Deep Learning and Multi-Modal MRI for the Segmentation of Sub-Acute and Chronic Stroke Lesions Authors

Lounès Meddahi, Stéphanie s Leplaideur, Arthur Masson, Isabelle Bonan, Elise Bannier, Francesca Galassi

TITLE

Deep Learning and Multi-Modal MRI for the Segmentation of Sub-Acute and Chronic Stroke Lesions

Authors

Lounès Meddahi[1], Stéphanie s Leplaideur[2, 4, 5] , Arthur Masson[1] , Isabelle Bonan[1,2], Elise Bannier[1,3], Francesca Galassi[1]

[1] Univ Rennes, Inria, CNRS, Inserm, IRISA UMR 6074, Empenn, Rennes, France

[2] CHU Rennes, Physical medicine and rehabilitation department, Rennes, France

[3] CHU Rennes, Radiology Department, Rennes, France

[4] CIC - Centre d'Investigation Clinique

[5] CMRRF - Centre Mutualiste de Rééducation et de Réadaptation Fonctionnelles de KERPAPE

ABSTRACT

Background: Stroke is a leading cause of morbidity and mortality worldwide. Accurate segmentation of sub-acute and chronic stroke lesions using MRI is crucial for assessing brain damage and developing effective rehabilitation plans. Manual segmentation is labor-intensive and error-prone, necessitating automated approaches. This study aims at improving sub-acute and chronic stroke lesion segmentation using deep learning and multi-modal MRI data. Both models are made available to the research community.

Methods: This study developed and evaluated two models for segmenting sub-acute and chronic stroke lesions using MRI: a single-modality model trained on the public ATLAS v2.0 dataset, and a dual-modality model adapted from the single-modality model by integrating T1-w and FLAIR MRI data from an internal dataset. Both models were trained using the nnU-Net framework, employing a preprocessing pipeline to improve the segmentation accuracy.

Results: The single-modality model achieved a mean Dice score of 83.0% on the ATLAS v2.0 dataset, and 68.8% on the internal test set. The dual-modality model significantly improved segmentation accuracy, yielding a mean Dice score of 75.6% and an F1 score of 72.6% on the internal test set. Additionally, volumetric analysis showed a high Pearson correlation coefficient (0.94) between predicted and actual lesion volumes.

Conclusions: The improved performance of the dual-modality model suggests the benefit of integrating FLAIR MRI to capture lesion characteristics in detecting and segmenting sub-acute and chronic stroke lesions. This could lead to more accurate assessment of brain damage and more effective rehabilitation plans for stroke patients. Future research should focus on larger multi-modal datasets and further investigate segmentation challenges, as well as clinical validation.

Keywords: chronic stroke, MRI, lesion segmentation, deep learning

INTRODUCTION

Stroke is a major global cause of morbidity and mortality, occurring when there is a sudden disruption of blood supply to the brain, leading to severe neurological impairments [1–4]. Medical imaging, particularly Magnetic Resonance Imaging (MRI), plays a crucial role in assessing the extent of brain damage caused by stroke. MRI is essential both during the acute phase, occurring immediately after the stroke, and the chronic phase, which may occur weeks or months later [5]. Rapid identification of brain lesions in the acute phase is vital for determining optimal treatment strategies and mitigating potential long-term effects, often utilizing techniques like diffusion-weighted MRI and perfusion MRI.

In the sub-acute and chronic phase, conventional MRI, typically including T1-weighted (T1-w) scans, is used to examine the consequences and underlying causes of stroke [6,7]. Accurately segmenting sub-acute and chronic brain lesions is crucial for physicians to assess the extent of damage, understand neurological sequelae, and establish a prognosis for the patient. It helps pinpoint the affected brain regions precisely, facilitating the development of specific rehabilitation plans to help recover lost or impaired functions. Moreover, by monitoring changes in brain lesions over time, segmentation helps track disease progression and treatment effectiveness, allowing for necessary therapeutic adjustments.

While manual segmentation is considered the gold standard, it is time-consuming and error-prone due to the complexity of brain anatomy, the varied patterns and locations of lesions, and differences between patients. Significant progress has been made in automating lesion segmentation in the acute phase [8–10], particularly with recent advancements in deep learning. However, solutions for chronic lesion segmentation remain limited due to several factors. Firstly, the complexity of chronic lesions, which are often less distinct and more diffuse than acute lesions, makes segmentation more challenging. Secondly, clinical imaging protocols for chronic lesions (i.e., MRI sequences) are not standardized, complicating the development and validation of segmentation methods. Additionally, annotated and shared datasets for chronic lesions are scarce, limiting researchers' ability to train and evaluate automatic segmentation models. Finally, chronic lesions can evolve over time, adding another layer of complexity to automatic segmentation.

The most recent review on chronic stroke lesion segmentation, conducted by Ahmed et al. [11], offers a thorough analysis and comparison of existing automated methods. These methods generally employ deep learning models trained and tested on the single-modality ATLAS dataset [12], which contains T1-w MRI

images from hundreds of subjects in the sub-acute and chronic stages of stroke, along with corresponding lesion segmentation masks. ATLAS v2.0, the latest iteration of this dataset, is regarded as a benchmark in the field for developing and validating segmentation models. Our study leverages ATLAS v2.0 to train and assess our models, directly comparing our approach with state-of-the-art methods proposed by Verma et al. [13] (Dice score of 0.65) and Huo et al. [14] (Dice score of 0.67).

