Semantic Segmentation of Spontaneous Intracerebral Hemorrhage, Intraventricular Hemorrhage and Associated Edema on CT Using Deep Learning

Manuscript type

AI in Brief

Summary

UNet-based networks accurately segment CT images of spontaneous intracerebral hemorrhage, with Focal loss function addressing intraventricular hemorrhage class imbalance.

Key Points

- A comparison of numerous deep learning networks for semantic segmentation of spontaneous intracerebral hemorrhage showed that UNet-based networks achieved significantly better performance than other network architectures for intracerebral hemorrhage and intraventricular hemorrhage segmentations ($p<.05$).
- Three-dimensional nnU-Net using the Focal loss function was able to address class imbalance in the dataset, providing significant performance improvement ($p<.05$) for segmentation of intraventricular hemorrhage present in approximately 30% of the training dataset (Dice score: 1.00 (IQR 0.87-1.00).

Abbreviations

ICH – intracerebral hemorrhage

PHE – perihematomal edema

IVH – intraventricular hemorrhage

DSC – Dice similarity coefficient

DiceCE – Dice and Cross-Entropy

Abstract

This study evaluated deep learning algorithms for semantic segmentation and quantification of intracerebral hemorrhage (ICH), perihematomal edema (PHE), and intraventricular hemorrhage (IVH) on non-contrast CT (NCCT) scans of patients with spontaneous ICH. Models were assessed on 1,732 annotated baseline NCCT scans obtained from the TICH-2 international multicenter trial (ISRCTN93732214), and different loss functions using three-dimensional nnU-Net were examined to address class imbalance (30% of participants with IVH in dataset). On the test cohort (*n*=174, 10% of dataset), the top-performing models achieved median Dice similarity coefficients of 0.92 (IQR, 0.89-0.94), 0.66 (0.58-0.71) and 1.00 (0.87-1.00), respectively for ICH, PHE and IVH segmentation. UNet-based networks showed comparable, satisfactory performances on ICH and PHE segmentations ($p>.05$), but all nnU-Net variants obtained higher accuracy than BLAST-CT and DeepLabv3+ for all labels ($p<.05$). The Focal model showed improved performance in IVH segmentation compared with Tversky, two-dimensional nnU-Net, UNet, BLAST-CT, and DeepLabv3+ ($p<.05$). Focal achieved concordance values of 0.98, 0.88, and 0.99 for ICH, PHE, and ICH volumes, respectively. The mean volumetric differences between ground truth and prediction were 0.32 mL (95% CI: −8.35 to 9.00), 1.14mL (-9.53 to 11.8) and 0.06mL (-1.71 to 1.84). In conclusion, UNet-based networks provide accurate segmentation on CT images of spontaneous ICH, and Focal loss can address class imbalance.

1 Introduction

Spontaneous intracerebral hemorrhage (ICH) is bleeding within the brain parenchyma in the absence of trauma or surgery, which may extend into the ventricles and subarachnoid space (1). Volumes of intracerebral haemorrhage (ICH), perihematomal edema (PHE) (2) and intraventricular hemorrhage (IVH) (3) are well-established biomarkers, consistent independent predictors of functional outcome and mortality of spontaneous ICH. Manual delineation and quantification of these biomarkers is labor

intensive and prone to human error. Thus, an efficient automated biomarker segmentation and quantification tool could provide quantitative outcome measures for clinical trials and accelerate studies in large cohorts of patients with spontaneous ICH.

Previous studies (4–10) have trained deep neural networks to perform ICH segmentation on CT scans, but most of these works were exclusively based on either or both ICH and PHE segmentation, as accurate delineation of IVH is challenging even for an experienced radiologist (11). Additionally, previous research in this area consists of single center studies with limited samples (12). Our work assesses the semantic segmentation and quantification of ICH, PHE, and IVH from a large multicenter dataset (from the Tranexamic acid for hyperacute primary intracerebral hemorrhage [TICH-2] trial, (13)).

This study compares the performance of existing deep learning approaches in the semantic segmentation and quantification of ICH, PHE and IVH. The best existing deep learning model was then refined by using different loss functions to address the class imbalance issue (unequal distribution of the lesion classes with little to no PHE or IVH pixels in a scan).

