BHSD: A 3D Multi-Class Brain Hemorrhage Segmentation Dataset

Biao Wu[1], Yutong Xie[1], Zeyu Zhang[1,3], Jinchao Ge[1], Kaspar Yaxley[2], Suzan Bahadir[2], Qi Wu[1], Yifan Liu[1,4], and Minh-Son To[2]

[1] Australia Institute of Machine Learning, University of Adelaide
[2] Flinders Health and Medical Research Institute, Flinders University
[3] Australian National University
[4] ETH Zürich

Abstract. Intracranial hemorrhage (ICH) is a pathological condition characterized by bleeding inside the skull or brain, which can be attributed to various factors. Identifying, localizing and quantifying ICH has important clinical implications, in a bleed-dependent manner. While deep learning techniques are widely used in medical image segmentation and have been applied to the ICH segmentation task, existing public ICH datasets do not support the multi-class segmentation problem. To address this, we develop the Brain Hemorrhage Segmentation Dataset (BHSD), which provides a 3D multi-class ICH dataset containing 192 volumes with pixel-level annotations and 2200 volumes with slice-level annotations across five categories of ICH. To demonstrate the utility of the dataset, we formulate a series of supervised and semi-supervised ICH segmentation tasks. We provide experimental results with state-of-the-art models as reference benchmarks for further model developments and evaluations on this dataset. The dataset and checkpoint is available at https://github.com/White65534/BHSD.

Keywords: Intracranial hemorrhage · Segmentation · Multi-class.

1 Introduction

Intracranial hemorrhage (ICH) refers to bleeding that occurs inside the skull or brain. There are various types, based on the anatomical relation of the bleeding with the brain and its surrounding membranes. These types include extradural hemorrhage (EDH), subdural hemorrhage (SDH), subarachnoid hemorrhage (SAH), intraparenchymal hemorrhage (IPH), and intraventricular hemorrhage (IVH). The causes of ICH are diverse [11,7,18], encompassing factors such as trauma, vascular malformations, tumors, hypertension, and venous thrombosis. The detection and characterization of ICH are typically done using non-contrast CT scans, which can reveal the type and distribution of bleeding. Accurate localization, quantification, and classification of hemorrhages are crucial as they have significant clinical implications, allowing clinicians to estimate severity, predict outcomes, and monitor progress [20,10,1,6]. Treatment options may also vary depending on the quantity and type of bleed.

(a) Comparison with other datasets.

(b) Example annotations.

Fig. 1: Overview of the BHSD. (a) Comparison of the BCIHM dataset, IN-STANCE dataset and BHSD in terms of data characteristics. The BHSD provides more data, and more annotation information at different levels. (b) Representative examples of the diverse hemorrhage annotations provided in the BHSD. The different colors correspond to different hemorrhage classes, as indicated by the legend.

The rapid development of deep learning and large-scale labeled dataset has accelerated the automation of medical image segmentation [25,15]. Despite the importance of delineating different classes of hemorrhage, many existing public ICH datasets focus either on hemorrhage classification [12] or single-class segmentation [16], specifically foreground versus background. Thus, it is essential to develop tools that can accurately segment different classes of hemorrhage rather than solely perform foreground versus background segmentation. Unfortunately, the lack of public datasets with multi-class, pixel-level annotations hinders the development of class-specific segmentation techniques.

The purpose of this work is to augment a large, public ICH dataset[5] to produce a 3D, multi-class ICH dataset with pixel-level hemorrhage annotations, hereafter referred to as the brain hemorrhage segmentation dataset (BHSD). Our approach leverages the existing high-quality slice-level annotations performed by neuroradiologists and subsequently relabels a subset of CT scans with multi-class pixel-level annotations. We demonstrate the utility of this dataset by performing a series of experiments and providing benchmarks on supervised and semi-supervised segmentation tasks.

2   Multi-Class Brain Hemorrhage Segmentation Dataset

2.1   Brain hemorrhage datasets

In this section, we describe existing, public brain hemorrhage datasets.