In this paper, we present two key contributions to the field of sub-acute and chronic stroke lesion segmentation. Firstly, we developed a single-modality model leveraging the ATLAS v2.0 dataset in conjunction with the nn-Unet framework [15], known for its superior performance in medical image segmentation tasks. Building on our previous work in multiple sclerosis (MS) lesion segmentation [16,17], we adapted its application to sub-acute and chronic stroke lesion segmentation. Secondly, we refined this model to generate a dual-modality model trained on both T1-w and FLAIR MRI data from an internal dataset, exploring the potential benefits of integrating a second modality compared to the single-modality approach. Preliminary results of our study were presented at the World Congress of Neurorehabilitation: WCNR 2024 [18].

MATERIAL AND METHODS

Datasets. A public single-modality dataset, ATLAS v2.0 dataset [12], was used to develop our baseline model, and our internal dataset was used to develop a dual-modality model. The internal dataset consists of two parts: Dataset A, chosen for fine-tuning the baseline model due to its homogeneous nature, and Dataset B, used to test the model. Table 1 summarizes the datasets' statistics, and Figure 1 illustrates the distribution of lesion volumes across each dataset. Below is a detailed description.

Public single-modality ATLAS v2.0 dataset. The ATLAS v2.0 dataset [12] comprises a *training* dataset with 655 T1-w MRI scans and corresponding lesion segmentation masks, a *test* dataset with 300 T1-w MRI scans, and a *hidden* test dataset with 316 T1-w MRI scans. All T1-w MRI scans are aligned to the MNI-152 standard template, with a voxel size of 1 x 1 x 1 mm. We used the *training* dataset for training and validating the baseline single-modality model. The average lesion volume in this dataset is 3.28×10^4 mm³, with individual lesion volumes ranging from 13 mm³ to 4.79×10^5 mm³.

Internal dual-modality datasets. The internal datasets were derived from two clinical studies: the

NeuroFB-AVC study (Dataset A) and the multi-centric AVCPOSTIM study (Dataset B)[1]. All subjects provided written consent; the studies were approved by the relevant ethics committee and complied with French data confidentiality regulations. Both studies enrolled patients in the sub-acute phase (7 days to 6 months post-stroke) and the chronic phase (over 6 months post-stroke), including both ischemic and hemorrhagic stroke types. The NeuroFB-AVC study was conducted on a 3T Magnetom Siemens Prisma scanner. The AVCPOSTIM study included data from various scanners: 3T Magnetom Siemens Prisma, 1.5T Siemens Avento, 3T Philips Medical Systems Achieva, and 3T GE Discovery. T1-w and FLAIR MRI modalities were acquired for each patient. Manual segmentation was performed on the FLAIR modality with the help of the T1-w modality by a neuroimaging expert and reviewed by a neuroradiologist.

Dataset A was used for fine-tuning the baseline model, initially on T1-w alone (*single-mod,* see Table 1), and subsequently on both T1-w and FLAIR (*dual-mod,* see Table 1). The T1-w modality has a mean voxel size of 1 x 1 x 1 mm, while the FLAIR modality has a mean voxel size of 0.75 x 0.75 x 3.3 mm. For Dataset B, the T1-w modality has mean voxel size of (0.942 ± 0.158, 0.942 ± 0.158, 1.463 ± 1.286) mm, and the FLAIR modality has mean voxel size of (0.680 ± 0.224, 0.680 ± 0.224, 1.286 ± 1.638) mm.

Preprocessing. Our approach builds upon our previously proposed framework [16,19], with adjustments to the preprocessing pipeline aimed at improving brain extraction. Specifically, we replaced the previous brain extraction step, which used Anima[2], with the state-of-the-art HD-BET deep learning-based tool [20]. Our preprocessing pipeline comprises the following steps:

1. Brain extraction: The HD-BET tool is used to remove the skull from the images.
2. Re-orientation: The volumes are re-oriented to the RAS (Right-Anterior-Superior) coordinates to ensure a consistent orientation across all images.
3. Registration: If both T1-w and FLAIR images are available for a subject, the T1-w image is rigidly registered to the corresponding FLAIR image using a block matching registration method (*animaPyramidalBMRegistration*). If only the T1-w modality is available, this step is skipped.
4. Bias correction: The bias due to spatial inhomogeneity is estimated using the N4 algorithm [21] and removed from the data (*animaN4BiasCorrection*).

[1] clinicaltrials.gov study IDs: NCT03766113, NCT01677091
[2] https://anima.irisa.fr/

5. Intensity Normalization: Image intensities are standardized by subtracting the mean voxel value and dividing by the standard deviation for each image [22].

These preprocessing steps prepare the data for the nnU-Net framework [15]. Prior to feeding the images into the U-Net model, this framework standardizes voxel spacing across all images. By default, it calculates a uniform target spacing for each axis using the median values from the training cases, and employs third-order spline interpolation for resampling. In the case of anisotropic images (maximum axis spacing / minimum axis spacing > 3) the approach differs: in-plane resampling uses third-order spline interpolation, while out-of-plane interpolation employs nearest neighbor interpolation to minimize resampling artifacts. Segmentation maps are converted to one-hot encodings, and each channel is then interpolated using linear interpolation; the final segmentation mask is obtained by applying the argmax operation. For anisotropic cases, nearest neighbor interpolation is specifically used on the low-resolution axis.

Model Training and Evaluation. We divided the *training* ATLAS v2.0 dataset into training (80%) and validation (20%) sets. After each training epoch, we evaluated model performance using the validation set. To mitigate overfitting, the final model, i.e., the baseline single-modality model, was chosen based on the lowest validation loss. All reported metrics pertain to the validation dataset.

