frontiers | Frontiers in Neurology

TYPE Original Research
PUBLISHED 13 December 2024
DOI 10.3389/fneur.2024.1490216

OPEN ACCESS

EDITED BY
Hsiao-Lung Chan,
Chang Gung University, Taiwan

REVIEWED BY
Tijana Nastasovic,
University of Belgrade, Serbia
Syeda Mahrukh Fatima Zaidi,
Dow University of Health Sciences, Pakistan

*CORRESPONDENCE
Dietmar Frey
✉ dietmar.frey@charite.de

RECEIVED 02 September 2024
ACCEPTED 29 November 2024
PUBLISHED 13 December 2024

CITATION
Kiewitz J, Aydin OU, Hilbert A, Gultom M, Nouri A, Khalil AA, Vajkoczy P, Tanioka S, Ishida F, Dengler NF and Frey D (2024) Deep learning-based multiclass segmentation in aneurysmal subarachnoid hemorrhage. *Front. Neurol.* 15:1490216. doi: 10.3389/fneur.2024.1490216

Deep learning-based multiclass segmentation in aneurysmal subarachnoid hemorrhage

Julia Kiewitz[1,2], Orhun Utku Aydin[1], Adam Hilbert[1], Marie Gultom[1], Anouar Nouri[1], Ahmed A. Khalil[3], Peter Vajkoczy[2], Satoru Tanioka[1], Fujimaro Ishida[4], Nora F. Dengler[2,5,6] and Dietmar Frey[1,2]*

[1]CLAIM – Charité Lab for AI in Medicine, Charité Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany, [2]Department of Neurosurgery, Charité Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany, [3]Center for Stroke Research Berlin, Charité Universitätsmedizin Berlin, Berlin, Germany, [4]Department of Neurosurgery, Mie Chuo Medical Center, Tsu, Mie, Japan, [5]Faculty of Health Sciences Brandenburg, Medical School Theodor Fontane, Bad Saarow, Germany, [6]Department of Neurosurgery, HELIOS Hospital Bad Saarow, Bad Saarow, Germany

Introduction: Radiological scores used to assess the extent of subarachnoid hemorrhage are limited by intrarater and interrater variability and do not utilize all available information from the imaging. Image segmentation enables precise identification and delineation of objects or regions of interest and offers the potential for automatization of score assessments using precise volumetric information. Our study aims to develop a deep learning model that enables automated multiclass segmentation of structures and pathologies relevant for aneurysmal subarachnoid hemorrhage outcome prediction.

Methods: A set of 73 non-contrast CT scans of patients with aneurysmal subarachnoid hemorrhage were included. Six target classes were manually segmented to create a multiclass segmentation ground truth: subarachnoid, intraventricular, intracerebral and subdural hemorrhage, aneurysms and ventricles. We used the 2d and 3d configurations of the nnU-Net deep learning biomedical image segmentation framework. Additionally, we performed an interrater reliability analysis in our internal test set (n = 20) and an external validation on a set of primary intracerebral hemorrhage patients (n = 104). Segmentation performance was evaluated using the Dice coefficient, volumetric similarity and sensitivity.

Results: The nnU-Net-based segmentation model demonstrated performance closely matching the interrater reliability between two senior raters for the subarachnoid hemorrhage, ventricles, intracerebral hemorrhage classes and overall hemorrhage segmentation. For the hemorrhage segmentation a median Dice coefficient of 0.664 was achieved by the 3d model (0.673 = 2d model). In the external test set a median Dice coefficient of 0.831 for the hemorrhage segmentation was achieved.

Conclusion: Deep learning enables automated multiclass segmentation of aneurysmal subarachnoid hemorrhage-related pathologies and achieves performance approaching that of a human rater. This enables automatized volumetries of pathologies identified on admission CTs in patients with subarachnoid hemorrhage potentially leading to imaging biomarkers for improved outcome prediction.

KEYWORDS

subarachnoid hemorrhage, deep learning, multiclass segmentation, interrater reliability, outcome prediction

1 Introduction

Aneurysmal subarachnoid hemorrhage (aSAH) is a severe subtype of stroke with an incidence of 8/100,000 person-years leading to a potential loss of many years of productive life (1, 2). Patient outcomes following aSAH show significant variability; while aSAH is lethal for one-third of patients, survivors may become permanently dependent on nursing care due to residual cognitive and functional impairment (3). In aSAH several clinical and imaging variables were shown to predict outcome. The initial imaging performed for diagnosing SAH is non-contrast CT imaging (NCCT) and delivers important information on the presence, volume, location and radiological appearance of the hemorrhage (4).

Different radiological scores are used to assess the extent of the subarachnoid hemorrhage and highlight different aspects of SAH. For example, SEBES (subarachnoid hemorrhage early brain edema score) predicts delayed cerebral ischemia (DCI) and unfavorable outcomes (5), the Graeb Score focuses on intraventricular bleeding (6) and the Hijdra Sum Score includes a grading system based on the amount of subarachnoid blood in different localizations of the brain (7). A different common scoring system is the Barrow Neurological Institute (BNI) scoring system which estimates the risk of vasospasm depending on the radiological extent of SAH in NCCT via measuring the thickness of the hemorrhage and distinguishing five different severity levels (8). Out of all radiological scores the modified Fisher Scale (9) is predominantly used in clinical practice due to its practicality. It quantifies the amount of subarachnoid blood to predict cerebral vasospasm, a condition leading to poor outcome and high mortality.