2 Materials and Methods

2.1 Study Patients

This retrospective analysis included baseline non-contrast CT (NCCT) scans from participants recruited to the prospective TICH-2 international randomized, placebo-controlled clinical trial (ISRCTN93732214) (13,14). The trial examined the effectiveness and safety of tranexamic acid in patients with acute spontaneous ICH within 8 hours of the onset of stroke symptoms. Ethical approval was granted from the UK Health Research Authority and the relevant national or local institutional review boards (non-UK sites), and written informed consent from patients or one of their relatives was obtained before enrollment. The full trial protocol is reported elsewhere (15). Our analysis included 1,732 eligible participants from the previously reported cohort (16) who had valid baseline scans (i.e. no incomplete or

missing slices). The previous work investigated radiomics-based features, whereas this study focuses on lesion segmentation using deep learning.

2.2 Image Acquisition and Ground Truth Delineation

The NCCT baseline scans were collected from 124 participating centers while complying with the local protocol. With a minimum requirement of axial image orientation, CT scans acquired from any scanner manufacturer, settings or slice thickness were included.

The anonymized ground truth segmentations of ICH, PHE and IVH were delineated on each scan by one of 3 independent trained raters (Z.K.L., vascular neurologist, with 15 years of experience; K.K., stroke physician, with 22 years of experience; A.A., CT radiographer, with 14 years of experience) using an active contour semi-automated segmentation algorithm on ITK-SNAP (version 3.6.0) (17), followed by manual editing if required. Additional inter- and intra-observer details are described in supplementary section 1.

The dataset was randomly split into a training cohort (n = 1558, 90%) (mean age, 69 ± [SD] 13 years; 872 men) and testing cohort (n = 174, 10%) (mean age, 68 ± 14 years; 102 men).

2.3 Deep Neural Network Selection for Comparison

We searched the best and most relevant neural networks for brain hemorrhage segmentation from Google Scholar and PubMed and shortlisted three approaches:

- nnU-Net (no-new-U-Net) (18): automated configuration method with state-of-the-art performance in many segmentation challenges including Medical Segmentation Decathlon, BraTS, and KiTS.
- BLAST-CT (brain lesion analysis and segmentation tool for CT) (19) pipeline based on DeepMedic: top performances in ISLES and BraTS
- DeepLabv3+ (20): ranks highly in the semantic segmentation of general objects and can outperform notable networks like FCN, SegNet and UNet in the segmentation of biomedical images

Finally, UNet, a widely used network for general medical image segmentation, is selected as baseline.

2.4 Network Implementation

We trained each model for 1800 epochs using the model pipeline default parameters and tested them on the independent test cohort. Both vanilla UNet and DeepLabv3+ models were implemented using the MONAI (Medical Open Network for Artificial Intelligence) framework (https://github.com/Project-MONAI/MONAI). We used the MONAI built-in three-dimensional (3D) BasicUNet and implemented the source code for the DeepLabv3+ model (https://github.com/janetkok/MONAI-DeepLabV3plus). The open-source frameworks, nnU-Net and BLAST-CT, can be found in (https://github.com/MIC-DKFZ/nnU-Net) and (https://github.com/biomedia-mira/blast-ct) respectively. General information and implementation details of these frameworks are described in supplementary sections 2 and 3 (Table S1), respectively.

2.5 Refinement of the Best Existing Model Through Loss Functions

We assumed that the default loss function in the best existing model—DiceCE (Dice and cross-entropy) (Table 2) would not be sufficiently sensitive to handle the extremely imbalanced target segmentation, low contrast, and heterogenous appearances of PHE and IVH lesions. Inspired by previous work (21), we evaluated Tversky, Focal, FocalTversky and DiceTopK loss using 3D nnU-Net to address the current model's limitations (code can be found on https://github.com/JunMa11/SegLoss.git (22)). These loss functions were selected based on their inherent capability to handle the class imbalance issue (see work by Ma et al (22) for full description of loss functions).

2.6 Performance Measures

Quantitative performance of the lesion volume was measured using automated-versus-human concordance and Bland–Altman plots. Accuracy overlay between ground truth and predicted lesion was quantified using the Dice similarity coefficient (DSC).

2.7 Statistical Analysis

Patient demographics were compared between the training and test samples using chi-square test or independent *t* test. The models' performances were statistically compared using the Kruskal-Wallis tests and corresponding Dunn post-hoc tests with false discovery rate correction. All statistical analyses were performed using RStudio (v1.4.1103), and $P < .05$ was considered significant.

3 Results

3.1 Patient Characteristics

Of the 1,732 included participants, we found no evidence of differences in characteristics between the training and test sets (Table 1).

3.2 Lesion Segmentation Performance

Note that all 3D nnU-Net loss function variants will be represented as the name of their loss functions. See supplementary section 4 for naming convention details.