For Dataset A, we employed a 5-fold cross-validation approach, splitting the dataset into five subsets, each containing 20% of the data. In each iteration, one subset acted as the validation set, while the remaining four subsets were used for training. The final evaluation of the model, whether fine-tuned single- or dual-modality, was based on the average performance metrics from each validation set. After completing cross-validation, we assessed the best-performing model on the test dataset, Dataset B, to assess its performance on unseen data.

Segmentation model architecture. Our segmentation core is built around the nnU-Net framework [15], which is based on the U-Net architecture [23]. This framework enables the training of a 3D model with deep supervision, combining soft Dice and cross-entropy loss functions. The model is optimized using stochastic gradient descent with a polynomial decay schedule to adjust the learning rate during training (initial rate = 0.01). Dropout-based regularization (removal rate = 0.2) and data augmentation techniques, such as isotropic rescaling (0.85 to 1.25), 3D rotation (-15° to 15°), and sagittal plane mirroring, are also incorporated.

After obtaining the predicted lesion masks, we applied a threshold of 0.2 to the output probability map, converting it into a binary map. Next, only connected components with a volume of at least 10 mm^3 were retained using a 26-connectivity criterion. The threshold value and minimum volume were determined empirically.

The baseline model was trained using the single-modality ATLAS v2.0 dataset. We then adapted this baseline model to create a dual-modality version that accepts both T1-w and FLAIR inputs. This adaptation involved doubling the input channels and reshaping the convolutional blocks to integrate the FLAIR modality. To prevent the model from losing the information learned from the single-modality dataset, we continued training with a reduced learning rate of 10^{-5}. This approach allowed the model to retain knowledge from the baseline training while adapting to the new dual-modality data.

To facilitate collaborative research and experimentation, we have made the adaptation script available in our Git repository[3]. This script enables the community to replicate our model adaptation process. Importantly, both the baseline single-modality and dual-modality models are available in the same repository.

Evaluation. To evaluate our models, we used a diverse set of metrics to provide a comprehensive assessment and comparison with existing methods [11]. For segmentation performance, we reported precision, sensitivity, Dice score, and Jaccard index [24]. We also included the average surface distance, i.e., the mean distance between boundaries. Lesion detection performance was assessed using the lesion-wise F1 score, where a candidate lesion was considered correctly detected if its connected voxels overlapped with the ground truth by at least 10%.

We evaluated volume prediction accuracy by comparing the ground truth with automatic segmentation results. Specifically, we used a regression line to analyze deviations from the volumes obtained through manual segmentation.

To determine the statistical significance of the observed improvements, we conducted a Wilcoxon signed-rank test. Statistical significance was defined as a *p-value* < 0.05.

RESULTS

[3] https://github.com/LounesMD/MMStrokeNet

Segmentation evaluation. In Table 2, we report the mean and standard deviations for each performance metric and model.

Baseline model. The baseline single-modality model achieved a mean Dice score of 83.0% (median: 85.2%) on the ATLAS v2.0 dataset, indicating strong spatial agreement between predicted lesion segmentations and ground truth. The model was both highly precise, with a mean voxel-wise precision of 83.3% (median: 84.8%), and highly sensitive, with a mean voxel-wise sensitivity of 83.6% (median: 86.8%). Similarly, the Jaccard index reflected high segmentation performance, with a mean value of 72.2% (median: 74.2%). Additionally, the low average surface distance of 2.21 mm indicated excellent spatial accuracy in comparison to the ground truth surfaces.

When evaluated on Dataset B, the baseline model model showed a higher rate of discarded valid lesion voxels, as indicated by a lower voxel-wise sensitivity of 65.1% (median: 75.0%). Consequently, the Dice score dropped to a mean value of 68.8% (median: 74.1%), although it remained competitive with reported performances in the literature. There was also a slight deterioration in both the Jaccard Index and the average surface distance, with the Jaccard index decreasing to a mean of 58.4% (median: 62.7%) and the average surface distance increasing to a mean of 4.45 mm (median: 2.5 mm).

In terms of lesion detection, the mean lesion-wise F1-score of 66.4% (median: 66.7%) on the ATLAS v2.0 dataset, demonstrating the model's ability in identifying lesions. On Dataset B, the F1-score experienced a slight decrease to a mean value of 63.2% (median: 57.1%). Despite this reduction, the performance remains high, suggesting the model's robustness in lesion detection across different datasets.

When comparing our baseline model to recent methods reported in the literature, which were trained and tested on the same ATLAS v2.0 dataset, our model achieves higher scores. Specifically, Verma et al.'s method reported a Dice score of 65% and an average surface distance of 12.04 mm [13], while Huo et al.'s achieved a Dice score of 67% and an F1 score of 56% [14].

Dual-modality model. In Table 3, we present the performance of the dual-modality model on Dataset A, using a 5-fold cross-validation approach. This method involved dividing the dataset into five subsets (folds), training the model on four subsets, and evaluating its performance on the remaining fold. This process was repeated five times, with each fold being the validation set once. The Dice and F1 scores consistently surpassed 80% across all folds, indicating that the model reliably detects and segments lesions in Dataset

A. The best-performing model was selected based on the average score over the five folds; results on Dataset B are reported in Table 3. Box plots for both the dual-modality and baseline models are shown in Figure 2.

The results demonstrate significant improvements in lesion segmentation and detection with the dual-modality model compared to the baseline. Specifically, the Dice score improved by 7%, reaching a mean value of 75.6% (median: 78.0%). The Jaccard score also increased from a mean of 55.5% (median: 58.9%) to 62.9% (median: 63.9%), and the average surface distance decreased from a mean value of 4.45 mm to 3.15 mm. These improvements were statistically significant, as indicated by the Wilcoxon test (p-value << 0.05).