A main limitation of radiological scores is their subjective nature and considerable interrater variability, which may limit their predictive value and clinical utility. For instance, the Fisher Scale does not include an exact definition of the various scoring criteria—thick subarachnoid clot or presence of intracerebral and intraventricular hemorrhage—determining the different severity levels (10, 11). The vague definition of imaging findings further reduces the interrater agreement which can be measured by weighted kappa score k_w reported for various scores (k_w for the Fisher Scale depending on different studies = 0.66/0.53/0.45) (10–14). The radiological scores only focus on specific aspects of the images like blood thickness, location, or volume, without providing a comprehensive assessment of available information in the medical images.

Recently, several studies have proposed quantifying the SAH volume using automated segmentation methods in NCCT (15–18). There are different methods for binary segmentation of subarachnoid hemorrhage including threshold-based (16), semi-automated (17), and deep learning-based methods (15, 19) not only quantifying the SAH-volume but also trying to predict parameters that are relevant to the patient's outcome—e.g., the prediction of vasospasm risk (17). Whereas a manual segmentation provided by an experienced human rater is considered the gold standard (17), this has several limitations including the high time effort and relatively low interrater reliability (15, 16). Deep learning-based methods can provide a more objective assessment of hemorrhage volumes in an automated way. Due to its high clinical relevance, hemorrhage segmentation is addressed as part of segmentation challenges such as the INSTANCE challenge 2022 and the upcoming MICCAI 2024 challenge MBH-Seg for segmentation of intracranial hemorrhage (20–22). One limitation of recently proposed models is that they only provide binary labels for intracranial hemorrhage subtypes or do not explicitly consider various co-occurring pathologies and relevant anatomical changes such as enlarged or shifted ventricles. Hence the automated segmentation can only differentiate between healthy brain tissue and hemorrhage and cannot classify the different types of hemorrhage. The distinction of the hemorrhage types is especially important because the presence of co-occurring pathologies—like ICH together with SAH—significantly worsens the outcome of aSAH patients (23, 24). One of the first studies that assess multiclass segmentation of intracranial hemorrhage in patients after traumatic brain injury was published as a preprint in 2023 by Wu et al. (25). In contrast to our study, Wu et al. only assessed pathologies, e.g., different types of intracranial hemorrhages and did not consider anatomical structures which, when altered, may also be of prognostic value.

Our study aims to develop a multiclass deep learning-based segmentation model tailored to aneurysmal subarachnoid hemorrhage that includes outcome-relevant structures and pathologies for rapid and accurate volume segmentation. The proposed method segments six different classes: basal and cortical SAH, intraventricular hemorrhage (IVH), ventricles, intracerebral hemorrhage (ICH), aneurysms visible on NCCT and subdural hematoma (SDH). Moreover, we assess expert interrater agreement and perform external validation of our model on patients with a different main pathology (primary ICH). Our model is made available open source with pre-trained weights to facilitate the extraction of outcome-related pathologies from NCCT images of aSAH patients for further research.

Abbreviations: SAH, subarachnoid hemorrhage; aSAH, aneurysmal SAH; ICH, intracerebral hemorrhage; IVH, intraventricular hemorrhage; SDH, subdural hematoma; NCCT, non-contrast CT scans; CNN, convolutional neural network; BNI, Barrow Neurological Institute; DICOM, Digital imaging and communications in medicine; NIfTI, Neuroimaging informatics technology initiative; TP, true positives; FP, false positives; FN, false negatives; DCI, delayed cerebral ischemia; SEBES, subarachnoid hemorrhage early brain edema score; VS, volumetric similarity; EDH, epidural hematoma; IPH, intraparenchymal hemorrhage; ICU, intensive care unit; AI, artificial intelligence; IQR, interquartile range.

2 Materials and methods

2.1 Patients

Out of 408 aSAH patients treated in the Department of Neurosurgery at Charité University Hospital in Berlin from 2009 to 2015, 73 patients meeting the inclusion criteria were randomly selected and retrospectively included. The selection of a limited subset of 73 patients meeting the inclusion criteria was due to time