Tables 2 shows the DSC of various models in our experiment. Box-violin plots showing the DSC distribution of various models are presented in Figure S1. Figures S2 and S3 present a qualitative overview of the segmentation (best and worst-case segmentation with respect to DiceCE) for the top-performing models: DiceCE, DiceTopK, Focal.

UNet-based networks achieved similarly good performance in ICH and PHE segmentations based on the average and median schemes, showing no evidence of differences between them ($p>.05$; Table 2). Compared with the lowest performers—BLAST-CT and DeepLabv3+, all nnU-Net variants had higher DSC for all lesion segmentation ($p<.05$). The segmentation quality of ICH was satisfactory across all models (Figure S2), whereas PHE segmentation was not desirable as the boundaries of the PHE label appeared to be smoothed (Figure S2B), indicating a lack of precision.

The nnU-Net variations, DiceCE, two-dimensional (2D)+3D, DiceTopK and Focal showed significant performance improvement in IVH segmentation compared with Tversky, 2D, UNet, BLAST-CT and DeepLabv3+ ($p<.05$). Given that the top-performing models for IVH segmentation, particularly Focal, DiceTopK and FocalTversky, are mainly designed with emphasis on hard samples to address high class imbalance, there was no evidence of statistical differences between the aforementioned models. Still, Focal had a notably higher average DSC and more consistent performance based on the small interquartile range in IVH segmentation (Table 2). Also, based on the qualitative performance shown in Figures S2A,B,C and S3C (note the lack of blue labels when using DiceCE and DiceTopK), Focal presented greater detection capability for small and low contrast IVH lesion compared with other top performing models such as DiceCE and DiceTopK. Furthermore, we observed that the DSC values of most low-performing networks in IVH segmentation have a bimodal distribution (Figure S1C). This suggests that their segmentation results are polarized, presumably caused by differences in lesion size and intensity.

3.3 Volume Quantification and Agreement

Figure 1 shows the concordance and Bland-Altman plot of agreements between the ground truth and predicted lesion volumes by Focal, the best overall network (based on the global mean of DSCs). Focal demonstrated high concordance and low mean difference in estimating all lesions, except PHE. Additional details are presented in supplementary section 7.

4 Discussion

We compared the segmentation performance of existing deep learning networks on a large dataset from the TICH-2 trial. The analysis demonstrated that UNet-based networks have immense potential in segmenting targeted lesions. We investigated how the application of a selected range of loss functions

could be a feasible technique to address the issue of class imbalance. We showed that Focal can address this problem and significantly improve IVH segmentation.

Based on the median DSC of ICH, 3D nnU-Net variant performance was similar to that found in the single center study by Zhao et al (12) and multicenter study by Sharrock et al. (23). Our reported average DSC for the best ICH segmentation was lower than the previously published Ψ-Net (0.95) (24), but it should be noted that this prior work had lower CT variability and a smaller dataset.

With regards to PHE segmentation, studies by Ironside et al (12) and Zhao et al (12) reported higher performances than our results. Nevertheless, our performances remain adequate considering the variability of our multicenter dataset and smaller PHE volume (larger lesion volumes are positively correlated with higher DSC (12,19)).

The best model for IVH segmentation (Focal) outperformed that in the single center study by Zhao et al (12). Of note, our initial assumption that PHE segmentation performance would be superior to that of IVH did not hold, albeit PHE was much larger in volume and was found in 99% of the training dataset. A possible explanation is that PHE demands resolution of indistinct low-contrast edges, whereas IVH typically shows as areas of high attenuation and often has sharp edges where the IVH contacts the ventricle wall.

This study had limitations. First, a scan not containing a lesion class that it was mislabelled by the model had a DSC of 0. In the same setting, but with no error made, the DSC was counted as 1 to include participant scans with no target lesion that were correctly predicted. Thus, the metric tends to favor non-existent lesions and can result in misrepresentation of model performance. To address this issue, we included two other metrics, volume intersection and false positives, in Table S2 for better performance comparison. Additionally, we are aware that we only performed a single split, and the performance evaluation could have been strengthened with k-fold cross validation. Also, we acknowledge that our ground truth segmentations contained segmentation errors in a few cases, as supported by the high but imperfect rates of inter- and intra-rater agreement. Inspection of the ‘worst-case’ segmentation in Figure

S3 shows that, in retrospect, network-based segmentations were more accurate than the manually edited semi-automated segmentations. That being said, the study would be enhanced if a radiologist was among the reviewers.