In terms of lesion detection, the dual-modality model achieved a mean F1 score of 72.6% (median: 66.7%), outperforming the baseline model's mean F1-score of 63.2%. This improvement was statistically significant, with a p-value lower than 0.05. Furthermore, the lower standard deviations across all metrics for the dual-modality model indicate improved consistency in performance compared to the baseline.

Single-modality model. The observed improvements in performance after fine-tuning may be attributed to the similar characteristics shared between Dataset A and test Dataset B, both annotated by the same expert, unlike the ATLAS v2.0 dataset. This suggests that the improvement may not only be due to the introduction of the second modality. To investigate this further, we conducted additional fine-tuning exclusively on the T1-w modality, using 5-fold cross validation, and evaluated the best resulting model on Dataset B. The performance scores on Dataset B are reported in Table 2. For a thorough comparison, box plots comparing the outcomes of the fine-tuned single-modality model with the dual-modality model are illustrated in Figure 3.

The results indicate significantly higher overlap scores for the dual-modality model compared to the fine-tuned single-modality model. This trend is also observed in the distance metric, where the dual-modality model consistently outperforms the single-modality model. Regarding lesion detection, the F1-score indicates comparable performances between the two models.

In Figure 4, we present qualitative results illustrating the lesions automatically segmented by the dual-modality model, and reported back in the native space.

Volumetric Analysis. We assessed the accuracy of volume predictions by comparing the ground truth with automatic segmentation results from the dual-modality model. The scatter plot in Figure 5 illustrates how closely the automatic segmentation aligns with the ground truth volumes. The proximity of most data points to the regression line indicates that the automatic method generally predicts lesion volumes accurately, with minimal deviation from manual segmentation. A high Pearson correlation coefficient of 0.94 signifies a strong positive linear relationship between volumes from manual and automatic segmentation methods, with a p-value of 2.5×10^{-16} suggesting the statistical significance of this correlation.

However, a few (labeled) points deviate from the general pattern. Specifically, patients P9 and P20 are instances where the automatic segmentation model predicted lower lesion volumes compared to manual segmentation. To date, inspection of these cases has not revealed any clear relationship to metadata or image quality that might explain this slight drop in performance. The lesions are extended and challenging to visually delineate. Suggestions for further investigation are discussed in the Discussion section.

DISCUSSION

With this work, we release two models for the segmentation of sub-acute and chronic stroke lesions: a single-modality model trained on the ATLAS v2.0 dataset and a dual-modality model fine-tuned on an internal dataset that includes both T1-w and FLAIR MRI scans. T1-w images, commonly acquired in chronic stroke follow-ups, provide detailed anatomical information, while FLAIR images appear more sensitive to stroke-related changes, such as contrast and hyperintensities. Our single-modality model demonstrated good performance, achieving a Dice score of 83.0%, surpassing previous benchmarks. The dual-modality model achieved higher Dice score and F1 score on our internal test set. This suggests the benefit of using both modalities to capture a wider range of chronic stroke lesion features for improved segmentation accuracy. This work is the first to investigate a dual-modality approach.

With the aim to share our research and ensure reproducibility of results, we provide a detailed description of our algorithm, including clear references to existing code. Additionally, we have made available the scripts developed for this project in a dedicated repository, together with the trained models. These scripts may facilitate the adaptation of a pre-trained single-modality nnU-Net model to a multi-modal architecture, allowing for the transfer of learned features from the source to the target model.

Currently, there is no publicly available dataset that combines both modalities. In the near future, we plan to build a large, well-annotated dual-modality dataset that will be made available to the research community.

While our study shows promising results, it has limitations. The internal dataset used for fine-tuning and testing was relatively small. Although the observed improvements are statistically significant, a larger dataset is necessary to confirm the generalizability of our model. We also noted a few instances of slight performance drops. Further investigation using a larger, more diverse dataset could provide insights into factors that may lead to underestimating lesion volumes, such as variations in image quality, lesion characteristics, or specific segmentation challenges. Lastly, our evaluation focused on detection and segmentation accuracy. Although these are standard metrics, it is also important to assess the ease of integration into clinical workflows and overall usability.

CONCLUSIONS

Our study provides accurate single-modality and dual-modality segmentation models for sub-acute and chronic stroke lesion segmentation. It suggests that deep learning models utilizing multi-modal MRI data can improve the accuracy of sub-acute and chronic stroke lesion segmentation. The dual-modality model surpasses single-modality models, suggesting the advantage of incorporating additional FLAIR MRI sequences. These findings are clinically significant, potentially improving stroke patient assessment and management. Future research should aim to validate these results in larger, diverse cohorts and investigate the integration of these models into clinical workflows.