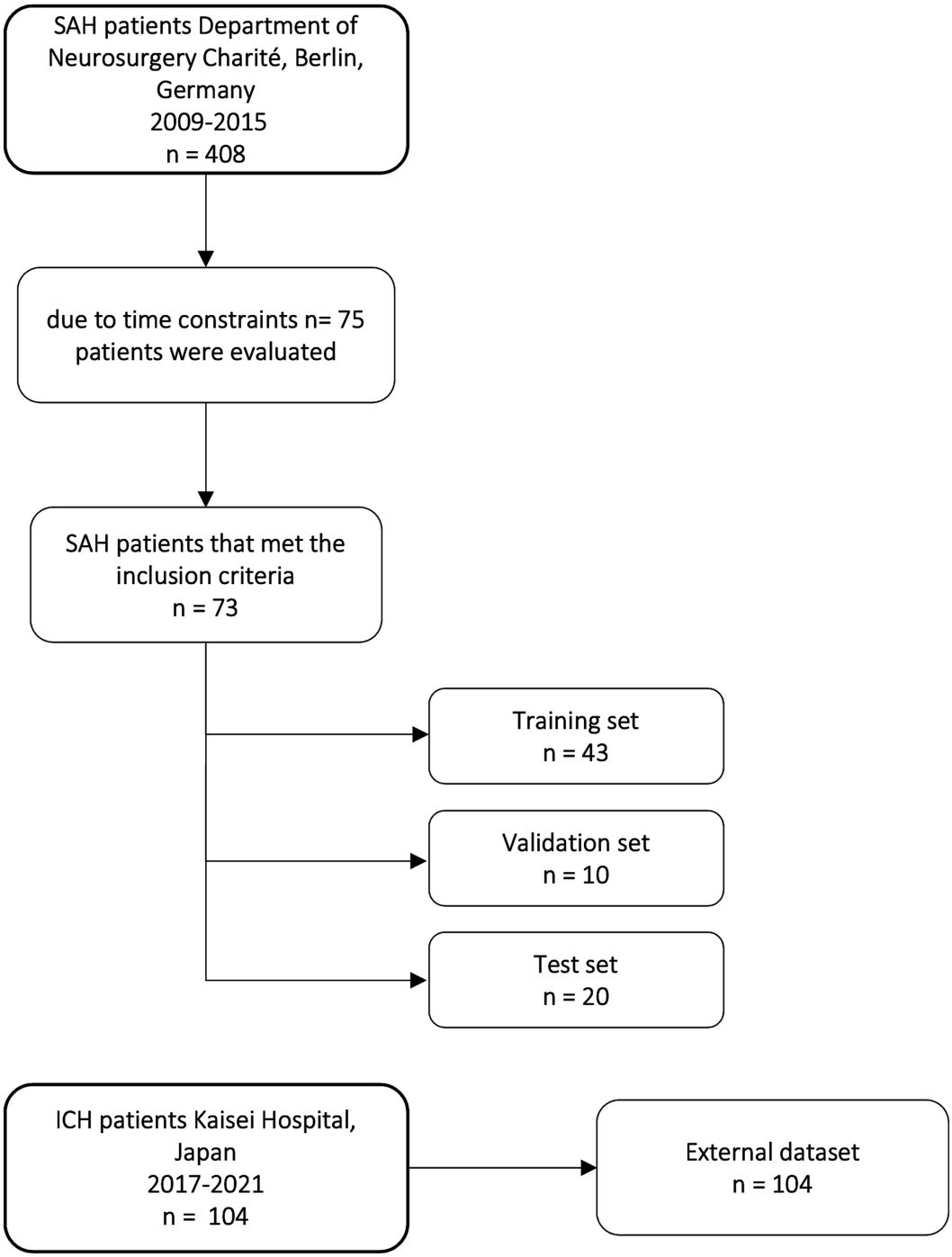

FIGURE 1
Flow chart of patients included in our study from the Department of Neurosurgery Charité University Hospital and distribution of training, validation, and test set and external dataset from Kaisei Hospital, Japan.

constraints regarding manual segmentation (Figure 1). Inclusion criteria were diagnosis of aSAH; CT scan at admission; CT scan slice thickness between 4 and 6 mm; and patients aged 18 or older. aSAH was diagnosed through non-contrast computed tomography (NCCT) scans, and if NCCT was negative, lumbar puncture was performed to assess bilirubin or other blood degradation products. 69 out of 73 patients had visible hemorrhage in NCCT and 4 patients were diagnosed with SAH by lumbar puncture because NCCT was negative.

To test for generalizability, we validated our final model on external data. For the external validation, we analyzed NCCT images of 104 patients presenting with the primary diagnosis of ICH at Kaisei Hospital, Japan.

2.2 Ethics statement

Ethics approval was granted by the local authorities of Charité University Hospital (EA1/291/14) and Kaisei Hospital (2020-05), respectively. The need for participant consent was waived by both ethics committees due to the analysis being retrospective. All data

TABLE 1 Distribution of hemorrhage and artifacts in datasets.

	Training set = 43	Validation set = 10	Test set = 20	*n* = 73
Basal and cortical SAH (%)	41 (95%)	10 (100%)	20 (100%)	71 (97%)
Intraventricular hemorrhage (IVH) (%)	25 (58%)	6 (60%)	11 (55%)	42 (58%)
Intracerebral hemorrhage (ICH) (%)	16 (37%)	2 (20%)	7 (35%)	25 (34%)
Visible aneurysm (%)	5 (12%)	1 (10%)	3 (15%)	9 (12%)
Subdural hematoma (SDH) (%)	2 (0.5%)	1 (10%)	2 (10%)	5 (7%)
Metal artifacts (%)	2 (5%)	1 (10%)	2 (10%)	5 (7%)
Shunt artifacts (%)	2 (5%)	1 (10%)	1 (5%)	4 (5%)

were fully anonymized before the analysis. The data of our main dataset was first accessed on December 1st, 2021. The data of the external set was first accessed on August 15th, 2023.