In summary, we compared numerous deep learning approaches for the segmentation and quantification of ICH, PHE and IVH in spontaneous ICH from a large scale, international multicenter dataset. We showed that UNet-based networks remain robust in medical imaging segmentation, demonstrating similarly high performances for both ICH and PHE lesions. We also investigated a selected range of loss functions on the 3D nnU-Net, but none of the networks had the best result in every lesion; however, Focal can address class imbalance and showed greater detection capability with significant performance improvement in IVH segmentation, a prominent yet rarely investigated lesion due to its complexity and scarcity. We believe that future development of a fully accurate and automated deep learning-based segmentation model can potentially eliminate human error in manual segmentation and provide early prediction of hematoma expansion and clinical outcome when combined with quantitative radiomic analysis (16).

Data Sharing Statement

Data generated or analyzed during the study are available upon written request to the TICH-2 trial CI Prof Nikola Sprigg (Nikola.Sprigg@nottingham.ac.uk). Proposals will be assessed by the CI (with advice from the TICH-2 trial Steering Committee if required) and a Data Transfer Agreement will be established before any data are shared.

Acknowledgements

We thank all the participants, staff and centres who took part in the TICH-2 trial. The TICH-2 trial was funded by a grant from the UK National Institute for Health Research Health Technology Assessment programme (project code 11_129_109).

References

1. Aguilar MI, Freeman WD. Spontaneous intracerebral hemorrhage. Semin Neurol. 2010;30(5):555–564. doi: 10.1055/s-0030-1268865.

2. Appelboom G, Bruce SS, Hickman ZL, et al. Volume-dependent effect of perihaematomal oedema on outcome for spontaneous intracerebral haemorrhages. J Neurol Neurosurg Psychiatry. BMJ Publishing Group Ltd; 2013;84(5):488–493. doi: 10.1136/jnnp-2012-303160.

3. Trifan G, Arshi B, Testai FD. Intraventricular Hemorrhage Severity as a Predictor of Outcome in Intracerebral Hemorrhage. Front Neurol. 2019;10. doi: 10.3389/fneur.2019.00217.

4. Dhar Rajat, Falcone Guido J., Chen Yasheng, et al. Deep Learning for Automated Measurement of Hemorrhage and Perihematomal Edema in Supratentorial Intracerebral Hemorrhage. Stroke. American Heart Association; 2020;51(2):648–651. doi: 10.1161/STROKEAHA.119.027657.

5. Islam M, Sanghani P, See AAQ, James ML, King NKK, Ren H. ICHNet: Intracerebral Hemorrhage (ICH) Segmentation Using Deep Learning. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer International Publishing; 2019. p. 456–463. doi: 10.1007/978-3-030-11723-8_46.

6. Chang PD, Kuoy E, Grinband J, et al. Hybrid 3D/2D Convolutional Neural Network for Hemorrhage Evaluation on Head CT. American Journal of Neuroradiology. American Journal of Neuroradiology; 2018;39(9):1609–1616. doi: 10.3174/ajnr.A5742.

7. Guo D, Wei H, Zhao P, et al. Simultaneous Classification and Segmentation of Intracranial Hemorrhage Using a Fully Convolutional Neural Network. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020. p. 118–121. doi: 10.1109/ISBI45749.2020.9098596.

8. Hu K, Chen K, He X, et al. Automatic segmentation of intracerebral hemorrhage in CT images using encoder–decoder convolutional neural network. Information Processing & Management. 2020;57(6):102352. doi: 10.1016/j.ipm.2020.102352.

9. Ironside N, Chen C-J, Mutasa S, et al. Fully Automated Segmentation Algorithm for Hematoma Volumetric Analysis in Spontaneous Intracerebral Hemorrhage. Stroke. 2019;50(12):3416–3423. doi: 10.1161/STROKEAHA.119.026561.

10. Ironside N, Chen C-J, Mutasa S, et al. Fully Automated Segmentation Algorithm for Perihematomal Edema Volumetry After Spontaneous Intracerebral Hemorrhage. Stroke. 2020;51(3):815–823. doi: 10.1161/STROKEAHA.119.026764.

11. Dowlatshahi D, Kosior JC, Idris S, et al. Planimetric hematoma measurement in patients with intraventricular hemorrhage: is total volume a preferred target for reliable analysis? Stroke. Am Heart Assoc; 2012;43(7):1961–1963. doi: 10.1161/STROKEAHA.113.003387.

12. Zhao X, Chen K, Wu G, et al. Deep learning shows good reliability for automatic segmentation and volume measurement of brain hemorrhage, intraventricular extension, and peripheral edema. Eur Radiol. 2021; doi: 10.1007/s00330-020-07558-2.