The BCIHM dataset consists of 82 non-contrast CT scans of patients with traumatic brain injury [12]. The dataset is provided in NIfTI format. Each slice of the scans was reviewed by two radiologists who recorded hemorrhage types if hemorrhage occurred or if a fracture occurred. Hemorrhage regions in each slice were also segmented, however, the annotations only support a foreground versus background segmentation. In contrast, the INSTANCE dataset, introduced as part of a MICCAI 2022 Challenge, consists of 200 volumes, of which 130 have foreground and background segmentation labels [16]. The volumes in this dataset are also provided in NIfTI format, but again all bleed types are combined into a single foreground class.

The CQ500 dataset consists of 491 CT scans with 193,317 slices in DICOM format [3]. The scans have been read by three radiologists, and the annotations provided indicate, at the scan level, the presence, type and location of hemorrhage. While this dataset does not support segmentation tasks, an augmentation of this dataset, BHX, provides bounding box annotations for five types of hemorrhage [19]. Specifically, BHX contains 39,668 bounding boxes in 23,409 images. Another key brain hemorrhage dataset was published by the Radiological Society of North America (RSNA) [5]. This dataset is a public collection of 874,035 CT head images in DICOM format from a mixed patient cohort with and without ICH. The dataset is multi-institutional and multi-national and includes slice-level expert annotations from neuroradiologists about the presence and type of bleed. This dataset was used for the RSNA 2019 Machine Learning Challenge for detecting brain hemorrhages, *i.e.,* a classification, not segmentation problem.

2.2   Reconstruction and annotation pipeline of the BHSD

The BHSD is a high-quality medical imaging dataset comprising 2192 high-resolution 3D CT scans of the brain, each containing between 24 to 40 slices of 512×512 pixels in size (Fig. 1a). The original RSNA dataset was provided as a collection of randomly sorted slices in DICOM format with slice-level annotations. Important contextual information in adjacent slices may be lost in single slices, hence the first step was to reconstitute 3D head scans. Since the anonymized patient identifiers were provided and the DICOM files retained geometric/positional data, the original 3D head scans could be reconstructed and converted to NIfTI format [14]. The slice-level hemorrhage labels provided with the original RSNA dataset were mapped to the corresponding slices in the reconstructed head scans. A subset of 192 scans with one or more of five categories of ICH, namely EDH, IPH, IVH, SAH, and/or SDH, was subsequently selected for further annotation.

Pixel-level annotations were performed by three medical imaging experts in two stages. Hemorrhages on individual head scans were independently segmented using ITK-SNAP [27] by two trained medical imaging experts and radiology residents, both with over one year of experience reading CT head scans, using the original image-level hemorrhage annotations as a guide. These annotations were then reviewed by a board-certified radiologist with over 5 years post-fellowship

(a) Number of bleeds (slice level)

(b) Number of bleeds (scan level)

(c) Type of bleed (slice level)

(d) Type of bleed (scan level)

Fig. 2: Summary composition of the BHSD, at the slice and scan levels, by number and type of bleed.

experience, ensuring the quality of the annotations. To supplement the 192 volumes with pixel-level annotations, we also collected corresponding image-level annotations from the RSNA dataset and provide an additional 2000 3D CT scans with slice-level annotations, including scans with no bleed. The composition of the BHSD is shown in Fig. 2.

By covering both image-level and pixel-level annotations, the BHSD allows a more comprehensive interrogation of brain hemorrhage imaging, and as we show, enables the development of more varied deep learning methods for ICH segmentation.

2.3   Segmentation applications using the BHSD

In this section, we describe two segmentation applications designed to leverage the proposed BHSD.

Supervised segmentation. Supervised multi-class segmentation refers to the classification of all individual pixels in an image into distinct classes, using segmentation mask annotations for supervision. For ICH, multi-class segmentation is clinically more significant while technically more challenging compared with foreground and background segmentation. Multi-class segmentation can more accurately identify and segment different types of ICH, which is important

for diagnosis, treatment planning and prognostication. However, multi-class segmentation also has many challenges since ICH come in different shapes, sizes, and densities, and multiple different types of bleeding may occur simultaneously. This task may therefore require more data than the foreground/background problem. We evaluate the performance of supervised segmentation methods under different conditions using the BHSD.