2.3 Image processing

The NCCT images were converted from DICOM (Digital imaging and communications in medicine) to NIfTI (Neuroimaging informatics technology initiative) file format using the dicom2niix command line tool in nipype (26). No gantry tilt correction was applied. Preprocessing of images, including reshaping, and resampling was performed by the default configuration of the nnU-Net framework. All CT scans had a slice thickness of 4–6 mm (median slice thickness 5.0002 mm) with 24–32 slices (median number of slices: 28). The median voxel spacing after preprocessing was 5 × 0.44 × 0.44 mm and the median shape of the volumes were 512 × 512 × 28. For the external dataset the images were preprocessed manually to have uniform slice thickness of 2 mm with a shape of 512 × 512 × 80 using 3D slicer (27).

2.4 Dataset and labeling

Seventy three head NCCT scans that met the inclusion criteria were manually segmented with ITK-Snap, an open-source 3d medical image analysis software (28) (ITK-SNAP Version 3.6.0). For the segmentation, six different target classes were determined: basal and cortical SAH, IVH, ventricles, ICH, aneurysms and SDH. All 73 CT scans were first manually segmented by a junior rater (JK) then checked and corrected by a senior rater (ND, senior physician neurosurgeon). The raters, except for JK, had no access to any other information about the patients besides the NCCT image. The manual segmentation of one CT scan took between 3 and 6 h depending on the complexity and volume of the hemorrhages.

For the deep learning models, 73 CT scans were split into three sets: training set (n = 43), validation set (n = 10), and test set (n = 20). The split was stratified by balancing the number of our six classes and the occurrence of metal artifacts and shunts in each set to avoid biases in the training of the nnU-Net models (Table 1). Table 1 shows the distribution of the classes in the different sets.

For quality assurance, we calculated the interrater agreement. Another junior rater (MG) independently segmented the test set (20 out of the 73 CT scans) and another senior rater (AK, neuroradiology resident) checked and approved the segmentations to create a comparison group. In 4 patients we were only able to segment the ventricles as there was no visible hemorrhage in the NCCT, 2 of these patients belonged to the training set, one belonged to the validation set and one patient belonged to the test set.

Due to the time effort of multiclass segmentation of 3–6 h per NCCT, we validated the results using an already-available binary hemorrhage segmentation dataset (hemorrhage or no hemorrhage) for the external validation of our models. The dataset was labeled by ST (senior physician neurosurgeon) using the software 3D slicer (27).

For the evaluation, the outputs of our multiclass 2d model were merged to create a single hemorrhage class.

2.5 Deep learning segmentation

nnU-Net is a state-of-the-art biomedical image segmentation tool with automated configuration, including preprocessing, network architecture, training, and post-processing which gives the framework high flexibility (29). We also tested BRAVENET, which is a multiscale 3-D convolutional neural network (CNN) model initially developed on a dataset of patients with cerebrovascular diseases. For brain artery segmentation tasks, the BRAVENET architecture showed a superior performance compared to a standard U-Net (30).

2.6 Training

The nnU-Net framework was trained in two configurations: (1) 3d and (2) 2d using axial slices. The models were trained for 1,000 epochs. The best-performing model on the validation set after 1,000 epochs was selected for evaluation on the test set. Model training was performed on an RTX Titan GPU with 24GB of VRAM, while the GPU VRAM argument was set to 8GB (default) for the training. The selected patch sizes, hyperparameters, and model architecture details can be found in Table 2. We used the 2d model for compatibility with various slice thickness values and across datasets and patients.

2.7 Evaluation

The segmentation performance was evaluated using the EvaluateSegmentation tool (31). The quantitative evaluation of performance on each class was performed separately, and additionally a general hemorrhage group was created where all hemorrhage classes were merged to test the general ability of our model to detect blood in NCCT. To assess the performance, different metrics were reported: the median Dice coefficient, calculated on individual patients, the global

TABLE 2 Model hyperparameters.

Model configuration	Patch size	Convolutional kernel size	Starting/maximum number of convolutional filters	Batch size	Number of convolutional stages
3d_fullres	320 × 320 × 16	3 × 3 × 3	32–320	2	7
2d	512 × 512	3 × 3 × 3	32–512	12	8

Dice coefficient, by considering all patients in the test set to be one single concatenated image, volumetric similarity and sensitivity. Classes where the model segmented less than 20 voxels were excluded from metric calculations.