References

1. Duncan PW, Zorowitz R, Bates B, Choi JY, Glasberg JJ, Graham GD, Katz RC, Lamberty K, Reker D. Management of Adult Stroke Rehabilitation Care. *Stroke*. 2005;36(9):e100–e143.
2. Hankey GJ, Jamrozik K, Broadhurst RJ, Forbes S, Anderson CS. Long-Term Disability After First-Ever Stroke and Related Prognostic Factors in the Perth Community Stroke Study, 1989–1990. *Stroke*. 2002;33(4):1034–1040.
3. Langhorne P, Bernhardt J, Kwakkel G. Stroke rehabilitation. *The Lancet*. 2011;377(9778):1693–1702.
4. Feigin VL, Stark BA, Johnson CO, Roth GA, Bisignano C, Abady GG, Abbasifard M, Abbasi-Kangevari M, Abd-Allah F, Abedi V, Abualhasan A, Abu-Rmeileh NM, Abushouk AI, Adebayo OM, Agarwal G, et al. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Neurology*. 2021;20(10):795–820.
5. Bernhardt J, Hayward KS, Kwakkel G, Ward NS, Wolf SL, Borschmann K, Krakauer JW, Boyd LA, Carmichael ST, Corbett D, Cramer SC. Agreed definitions and a shared vision for new standards in stroke recovery research: The Stroke Recovery and Rehabilitation Roundtable taskforce. *International Journal of Stroke*. 2017;12(5):444–450.
6. Bonan I, Leplaideur S, Carson P. Rééducation de l'équilibre après accident vasculaire cérébral. In: Davenne B, Le Breton F, eds. *Accident vasculaire cérébral et médecine physique et de réadaptation: Actualités en 2010*. Paris: Springer; 2010:37–44.
7. Jamal K, Cordillet S, Leplaideur S, Rauscent H, Cogné M, Bonan I. Reliability and minimal detectable change of body-weight distribution and body sway between right and left brain-damaged patients at a chronic stage. *Disability and Rehabilitation*. 2023;45(2):260–265.
8. Hs M, L H, A M, B O-G, M H, Ct B, W F, A B. Automated multimodal segmentation of acute ischemic stroke lesions on clinical MR images. *Magnetic resonance imaging*. 2022;92.
9. Wong KK, Cummock JS, Li G, Ghosh R, Xu P, Volpi JJ, Wong STC. Automatic Segmentation in Acute Ischemic Stroke: Prognostic Significance of Topological Stroke Volumes on Stroke Outcome. *Stroke*. 2022;53(9):2896–2905.
10. Hernandez Petzsche MR, de la Rosa E, Hanning U, Wiest R, Valenzuela W, Reyes M, Meyer M, Liew S-L, Kofler F, Ezhov I, Robben D, Hutton A, Friedrich T, Zarth T, Bürkle J, et al. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*. 2022;9(1):762.
11. Ahmed R, Al Shehhi A, Hassan B, Werghi N, Seghier ML. An appraisal of the performance of AI tools for chronic stroke lesion segmentation. *Computers in Biology and Medicine*. 2023;164:107302.
12. Liew S-L, Anglin JM, Banks NW, Sondag M, Ito KL, Kim H, Chan J, Ito J, Jung C, Khoshab N, Lefebvre S, Nakamura W, Saldana D, Schmiesing A, Tran C, et al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Scientific Data*. 2018;5(1):180011.
13. Verma K, Kumar S, Paydarfar D. Automatic Segmentation and Quantitative Assessment of Stroke Lesions on MR Images. *Diagnostics*. 2022;12(9):2055.
14. Huo J, Chen L, Liu Y, Boels M, Granados A, Ourselin S, Sparks R. MAPPING: Model Average with Post-processing for Stroke Lesion Segmentation. 2022.
15. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*. 2021;18(2):203–211.
16. Masson A, Le Bon B, Kerbrat A, Edan G, Galassi F, Combes B. A nnUnet implementation of new lesions segmentation from serial FLAIR images of MS patients. In: *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*.; 2021.
17. Combès B, Kerbrat A, Pasquier G, Commowick O, Le Bon B, Galassi F, L'Hostis P, El Graoui N, Chouteau R, Cordonnier E, Edan G, Ferré J-C. A Clinically-Compatible Workflow for Computer-Aided Assessment of Brain Disease Activity in Multiple Sclerosis Patients. *Frontiers in Medicine*. 2021;8:740248.
18. Meddahi L, Leplaideur S s, Masson A, Bonan I, Bannier E, Galassi F. Enhancing stroke lesion detection and segmentation through nnU-net and multi-modal MRI Analysis. In: ; 2024:1.
19. Commowick O, Cervenansky F, Cotton F, Dojat M. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure.
20. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, Wick A, Schlemmer H-P, Heiland S, Wick W, Bendszus M, Maier-Hein KH, Kickingereder P. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*. 2019;40(17):4952–4964.

21. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*. 2010;29(6):1310–1320.
22. Carré A, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, Ammari S, Reuzé S, Alvarez Andres E, Estienne T, Niyoteka S, Battistella E, Vakalopoulou M, Dhermain F, Paragios N, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Scientific Reports*. 2020;10(1):12340.
23. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.
24. Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, Pop SC, Girard P, Améli R, Ferré J-C, Kerbrat A, Tourdias T, Cervenansky F, Glatard T, Beaumont J, et al. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Scientific Reports*. 2018;8(1):13650.

Table 1: Overview of the datasets used in our model development. For each dataset: the number of patients, mean lesion volume, minimum lesion volume, and maximum lesion volume.

Dataset	*Number of Patients*	*Mean Lesion Volume (mm^3)*	*Min Lesion Volume (mm^3)*	*Max Lesion Volume (mm^3)*
ATLAS v2.0	655	$3.28 \pm 6.09 \times 10^4$	1.3×10	4.79×10^5
Dataset A	27	$3.51 \pm 4.66 \times 10^4$	7.95×10^2	2.02×10^5
Dataset B	39	$6.01 \pm 6.26 \times 10^4$	3.89×10^2	2.81×10^5

Table 2: Summary of Performance Metrics. The baseline model was trained on ATLAS v2.0 and tested on both ATLAS v2.0 and Dataset B. The single-mod model was obtained by fine-tuning the baseline model on Dataset A using only the T1-w modality, while the dual-mod model was obtained by fine-tuning the baseline on Dataset A on both T1-w and FLAIR. Scores that are significantly better across the three models are in ***bold****.*