Semi-supervised with pixel-labeled and unlabeled data. Acquiring labeled medical imaging data can be a costly and challenging process, whereas unlabeled data is usually more obtainable. Annotating medical images also requires specialized domain expertise, which can pose a significant barrier to the widespread development of deep learning methods for clinical practice. Semi-supervised learning (SSL) addresses this challenge by using a small amount of labeled data and a large amount of unlabeled data for model training. Using the BHSD, we simulate the SSL scenario by discarding the image-level annotations to build an unlabeled dataset. Our approach combines data with pixel-level annotations and unlabeled data to evaluate the performance of SSL methods.

3 Experiments and Benchmarking Methods

To demonstrate the utility of the BHSD, we perform a series of segmentation experiments under different conditions and provide benchmarks for future model evaluations using this dataset.

Evaluation metrics Segmentation performance was evaluated using the Dice similarity coefficient (DSC) [4], which compares the similarity between the predicted and true segmentation.

Implementation details All experiments were performed on a single A6000 GPU. Following nnUnet [13], we first truncated the HU values of each scan using the range of [-40,120], and then normalized the truncated voxel values by subtracting 40 and dividing by 80. We randomly cropped sub-volumes of size [32,128,128] from CT scans as the input. Other parameters in different models retain the official default settings. In the BHSD, the 192 volumes were divided evenly into training and testing sets. The sets were balanced in terms of the number and types of bleeds, each containing 96 volumes. Furthermore, by verifying the original patient identifiers, no patient was contained in both sets.

3.1 Supervised 3D segmentation

In Experiment 1, we conducted a comprehensive evaluation of state-of-the-art 3D semantic segmentation models using the BHSD dataset. We evaluate five 3D models with SOTA backbones designed for supervised semantic segmentation, namely UNETR [9], Swin UNETR [8,17], CoTr [26], nnFormer [28,22], and nnUNet [13] (Table 1).

Table 1: Benchmark 3D supervised segmentation performance using the BHSD.

	EDH	IPH	IVH	SAH	SDH	Mean
UNETR	1.64	28.28	22.08	4.36	3.63	11.99
Swin UNETR	2.53	34.18	29.28	10.07	8.43	16.89
CoTr	1.63	48.62	53.55	17.88	15.44	27.43
nnFormer	0.00	69.75	25.78	25.94	10.31	29.19
nnUnet3D	4.81	54.12	51.48	21.57	15.23	29.44

The results indicated that nnUnet achieved superior performance compared to other models with an average Dice of 29.44. However, it is important to acknowledge the limitation of the dataset, specifically the low occurrence of epidural hematoma *i.e.,* EDH class. Consequently, the segmentation performance for this class was considerably inferior to other classes. Further refinement and adaptation are necessary to enhance the segmentation performance for the EDH class, considering its limited representation in the dataset. Confusion matrix analysis may also allow identification of imbalances and biases in model predictions (Supplementary Fig. 1).

3.2   Incorporation of scans with no hemorrhage

In Experiment 2, we sought to enhance the model's performance through a gradual augmentation of the training set with negative samples within a supervised experimental setup. This also allowed us to address the issue of false positives. The model achieved optimal performance with 200 negative samples (Table 2), both in terms of hemorrhage segmentation performance and suppression of false positives.

Table 2: Incorporation of scans with (B) and without bleeding (NB). The false positive (FP) rate is also indicated.

Heading level	EDH	IPH	IVH	SAH	SDH	Mean	FP
96B	4.81	54.12	51.48	21.57	15.23	29.44	1.41
96B + 200NB	9.77	56.90	58.53	29.98	21.73	35.38	0.84
96B + 400NB	4.46	39.45	39.90	15.43	9.30	21.71	1.30
96B + 600NB	3.98	36.24	26.25	6.41	6.59	15.90	1.85
96B + 800NB	2.14	28.21	29.57	4.47	2.78	13.43	2.41

This outcome suggests the presence of a balance point, where an insufficient number of negative samples hampers the model's ability to effectively learn the discriminating features, leading to inadequate suppression of false positives, while excessive negative samples may cause the model to overly focus on them, resulting in difficulties distinguishing between target and non-target categories

(Supplementary Fig. 2). It is worth noting that selection of the optimal number of negative samples requires a comprehensive consideration of the model's recognition and generalization abilities, as well as the characteristics and requirements of the dataset. Future research can explore combinations of varying sample quantities to gain a deeper understanding and achieve more precise performance optimization.