The Dice coefficient was used as the main metric for performance assessment. In addition to reporting mean and median values for the Dice coefficient a global Dice coefficient was calculated to provide a general overview of the models' performance on the whole dataset (Equation 1), since in patients with smaller SAHs even small errors in segmentations can easily lead to loss of overlap and impact mean Dice values. The Dice coefficient was calculated according to following formula:

$$\text{Dice coefficient} = \frac{2 * TP}{2 * TP + FP + FN} \quad (1)$$

Equation 1: Calculation of Dice coefficient and global Dice coefficient (31).

Where TP (True Positives), FP (False Positives) and FN (False Negatives) are used from each individual patient. For calculating the global Dice coefficient, the same equation applies as the FP, TP and FN voxels across the entire test set were used.

$$\text{Volumetric similarity} = 1 - \frac{|FN - FP|}{2 * TP + FP + FN} \quad (2)$$

Equation 2: Calculation of volumetric similarity (31).

While the global Dice coefficient provides valuable insight into the spatial overlap between segmented and ground truth regions, we also calculated the volumetric similarity (Equation 2). Volumetric similarity directly emphasizes the portion of the segmented volume in relation to the reference volume and is an important measure for outcome prediction (31). Detailed descriptions and formulas of the metrics can be found in the study by Taha et al. (31). Bland Altman Plots were created to show the volumetric similarity and bias of the models for the SAH class for the 3d and 2d nnU-Net model and Rater 2 and the merged hemorrhage class. We also report the cases where the different models and the raters agreed or disagreed on the presence of a certain class.

3 Results

We assessed 73 head CT scans from patients with aSAH at admission. Median age was 48 years [min; max = 22; 92 years] and 63% of the patients were female (Table 3). Out of 73 CT scans, there were 71 with basal and/or cortical SAH, 42 with IVH, 25 with ICH, 9 with visible aneurysms, and 5 with SDH (Table 1). Out of 73 patients, 28 patients had image artifacts or additional treatment material: 4 had an intraventricular shunt, 19 had movement artifacts in the form of blur and 5 had metal artifacts due to previous aSAH therapy in the form of aneurysmal clipping or coiling.

3.1 nnU-Net segmentation results

In general, the nnU-Net models achieved human-level performance for the SAH, the ICH, the ventricle and the hemorrhage segmentation class. For other classes such as the IVH, aneurysm, and the SDH class, the automated segmentation performance was lower than human-level performance (Table 4). Overall, the 3d and 2d models performed similarly. The 3d model achieved a global Dice coefficient in the hemorrhage class of 0.730 and the 2d model achieved a global Dice coefficient of 0.736 (Table 4). For the SAH class, the 3d and 2d models achieved the same result with a global Dice coefficient of 0.686. The models had a slightly better performance for the ICH and IVH classes with a global Dice coefficient of 0.750 (IVH 3d and 2d model) and 0.743 (ICH 3d model) respectively 0.765 (ICH 2d model). The segmentation performance of the aneurysm class varied the most (3d model: global Dice coefficient = 0.037; 2d model: global Dice coefficient = 0.366). The results of the automatic segmentation of the 3d and 2d models of the SDH class were similar to the segmentation results of the IVH and ICH classes. The ventricle class segmentation achieved the highest global Dice coefficient of 0.875/0.872 (3d model/2d model).

The checkpoints of the segmentation models can be found in the following GitHub repository: https://github.com/claim-berlin/aSAH-multiclass-segmentation.

The experimentation we performed with additional hyperparameter configurations and another in-house model architecture coined BRAVENET (30) can be found in the Appendix Tables 7, 8. Overall, the nnU-Net configured model outperformed the BRAVENET architecture. For reporting our results in the following sections, we only focused on the 2d and 3d nnU-Net models because no significant performance gains were observed by changing the architecture or tuning its hyperparameters.

Figures 2, 3 show a few examples of the segmentations of our model on different hemorrhage classes in comparison to the ground truth (Rater 1). The failure cases of the nnU-Net models are shown in Figure 4.

3.2 External dataset results

In the external dataset the median age was 71 years [min; max = 36; 98 years] and 46% of the patients were female. The external dataset included mainly patients with ICH as primary diagnosis, but many patients had co-occurring pathologies like SAH, IVH, and SDH. The automated segmentation of the external dataset achieved a

10.3389/fneur.2024.1490216

TABLE 3 Patient characteristics.

	Training set *n* = 43	Validation set *n* = 10	Test set *n* = 20	Total *n* = 73	External testing *n* = 104
Sex					
Male (%)	17 (40%)	0 (0%)	10 (50%)	27 (37%)	56 (54%)
Female (%)	26 (60%)	10 (100%)	10 (50%)	46 (63%)	48 (46%)
Age					
Mean age	50	62	49	52	70
Median age	46 (IQR = 19)	62 (IQR = 24)	49 (IQR = 12)	48 (IQR = 17)	71 (IQR = 23)
Glasgow coma scale					
GCS < 9	25	5	7	37	15

TABLE 4 Comparison of results.