Model	*Test Dataset*	*Dice*	*Sensitivity*	*Precision*	*Jaccard*	*AvgSurfDist*	*F1*
Baseline	*ATLAS v2.0*	*0.830 (0.119) 0.852*	*0.836 (0.135) 0.868*	*0.833 (0.109) 0.848*	*0.722 (0.132) 0.742*	*2.210 (4.528) 1.058*	*0.664 (0.274) 0.667*
Baseline	Dataset B	0.688 (0.204) 0.741	0.651 (0.248) 0.750	0.796 (0.162) 0.834	0.555 (0.212) 0.589	4.448 (5.433) 2.5	0.632 (0.273) 0.571
Single-mod	Dataset B	0.712 (0.198) 0.769	0.705 (0.243) 0.807	0.775 (0.178) 0.842	0.584 (0.211) 0.627	4.769 (8.899) 2.219	0.714 (0.288) 0.667
Dual-mod	Dataset B	**0.756 (0.158)** 0.780	**0.762 (0.199)** 0.823	0.793 (0.117) 0.819	**0.629 (0.173)** 0.639	**3.148 (3.060)** 1.954	**0.726 (0.240)** 0.667

Table 3: Performance metrics across folds. Evaluation metrics are reported for each fold (0-4), along with their calculated averages.

Fold	*Dice*	*Sensitivity*	*Precision*	*Jaccard*	*AvgSurfDist*	*F1*
0	*0.862*	*0.895*	*0.851*	*0.761*	*1.402*	*0.813*
1	*0.907*	*0.929*	*0.887*	*0.833*	*0.721*	*0.893*
2	*0.859*	*0.926*	*0.825*	*0.762*	*1.036*	*0.933*
3	*0.818*	*0.821*	*0.838*	*0.702*	*5.159*	*0.861*
4	*0.832*	*0.848*	*0.818*	*0.731*	*3.359*	*0.917*
Average	0.856	0.884	0.844	0.758	2.335	0.883

Figure 1: Distribution of lesion volumes for individual subjects in each dataset. The mean score is represented by a triangle, and the median score is represented by a yellow bar.

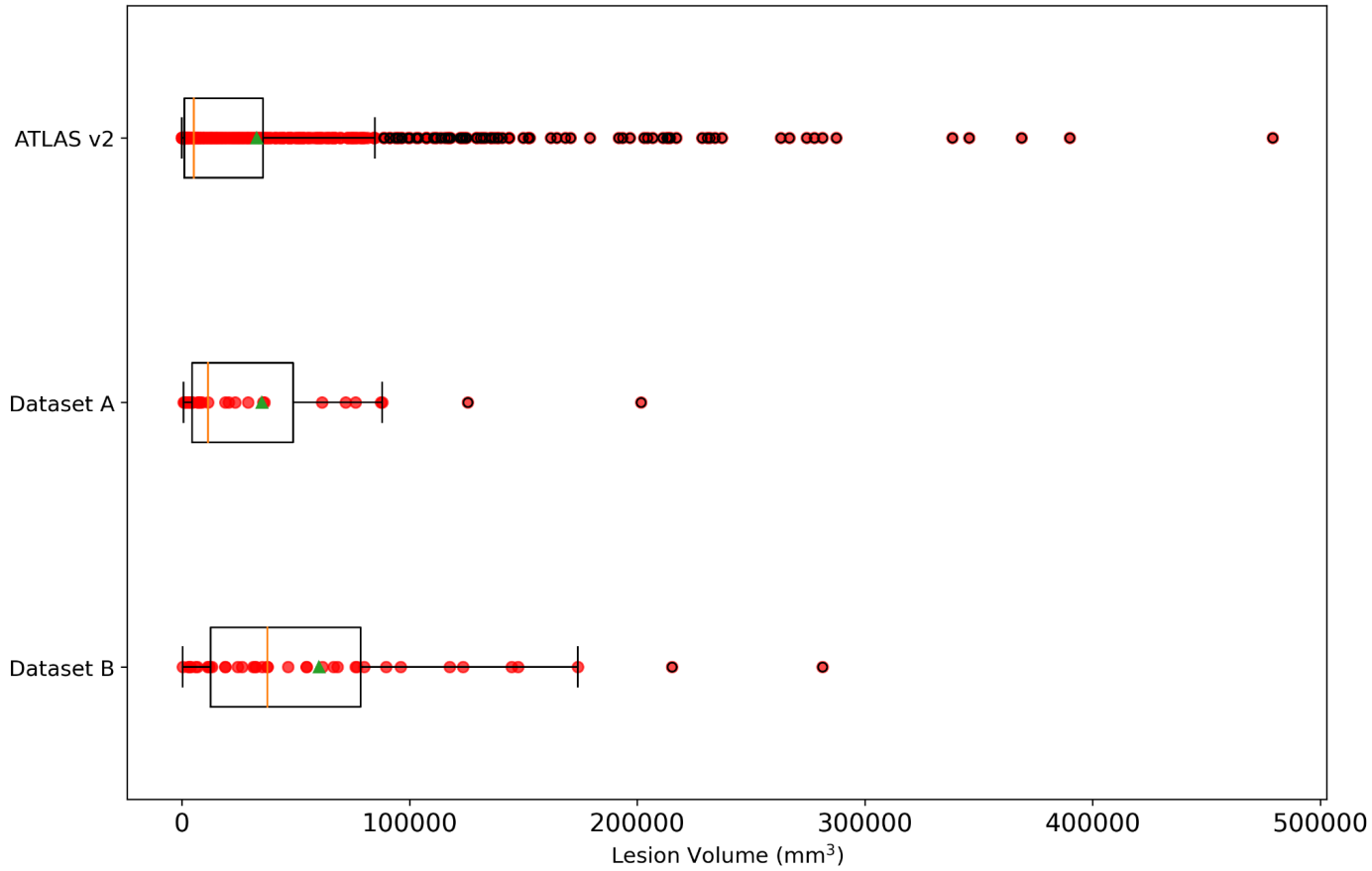

Figure 2: Comparison of the baseline model ATLAS v2.0 versus the dual-modality model.