13. Sprigg N, Flaherty K, Appleton JP, et al. Tranexamic acid for hyperacute primary IntraCerebral Haemorrhage (TICH-2): an international randomised, placebo-controlled, phase 3 superiority trial. The Lancet. Elsevier; 2018;0(0). doi: 10.1016/S0140-6736(18)31033-X.

14. Sprigg N, Flaherty K, Appleton JP, et al. Tranexamic acid to improve functional status in adults with spontaneous intracerebral haemorrhage: the TICH-2 RCT. Health Technol Assess. 2019;23(35):1–48. doi: 10.3310/hta23350.

15. Sprigg N, Robson K, Bath P, et al. Intravenous tranexamic acid for hyperacute primary intracerebral hemorrhage: Protocol for a randomized, placebo-controlled trial. International Journal of Stroke. SAGE Publications; 2016;11(6):683–694. doi: 10.1177/1747493016641960.

16. Pszczolkowski S, Manzano-Patrón JP, Law ZK, et al. Quantitative CT radiomics-based models for prediction of haematoma expansion and poor functional outcome in primary intracerebral haemorrhage. Eur Radiol. 2021; doi: 10.1007/s00330-021-07826-9.

17. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage. 2006;31(3):1116–1128. doi: 10.1016/j.neuroimage.2006.01.015.

18. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–211. doi: 10.1038/s41592-020-01008-z.

19. Monteiro M, Newcombe VFJ, Mathieu F, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. The Lancet Digital Health. Elsevier; 2020;2(6):e314–e322. doi: 10.1016/S2589-7500(20)30085-6.

20. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv:180202611 [preprint] http://arxiv.org/abs/1802.02611. Posted August 22, 2018. Accessed November 29, 2020.

21. Ma J. Loss Ensembles for Extremely Imbalanced Segmentation. arXiv:210110815 [preprint] http://arxiv.org/abs/2101.10815. Posted December 31, 2020. Accessed March 7, 2021.

22. Ma J, Chen J, Ng M, et al. Loss odyssey in medical image segmentation. Medical Image Analysis. 2021;71:102035. doi: 10.1016/j.media.2021.102035.

23. Sharrock MF, Mould WA, Ali H, et al. 3D Deep Neural Network Segmentation of Intracerebral Hemorrhage: Development and Validation for Clinical Trials. Neuroinform. 2020; doi: 10.1007/s12021-020-09493-5.

24. Kuang Z, Deng X, Yu L, Wang H, Li T, Wang S. Ψ-Net: Focusing on the border areas of intracerebral hemorrhage on CT images. Computer Methods and Programs in Biomedicine. 2020;194:105546. doi: 10.1016/j.cmpb.2020.105546.

25. McBride GB. A proposal for strength of agreement criteria for Lin's concordance correlation coefficient. NIWA Client Report HAM2005-062. Hamilton, New Zealand: National Institute of Water & Atmospheric Research Ltd, 2005.

Figure Legend

Figure 1: (A) Bland-Altman and (B) concordance plot of agreements between ground truth and predicted lesion volumes in the test cohort by the Focal model. CCC = concordance correlation coefficient, ICH = intracerebral hemorrhage, PHE = perihematomal edema, IVH = intraventricular hemorrhage

Tables

Table 1:Characteristics for Training and Test Cohorts

Characteristic	Training (n=1558)	Test (n=174)	p value
Age, years*	69±14 (20-101)	68±13 (35-92)	.25
Sex†			.56
Men	872 (56)	102 (59)	
Women	686 (44)	72 (41)	
Onset to CT, h ‡	1.9 (1.4-2.9)	1.8 (1.3-2.9)	.58
ICH			.91
Count†	1558 (100)	174 (100)	
Volume, mL ‡	11.99 (5.32-27.84)	12.64 (5.19-27.58)	
Volume, mL*	20.95±23.47 (0.50-158.64)	19.43±21.40 (0.50-128.55)	
PHE			.93
Count†	1542 (99)	172 (99)	
Volume, mL ‡	7.15 (3.53-13.63)	6.73 (3.41-14.37)	
Volume, mL*	11.62±14.43 (1.19×10^{-4}-152.10)	10.94±11.58 (0.51-61.10)	
IVH			.82
Count†	472 (30)	48 (28)	
Volume, mL ‡	5.63 (1.88-13.38)	5.84 (1.90-15.09)	
Volume, mL*	9.47±10.92 (7.73×10^{-3}-77.49)	9.45±9.28 (0.27-33.57)	

Note.— ICH = intracerebral haemorrhage, PHE = perihematomal edema, IVH = intraventricular hemorrhage