3.3   Single class, multiple models versus multiple class, single model

The utilization of existing brain hemorrhage data for multi-class semantic segmentation necessitates training a single-class detection model and merging the prediction outcomes. To further investigate the advantages of the BHSD dataset compared to existing datasets, in Experiment 3 we performed separate training iterations on BHSD, focusing on one category at a time. We repeated this process five times, resulting in distinct models capable of recognizing individual categories. Subsequently, we combined the inference results from these five models. Through comparative analysis, we observed that the multi-class semantic method surpassed its predecessor in the fusion result (Table 3). However, it is important to note that the application of model fusion for single-category segmentation yielded considerable challenges, such as extensive overlapping and conflicting prediction outcomes (Supplementary Fig. 3). This finding underscores the ineffectiveness of the fusion approach for multi-class semantic segmentation.

Table 3: Multiple single class models versus single multi-class model.

Heading level	EDH	IPH	IVH	SAH	SDH	Mean
1class-EDH	5.15	-	-	-	-	-
1class-IPH	-	37.55	-	-	-	-
1class-IVH	-	-	23.60	-	-	-
1class-SAH	-	-	-	14.50	-	-
1class-SDH	-	-	-	-	17.59	
5class-Merge	1.25	3.42	2.57	2.43	2.16	19.68
5class-Single	4.81	54.12	51.48	21.57	15.23	29.44

3.4   Semi-supervised segmentation

Experiment 4 was conducted with the aim of investigating the utilization of unlabelled data from clinical settings to enhance the performance of the dataset. Through the implementation of semi-supervised experiments, we sought to assess the potential for improvement. The unlabelled data introduced in this experiment were randomly sampled, with no available information regarding the health status or presence of hemorrhage. To evaluate this setting, four methods were applied to the BHSD, namely mean teacher [21], cross pseudo supervision (CPS) [2], entropy minimization [24], and interpolation consistency [23]. In these experiments, the training set of 96 volumes was supplemented by 500 unlabeled

data. The same test set of 96 volumes was retained. For this task, we also merged hemorrhage labels to a single foreground mask.

The experimental findings revealed that the model's performance could indeed be further enhanced by employing an appropriate semi-supervised approach (Table 4). This observation highlights the efficacy of incorporating unlabelled data in clinical settings. Moreover, the merging of all categories served to accentuate the extent of performance improvement achieved through this approach. While there was mixed performance, we find that CPS improves on supervised learning and achieves 49.50% DSC in the binary segmentation task. The CPS method [2] uses pseudo-labels generated by one model to train another model, and then using this new model to generate new pseudo-labels and so forth, iterating this process to improve the accuracy of the pseudo-labelled data and improve the performance of the model segmentation.

Table 4: Semi-supervised performance based on nnUNet. To highlight the advantages of semi-supervision, we merged all the hemorrhage categories and report foreground and background semantic segmentation results.

Method	unlabeled samples	Dice
SupOnly	0	45.10 ±0.21
Entropy Minimization	500	36.91 ±0.16
Mean Teacher	500	44.63 ±0.18
Interpolation Consistency	500	45.38 ±0.33
Cross Pseudo Supervision	500	49.50 ±0.19

4   Conclusion

We describe a 3D CT head dataset, the BHSD, for intracranial hemorrhage segmentation. This dataset includes a diverse mix of head scans with pixel-level and slice-level annotations, as well as scans with and without hemorrhage. To qualitatively and quantitatively scrutinize the characteristics of the BHSD, we compare popular SOTA models and diverse training techniques and draw three key insights from our benchmarking experiments. Firstly, the BHSD can significantly enhance the performance of SOTA models for multi-class segmentation of ICH. Multi-class segmentation significantly outperforms the fusion of single-class segmentation models. Secondly, incorporation of scans without hemorrhage can enhance segmentation performance. However, the right balance needs to be found. Third, the BHSD improves model performance using a semi-supervised approach even when the volumes are not annotated at the pixel-level. Hence, the BHSD is a valuable dataset for ICH segmentation models, providing an opportunity to study how segmentation tasks can make better use of unlabeled and weakly labeled data. This will in turn facilitate the development and validation of computer-aided diagnostic tools for clinical practice.