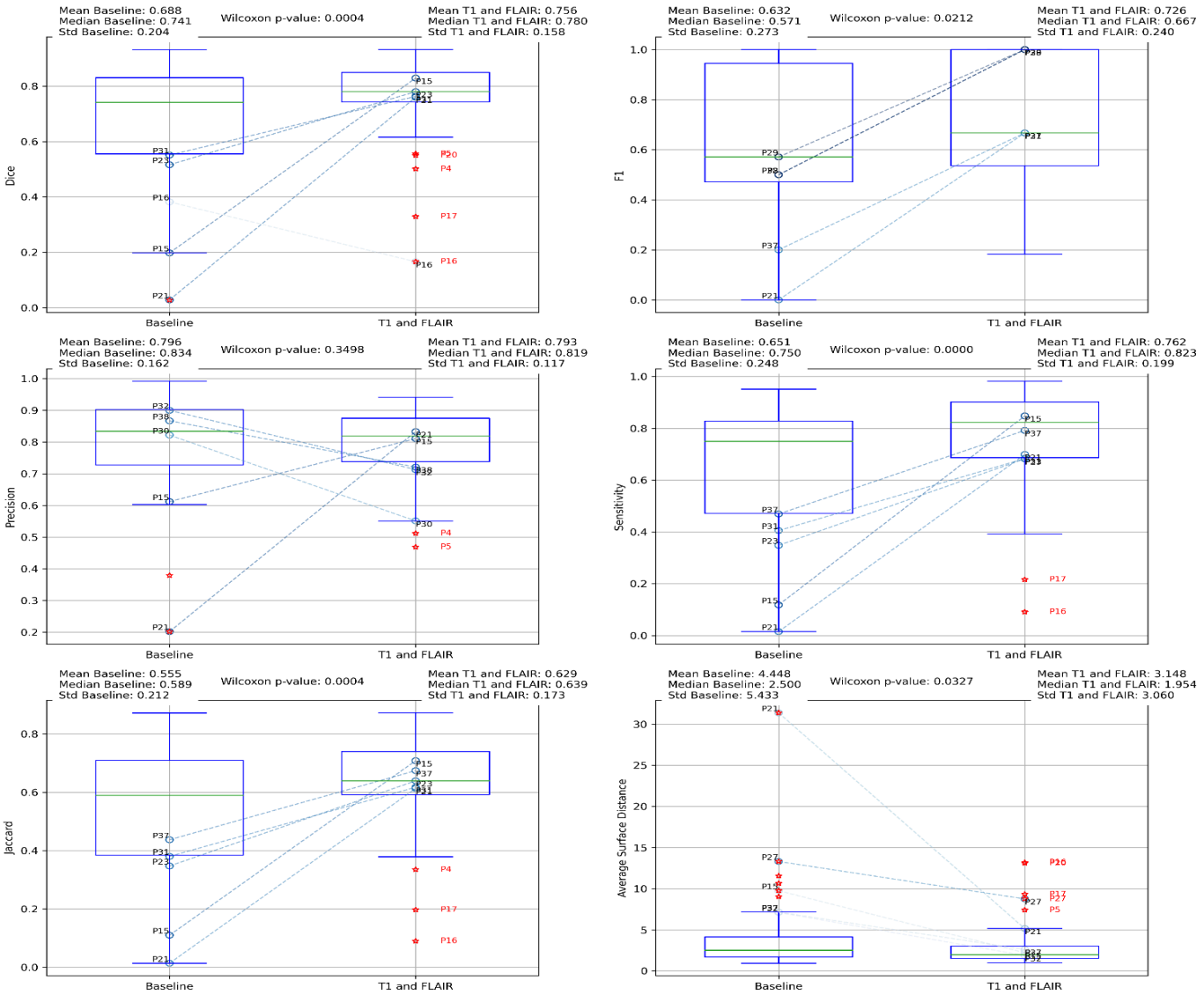

Figure 3: Comparison of the single-modality model fine-tuned on our internal dataset versus the dual-modality model.