	SAH	IVH	Ventricles	ICH	Aneurysm	SDH	Hemorrhage class
Global dice							
3d_fullres	0.686	0.750	0.875	0.743	0.037	0.758	0.730
2d	0.686	0.750	0.872	0.765	0.366	0.766	0.736
Rater 2	0.663	0.808	0.839	0.750	0.877	0.902	0.732
External dataset	–	–	–	–	–	–	0.838
Median dice (Q1/Q3)							
3d_fullres	0.581 (0.515/0.691)	0.568 (0.197/0.664)	0.862 (0.815/0.888)	0.622 (0.565/0.729)	0.208 (0.104/0.275)	0.758 (0.758/0.758)	0.664 (0.545/0.735)
2d	0.616 (0.489/0.689)	0.515 (0.232/0.765)	0.855 (0.832/0.874)	0.692 (0.530/0.855)	0.541 (0.435/0.646)	0.766 (0.766/0.766)	0.673 (0.498/0.754)
Rater 2	0.626 (0.511/0.683)	0.792 (0.699/0.822)	0.808 (0.753/0.854)	0.834 (0.613/0.873)	0.833 (0.741/0.861)	0.896 (0.891/0.901)	0.659 (0.560/0.764)
External dataset	–	–	–	–	–	–	0.831 (0.724/0.872)
Median volumetric similarity (Q1/Q3)							
3d_fullres	0.922 (0.833/0.973)	0.698 (0.365/0.973)	0.947 (0.924/0.979)	0.741 (0.593/0.932)	0.233 (0.118/0.416)	0.775 (0.775/0.775)	0.892 (0.820/0.960)
2d	0.843 (0.731/0.909)	0.723 (0.533/0.852)	0.938 (0.900/0.966)	0.898 (0.701/0.951)	0.660 (0.502/0.818)	0.786 (0.786/0.786)	0.879 (0.729/0.914)
Rater 2	0.835 (0.704/0.927)	0.860 (0.719/0.927)	0.890 (0.849/0.919)	0.879 (0.670/0.889)	0.920 (0.824/0.934)	0.928 (0.927/0.929)	0.831 (0.779/0.915)
External dataset	–	–	–	–	–	–	0.906 (0.835/0.954)
Median sensitivity (Q1/Q3)							
3d_fullres	0.589 (0.509/0.683)	0.479 (0.111/0.573)	0.858 (0.821/0.910)	0.616 (0.402/0.942)	0.118 (0.059/0.181)	0.619 (0.619/0.619)	0.628 (0.553/0.698)
2d	0.583 (0.494/0.692)	0.484 (0.138/0.668)	0.819 (0.777/0.842)	0.576 (0.533/0.899)	0.466 (0.333/0.600)	0.632 (0.632/0.632)	0.662 (0.492/0.705)
Rater 2	0.492 (0.411/0.614)	0.720 (0.546/0.850)	0.745 (0.627/0.773)	0.790 (0.585/0.832)	0.792 (0.651/0.879)	0.836 (0.832/0.840)	0.578 (0.446/0.646)
External dataset	–	–	–	–	–	–	0.765 (0.684/0.830)

Comparison of different metrics for different labels for 3d model, 2d model, Rater 2 and the external set. Values were only calculated when there was an agreement on the presence of one class.