** Data presented as mean ± SD (range)*

† Data presented as number of participants (percentage)

‡ Data presented as median (IQR)

Table 2: Dice Score Performances of Existing Models and 3D nnU-Net Loss Function Variants

	Average Dice score				Median Dice score (IQR)			
	ICH	PHE	IVH	Mean	ICH	PHE	IVH	Mean
BLAST-CT	0.850†	0.567	0.407	0.608	0.891 (0.846-0.922)	0.602 (0.494-0.659)	0.007 (0.000-0.891)	0.500
DeepLabv3+	0.857	0.522†	0.366†	0.582†	0.888 (0.844-0.912)†	0.553 (0.438-0.628)†	0.000 (0.000-0.814)†	0.480†
UNet	0.891	0.602	0.701	0.731	0.913 (0.875-0.933)	0.625 (0.541-0.689)	1.000 (0.411-1.000)	0.846
Default nnU-Net variants								
2D	0.894	0.610	0.614	0.706	0.911 (0.881-0.934)	0.633 (0.539-0.710)	0.851 (0.000-0.000)	0.798
3D/DiceCE	0.904	0.627*	0.811	0.781	0.916 (0.887-0.935)	0.657 (0.578-0.710)*	1.000 (0.826-1.000)	0.858*
2D + 3D	0.892	0.618	0.794	0.768	0.914 (0.884-0.935)	0.647 (0.567-0.715)	1.000 (0.798-1.000)	0.854
3D nnU-Net loss function variants								
Tversky	0.894	0.608	0.659	0.720	0.913 (0.885-0.932)	0.633 (0.536-0.704)	0.853 (0.000-1.000)	0.799
DiceTopK	0.905*	0.626	0.846	0.792	0.916 (0.888-0.936)*	0.651 (0.578-0.708)	1.000 (0.849-1.000)	0.856
FocalTversky	0.900	0.608	0.783	0.764	0.912 (0.883-0.931)	0.632 (0.543-0.711)	1.000 (0.777-1.000)	0.848
Focal	0.904	0.612	0.885*	0.800*	0.915 (0.888-0.935)	0.639 (0.550-0.705)	1.000 (0.867-1.000)*	0.851

Note.— 2D = two-dimensional, 3D = three-dimensional, BLAST-CT = brain lesion analysis and segmentation tool for CT, DiceCE = Dice and Cross-Entropy loss, ICH = intracerebral hemorrhage, IVH = intraventricular hemorrhage, nnU-Net = no-new-U-Net, PHE = perihematomal edema

+ indicates an ensemble between models

** Best performance*

† Worst performance

Supplementary Material

1 Inter- and intra-observer agreement for ground truth segmentations

Inter- and intra-observer agreement for intracerebral hemorrhage (ICH), intraventricular hemorrhage (IVH) and perihematomal edema (PHE) measurement was assessed in a subset of 20 scans measured independently by two of the raters on two occasions separated by at least one week using type A intraclass correlation coefficients using an absolute agreement definition. High intraclass correlation coefficients were achieved for ICH volumes for both inter-observer (ICC=0.98, 95% confidence intervals (CI) 0.92, 0.99) and intra-observer agreement (ICC=0.99, 95%CI 0.99, 1.00 for both raters), for IVH volumes for both inter-observer (ICC=0.98, 95%CI 0.93, 0.99) and intra-observer agreement (ICCs 0.99, 95%CI 0.98, 1.00, and 0.98, 95%CI 0.94-0.99), and for PHE volumes for both inter-observer (ICC=0.99, 95%CI 0.96, 1.00) and intra-observer agreement (ICCs 0.99, 95%CI 0.99-1.00, and 0.83, 95%CI 0.58, 0.93).

2 General information of the selected models for comparison

2.1 nnU-Net

The nnU-Net (18) is a self-adapting framework on the basis of vanilla UNet. It uses a set of heuristic rules to infer the data-dependent hyperparameters based on the dataset properties. The two-dimensional (2D) UNet and three-dimensional (3D) full resolution UNet configurations and the ensemble of these configurations were evaluated in this study.

2.2 BLAST-CT

The BLAST-CT (19) pipeline is constructed based on the DeepMedic model. To account for the CT modality from the CENTER-TBI (Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury) dataset, Miguel et al. used intensity windowing to replace skull-stripping as the latter is susceptible to failure. Instead of using a dual pathway architecture, Miguel et al. employed a 3 parallel pathways network that handles image patches at full, three-times and five-times downsampled

resolution. Furthermore, the network utilises residual connections and pre-activation blocks to keep pace with the current advanced approaches.