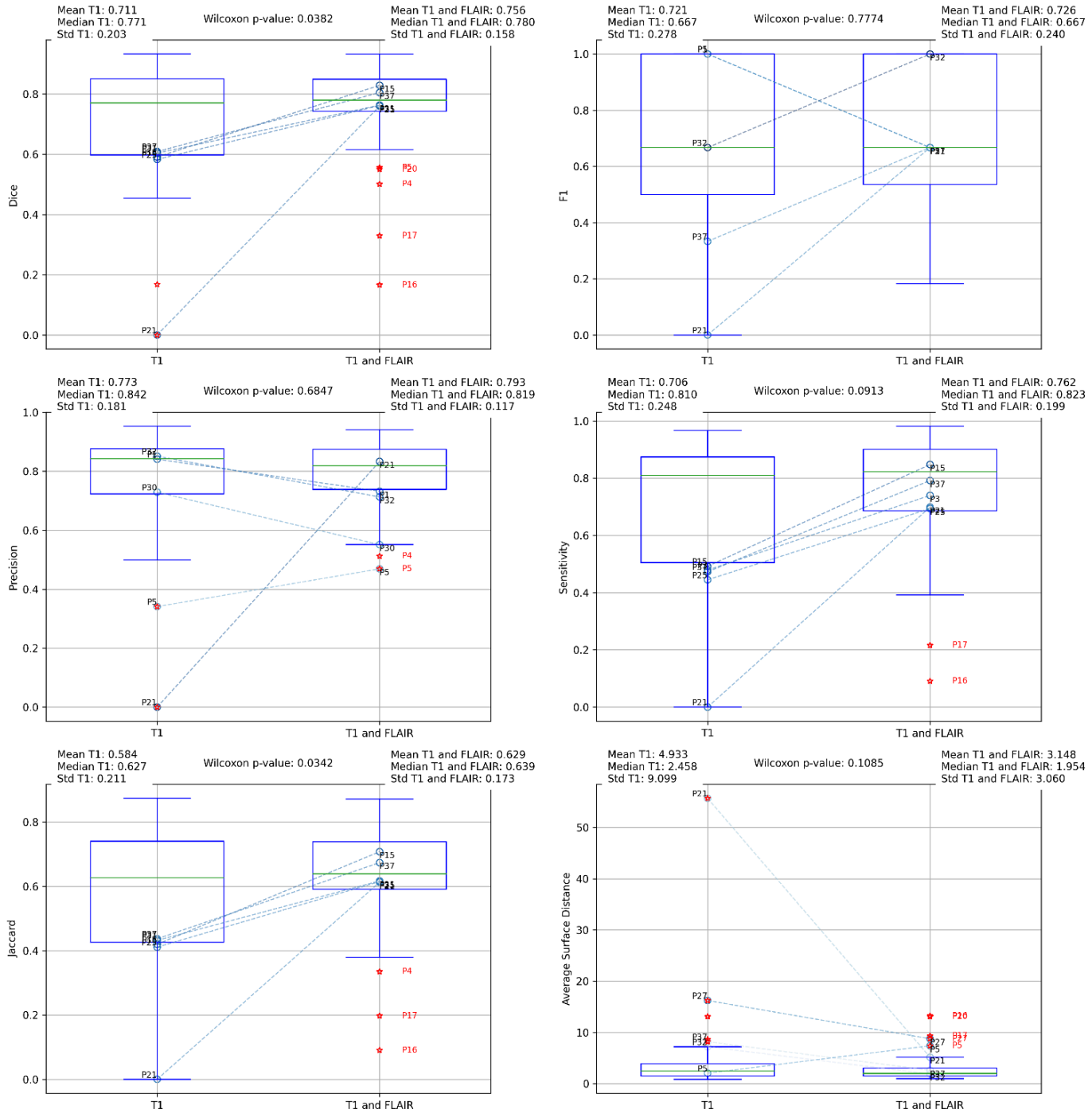

Figure 4: Sagittal (a), coronal (b), and axial (c) sections showing the original FLAIR image (top) and our segmentation output (bottom).

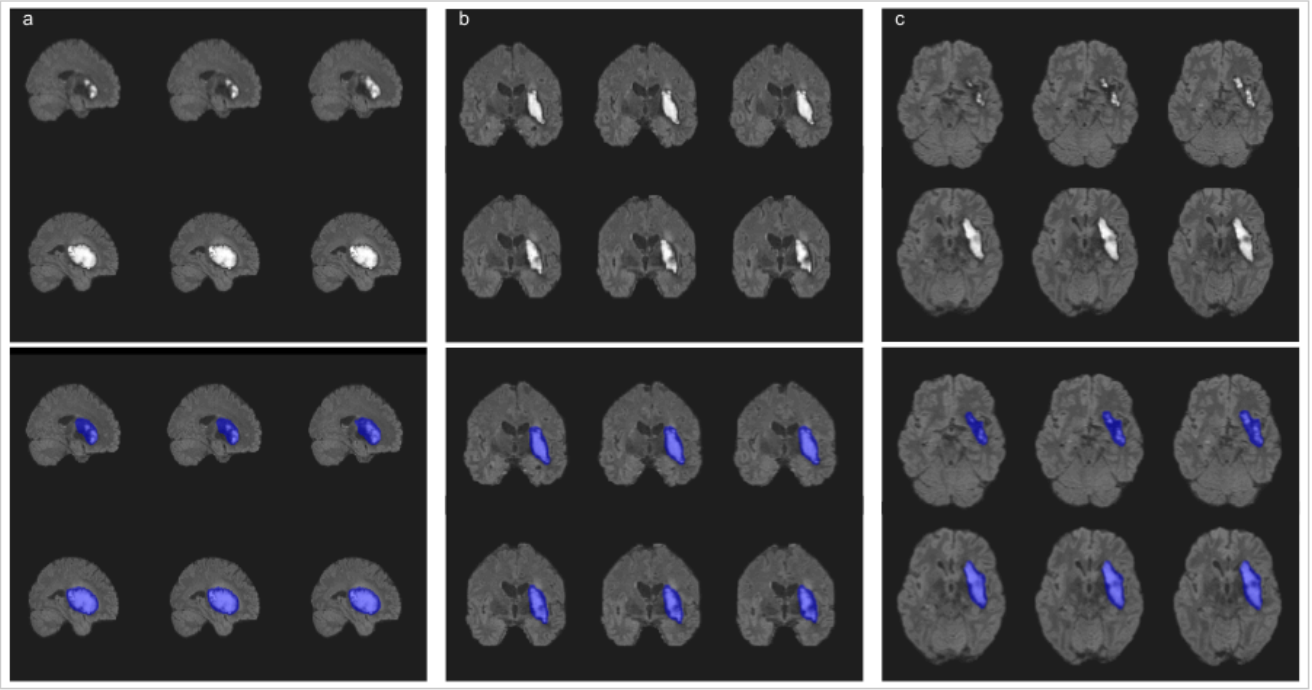

Figure 5: Comparison of lesion volumes between Ground Truth (GT) and model output, with regression line.