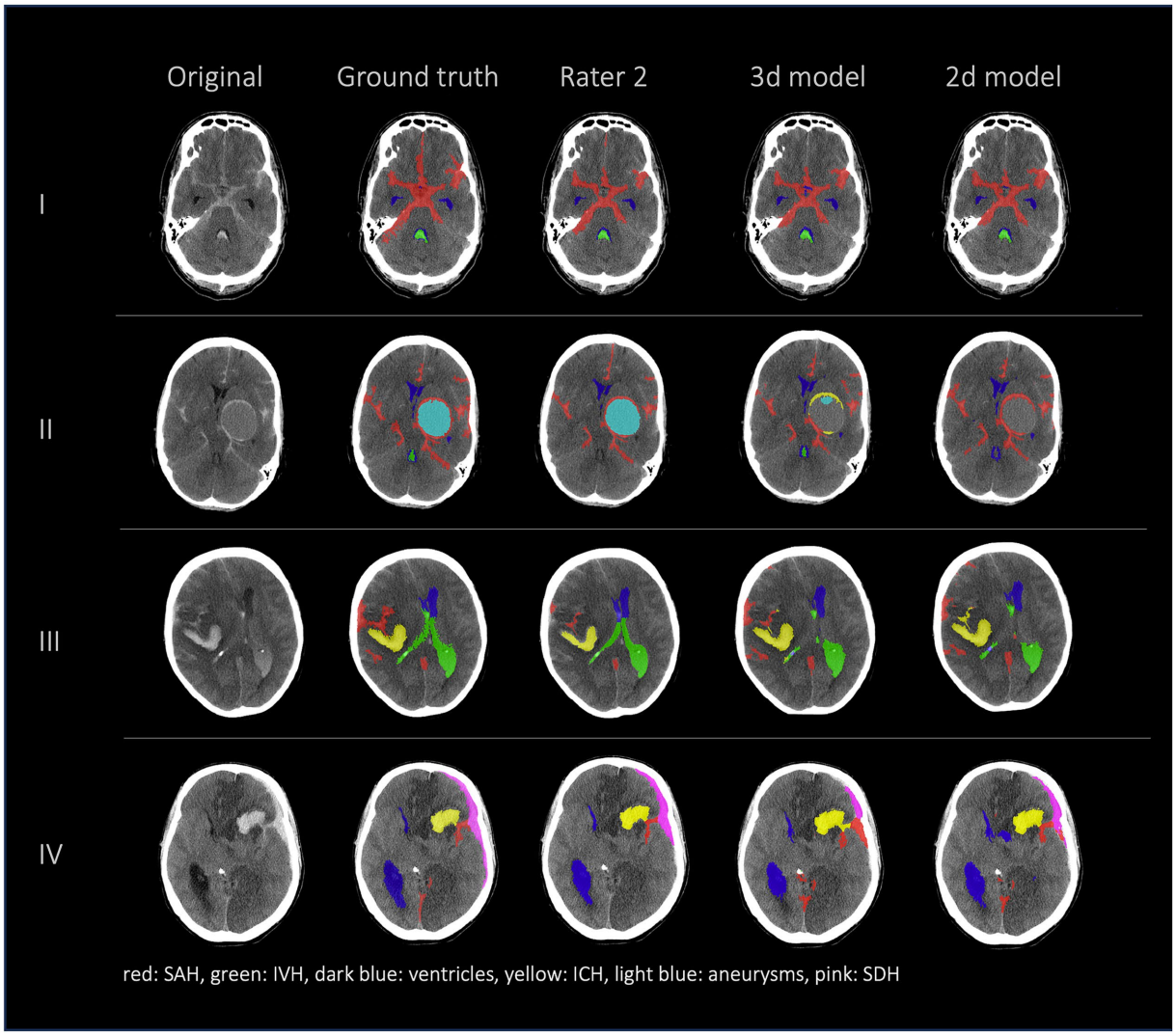

FIGURE 2
Comparison of 4 patients' segmentations in the following order from left to right: first column shows the original NCCT, second column shows the ground truth (segmentation of Rater 1), third column shows the segmentation of Rater 2 and the fourth and fifth columns show the results of our nnU-Net models (fourth column: 3d model, fifth column: 2d model). This figure shows that the nnU-Net is able to detect hemorrhage but struggles to assign the segmentations to the different classes. The second row (II) effectively illustrates our findings that in the aneurysm class the results of the nnU-Net models and Rater 2 differed the most.

global Dice coefficient of 0.838 for the merged hemorrhage class (Table 4 and Figure 5).

As shown in Table 5 both nnU-Net models agreed with the ground truth of Rater 1 on the presence of SAH in all cases, whereas Rater 1 and Rater 2 did not agree on the presence of SAH in two patients. The two different nnU-Net models agreed with the ground truth on the presence of IVH, ICH, and SDH in 8/5/1 cases. The two nnU-Net models disagreed with the ground truth mostly on the presence of IVH and ICH (3d model: 3/4; 2d model: 4/4), whereas Rater 2 disagreed on a relatively similar number of cases in the IVH, ICH, aneurysm and SDH class (2/3/2/1). The ventricle class showed no disagreements between the ground truth, the 3d model, the 2d model, and Rater 2. The agreements and disagreements for all classes are shown in Table 5.

In this study we not only assessed the Dice coefficient but also calculated the volumetric similarity (*VS*) to create a connection between the correctly segmented volume and the overall hemorrhage volume. Figure 6 shows the volumetric similarity for the SAH class using Bland Altman Plots and therefore the absolute and relative difference between the ground truth, the segmented SAH volumes by Rater 2 and by the 3d and 2d nnU-Net model. It illustrates that even though smaller volumes have smaller differences in segmented volume (Figure 6: left column) relatively speaking, the deviation based on the total segmented volume is larger (Figure 6: right column).

4 Discussion

We present in the current work a deep learning-based analysis of automated multiclass segmentation for SAH-related pathologies from NCCTs of patients with SAH. Our framework demonstrates high quantitative performance based on the Dice coefficient and volumetric similarity that is close to the performance of a human rater. We show