2.3 DeepLabV3+

Chen et al. introduced DeepLabV3 which integrates the benefits of both dilated convolutions and feature pyramid pooling. Using the Xception model as a backbone, DeepLabV3+ (20) extends the network by adding a decoder module to fine-tune the segmentation results, specifically along the object boundaries.

2.4 UNET

UNet comprises of a encoder and decoder part that are linked up through skip connections. The encoder encodes the high-resolution features by gradually reducing the spatial information and then the decoder is used to enable precise localisation using upsampling convolutions and the help of skip connections.

3 Implementation details of selected model pipelines used for comparison

All experiments were implemented using Python and PyTorch on a Linux workstation with NVIDIA GeForce RTX 2080 Ti.

Table S1 Implementation Details of Selected Model Pipelines Used for Comparison: UNet, DeepLabv3+, BLAST-CT and nnU-Net

Model Framework	MONAI	MONAI	BLAST-CT	nnU-Net
Network	3D BasicUNet	3D DeepLabv3+ (Modified Aligned Xception as backbone)	DeepMedic	2D UNet, 3D UNet
Training type	Patch-wise training at full resolution data		3 path architecture patches at full, three-times and five-times downsampled resolution	Patch-wise training at full resolution data
Loss function	Dice and cross-entropy		Cross-entropy	Dice and cross-entropy
Optimizer	Adam		RMSprop	SGD with Nesterov momentum
Data augmentation	-		Elastic Deformation Coarse Perlin Noise, Histogram Deformation, Patch Rotation, Patch Flip	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring
Inference	Sliding window overlapping patches, Gaussian patch centre weighting.		Sliding window with overlapping patches.	Sliding window with overlapping patches, Gaussian patch centre weighting.
Intensity normalisation	Intensity normalization based on mean and standard deviation calculated on each channel separately.		Bounded the intensities between -15 and 100 Hounsfield units (HU) before scaling the range between -1 and 1.	Clipping images to 0.5 and 99.5 percentiles, followed by subtraction of the global mean and division by the global standard deviation.
Resampling strategy	Resample images to an isotropic resolution of 1×1×1 mm.			Median spacing is computed independently for each axis. Then, resampling with third-order spline(data) and linear interpolation (annotation).
Postprocessing	-		3D conditional random field	Opt for non-largest component suppression if achieve performance gain by first treating all foreground class as one component and then reiterate for individual classes.
Ensemble	N/A			Ensemble of 2D and 3D UNet

Note.— MONAI = Medical Open Network for Artificial Intelligence, 2D = two-dimensional, 3D = three-dimensional, BLAST-CT = brain lesion analysis and segmentation tool for CT, nnU-Net = no-new-U-Net

4 Naming convention for 3D nnU-Net loss function variants

3D nnU-Net loss function variants will be represented as the name of their loss functions:

DiceCE: 3D nnU-Net using Dice and Cross-Entropy as loss function

Tversky: 3D nnU-Net using Tversky as loss function

DiceTopK: 3D nnU-Net using DiceTopK as loss function

FocalTversky: 3D nnU-Net using FocalTversky as loss function

Focal: 3D nnU-Net using Focal as loss function

5 Quantitative performance

Figure S1: Box-violin plots of Dice Similarity Coefficient of various models for ICH, PHE and IVH, sorted by mean score. DSC = Dice Similarity Coefficient; ICH = intracerebral hemorrhage; PHE = perihematomal oedema; IVH = intraventricular hemorrhage; DiceCE = Dice and Cross-Entropy loss;

6 Qualitative performances of best networks

Figure S2: Best segmentation results of networks (DiceCE, DiceTopK, Focal) with respect to DiceCE. ICH is shown in red, PHE in green and IVH in blue. Dice scores of each model are shown on top of each image. DiceCE = Dice and Cross-Entropy; ICH = intracerebral hemorrhage; PHE = perihematomal edema; IVH = intraventricular hemorrhage.

Figure S3: Worst segmentation results of networks (DiceCE, DiceTopK, Focal) with respect to DiceCE. ICH is shown in red, PHE in green and IVH in blue. Dice scores of each model are shown on top of each image. DiceCE = Dice and Cross-Entropy; ICH = intracerebral hemorrhage; PHE = perihematomal edema; IVH = intraventricular hemorrhage.

7 Volume quantification and agreement

We assessed the concordance correlation coefficient and Bland–Altman plot of agreements between the ground truth and predicted lesion volumes. For reference, McBride (25) suggested that concordance correlation coefficient less than 0.90 is poor, 0.90-0.95 is moderate, 0.95-0.99 is substantial and greater than 0.99 is excellent.

