for 3D standard, Residual, and Attention U-Net architectures.

Method	**Wilcoxon Test**		**Pearson Correlation**	
	U-statistic	p-value	statistic	p-value
U-Net [32]	6353.0	≈ 0.0	0.949	≈ 0.0
Residual U-Net [44]	5729.5	0.001	0.962	≈ 0.0
Attention U-Net [27]	4279.0	0.540	0.844	≈ 0.0

with varying numbers of channels (64, 128, 256, 512, 1024) to effectively capture diverse and intricate features in the data. However, the U-Net Transformer model has (64, 128, 256, and 512) channels. Notably, data augmentations were not applied to the all 2D U-Net models.

3D Parameter description: Here we implemented three state-of-the-art 3D models for image segmentation: U-Net 3D, Res U-Net 3D, and Attention U-Net 3D. Each model's number of learnable parameters was meticulously assessed, revealing U-Net 3D to possess 1.40 M parameters, Res U-Net 3D with 1.42 M parameters, and Attention U-Net 3D featuring 1.61 M parameters. These models were intricately designed with 3 levels of up-and-down samplings. Regarding input image and mask dimensions, both U-Net 3D and Res U-Net 3D were configured with image dimensions of $144 \times 172 \times 128$, while Attention U-Net 3D adopted the dimensions of $144 \times 176 \times 128$.

The Adam optimizer is employed for training the models, with an initial learning rate of 0.001 to facilitate convergence. The learning rate schedule "CosineAnnealingLR" is applied to all models, with ADAM optimization. For training and validation batches, a batch size of 4 is utilized for all models. Additionally, the early stopping criterion is set at 100 epochs for all models to control training duration. Regarding the number of channels used, the models incorporate 16, 32, 64, and 128 channels, allowing them to capture diverse and intricate image features. To prevent overfitting, a weight decay value of 0.0001 is employed for all models. No data augmentations are applied during training.

Visualization Source: For, the 3D visualizations the authors have used Ni-Learn framework: `https://nilearn.github.io/dev/index.html` and for 2D visualizations Matplotlib framework: `https://matplotlib.org/`.

Concerns of HD95 Evaluation Metric: We are thankful and are obliged from the fruitful reviews provided by anonymous reviewers. One of the reviewers suggested using the HD95 score as one of the evaluation metrics. However, the evaluation metrics used in our study are well-established in the literature and accepted across various domains in medical image segmentation. Apart from these metrics, HD95 provides surface-based information and provides additional

insights into the delineation of stroke. Due to time constraints, we were not able to provide them. But, we are very much willing to add those in our future work.