References

1. Auer, L.M., Deinsberger, W., Niederkorn, K., Gell, G., Kleinert, R., Schneider, G., Holzer, P., Bone, G., Mokry, M., Körner, E., et al.: Endoscopic surgery versus medical treatment for spontaneous intracerebral hematoma: a randomized study. Journal of neurosurgery **70**(4), 530–535 (1989)
2. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021)
3. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Development and validation of deep learning algorithms for detection of critical findings in head ct scans (2018)
4. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
5. Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P., et al.: Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. Radiology: Artificial Intelligence **2**(3), e190211 (2020)
6. Frontera JA, Claassen J, S.J.W.K.T.R.C.E.J.M.R., SA., M.: Prediction of symptomatic vasospasm after subarachnoid hemorrhage: the modified fisher scale. Neurosurgery **59**(1), 21–27 (2006)
7. Grønbæk, H., Johnsen, S.P., Jepsen, P., Gislum, M., Vilstrup, H., Tage-Jensen, U., Sørensen, H.T.: Liver cirrhosis, other liver diseases, and risk of hospitalisation for intracerebral haemorrhage: a danish population-based case-control study. BMC gastroenterology **8**, 1–6 (2008)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I. pp. 272–284. Springer (2022)
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
10. Hemphill III, J.C., Greenberg, S.M., Anderson, C.S., Becker, K., Bendok, B.R., Cushman, M., Fung, G.L., Goldstein, J.N., Macdonald, R.L., Mitchell, P.H., et al.: Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the american heart association/american stroke association. Stroke **46**(7), 2032–2060 (2015)
11. Howard, G., Cushman, M., Howard, V.J., Kissela, B.M., Kleindorfer, D.O., Moy, C.S., Switzer, J., Woo, D.: Risk factors for intracerebral hemorrhage: the reasons for geographic and racial differences in stroke (regards) study. Stroke **44**(5), 1282–1287 (2013)
12. Hssayeni, M., Croock, M., Salman, A., Al-khafaji, H., Yahya, Z., Ghoraani, B.: Computed tomography images for intracranial hemorrhage detection and segmentation. Intracranial Hemorrhage Segmentation Using A Deep Convolutional Model. Data **5**(1), 14 (2020)
13. 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)

14. Larobina, M., Murino, L.: Medical image file formats. Journal of digital imaging **27**, 200–206 (2014)
15. Lee, H., Kim, M., Do, S.: Practical window setting optimization for medical image deep learning. arXiv preprint arXiv:1812.00572 (2018)
16. Li, X., Luo, G., Wang, K., Wang, H., Li, S., Liu, J., Liang, X., Jiang, J., Song, Z., Zheng, C., et al.: The state-of-the-art 3d anisotropic intracranial hemorrhage segmentation on non-contrast head ct: The instance challenge. arXiv preprint arXiv:2301.03281 (2023)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
18. McCarron, M.O., Nicoll, J.A., Ironside, J.W., Love, S., Alberts, M.J., Bone, I.: Cerebral amyloid angiopathy–related hemorrhage: Interaction of apoe $\varepsilon 2$ with putative clinical risk factors. Stroke **30**(8), 1643–1646 (1999)
19. Reis, E.P., Nascimento, F., Aranha, M., Machado, B., Felix, M., Stein, A., Amaro, E.: Brain hemorrhage extended (bhx): Bounding box extrapolation from thick to thin slice ct images. PhysioNe (2020)
20. Steiner, T., Salman, R.A.S., Beer, R., Christensen, H., Cordonnier, C., Csiba, L., Forsting, M., Harnof, S., Klijn, C.J., Krieger, D., et al.: European stroke organisation (eso) guidelines for the management of spontaneous intracerebral hemorrhage. International journal of stroke **9**(7), 840–855 (2014)
21. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
22. 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)
23. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks **145**, 90–106 (2022)
24. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
25. Wang, X., Shen, T., Yang, S., Lan, J., Xu, Y., Wang, M., Zhang, J., Han, X.: A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head ct scans. NeuroImage: Clinical **32**, 102785 (2021)
26. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 171–180. Springer (2021)
27. Yushkevich, P.A., Gerig, G.: Itk-snap: an intractive medical image segmentation tool to meet the need for expert-guided segmentation of complex medical images. IEEE pulse **8**(4), 54–57 (2017)
28. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)