Based on Figure 1, Focal exemplified superior performance in estimating both ICH and IVH volumes with 'substantial' concordance of 0.98 and 0.99 respectively and mean differences of 0.32 mL (95% CI −8.35 to 9.00) and 0.06mL (95% CI -1.71 to 1.84). Conversely, Focal showed poorer performance in estimating the PHE volume: its mean difference was 1.14mL (95% CI -9.53 to 11.8) and concordance was 'poor' (0.88).

8 Intersection between ground truth and predicted volumes, False positives in the predicted volume

Based on Table 2 , the baseline UNet model remains as a robust architecture for brain segmentation as all nnU-Net variants showed no significant performance improvement ($p>0.05$) in ICH and PHE segmentations. However, UNet tends to have more false positives (Table S2) for IVH segmentation compared to DiceCE (Dice and cross-entropy) and 2D+3D, hence it has significantly poorer performance than the latter for IVH ($p<0.05$, Table 2). As a side note, Focal was able to outperform DiceCE (the best default nnU-Net variant for IVH segmentation) although it tends to have slightly more false positives because Focal can correctly predict the IVH volumes for most of the time compared to DiceCE.

DeepLabv3+ and BLAST-CT are the worst performers. They had significantly lower DSC for all segmentations due to their low mean intersection volume as depicted in Table S2. Also, we noticed that all networks were likely to have more false positive results when predicting PHE (Tables S2) as opposed to ICH and IVH because the PHE is typically low in contrast and has obscured boundaries.

Table S2: Mean of the Intersection between Ground Truth and Predicted Volume and Mean of the False Positives in the Predicted Volume.

	Intersection between Ground Truth and Predicted Volume (mL)				False Positives in Predicted Volume (mL)			
	ICH	PHE	IVH	Mean	ICH	PHE	IVH	Mean
BLAST-CT	17.313	6.569	1.900†	8.594	3.412†	4.317	0.641	2.790
DeepLabv3+	17.090†	5.077†	1.995	8.054†	2.464	2.570	0.785†	1.940
UNet	17.234	7.166	2.059	8.820	1.645	4.533†	0.554	2.244
Default nnU-Net variants								
2D	17.125	7.374	2.158	8.885	1.406	4.275	0.599	2.093
3D/DiceCE	17.357	7.519	1.994	8.956	1.513	4.252	0.418	2.061
(2D + 3D)	17.228	7.392	1.985	8.869	1.400*	4.079	0.402*	1.960
3D nnU-Net loss function variants								
Tversky	17.523	8.547*	2.308*	9.459*	1.936	6.973	0.965	3.291†
DiceTopK	17.452	7.152	1.999	8.868	1.560	3.545	0.408	1.838
FocalTversky	17.603*	8.539	2.107	9.417	1.981	6.952	0.669	3.201
Focal	17.485	6.662	2.116	8.754	1.622	3.023*	0.428	1.691*

Note.— A good model is a model with a balanced high intersection and low false positive volume. 2D = two-dimensional, 3D = three-dimensional, BLAST-CT = brain lesion analysis and segmentation tool for CT, DiceCE = Dice and Cross-Entropy loss, ICH = intracerebral hemorrhage, IVH = intraventricular hemorrhage, nnU-Net = no-new-U-Net, PHE = perihematomal edema

+ indicates an ensemble between models

** Best performance*

† Worst performance

Figure S1: Box-violin plots of Dice Similarity Coefficient of various models for ICH, PHE and IVH, sorted by mean score. DSC = Dice Similarity Coefficient; ICH = intracerebral hemorrhage; PHE = perihematomal oedema; IVH = intraventricular hemorrhage; DiceCE = Dice and Cross-Entropy loss;

Figure S2: Best segmentation results of networks (DiceCE, DiceTopK, Focal) with respect to DiceCE. ICH is shown in red, PHE in green and IVH in blue. Dice scores of each model are shown on top of each image. DiceCE = Dice and Cross-Entropy; ICH = intracerebral hemorrhage; PHE = perihematomal edema; IVH = intraventricular hemorrhage.

Figure S3: Worst segmentation results of networks (DiceCE, DiceTopK, Focal) with respect to DiceCE. ICH is shown in red, PHE in green and IVH in blue. Dice scores of each model are shown on top of each image. DiceCE = Dice and Cross-Entropy; ICH = intracerebral hemorrhage; PHE = perihematomal edema; IVH = intraventricular hemorrhage.