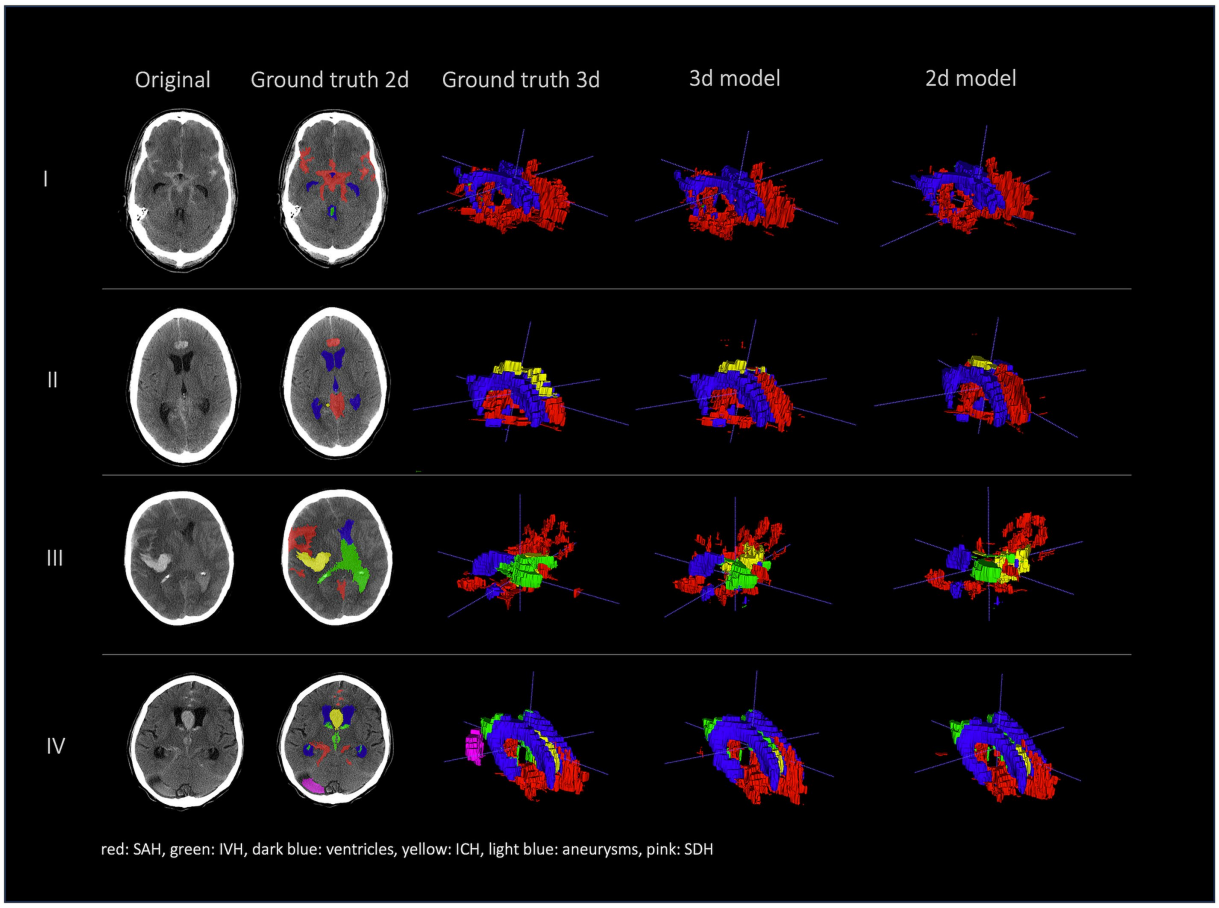

FIGURE 3
Comparison of segmentations of 4 patients in the following order from left to right: first column shows original NCCT, second column shows ground truth (Rater 1 segmentation) in 2d, third column shows 3d representation of ground truth (Rater 1), fourth column shows 3d representation of the 3d model and fifth column shows 3d representation of the 2d model. This figure shows a 3d reconstruction of the hemorrhage distribution which might be useful in clinical practice to get an overview of the extent of the hemorrhage.

that our nnU-Net models are able to do a precise and fast segmentation of aSAH related classes that are relevant to assess patient outcome.

4.1 Our results

We evaluated our model with quantitative measures, compared performance against interrater variability of medical expert raters, and performed an external validation. Overall, despite a limited training dataset, we attain highly promising results in most classes and demonstrate generalizability on an external dataset featuring ICH as primary hemorrhage class.

We observe varied performance across the different classes and identify important findings that might explain these differences in performance. Overall, the two deep learning models, 3d model and 2d model, show a similar performance. Segmentations of SAH, ventricles, ICH and the hemorrhage segmentation are comparable in 3d model, 2d model and Rater 2 whereas the aneurysm and SDH class show the highest variability. The difference between the performance of the nnU-Net models in comparison to Rater 2 is relatively high in the aneurysm and SDH class. This is most likely due to the limited occurrence of these classes in our training set and the limited size of our internal test set. Compared to the segmentation of the SAH class, better results were achieved for the IVH, ICH and SDH class by the nnU-Net models. The interrater reliability was assessed to create human benchmarks which can be used to put our models' results into context. A comparison of the models' results to human benchmarks is necessary because absolute values of metrics are challenging to interpret and thus model performance can be over−/underestimated. For instance, from an absolute value point of view, a global Dice coefficient of 0.686 for SAH is lower than a global Dice coefficient of 0.758 (3d model)/0.766 (2d model) for SDH but actually compared to human benchmarks (Rater 2 global Dice coefficient of 0.665 for SAH and 0.902 for SDH) the overall automated segmentation performance for SAH is closer to the human performance and is considered a better result when aiming for human level performance. In this sense, optimizing for higher performance—with an improvement of the global Dice coefficient—with architectural changes or optimizing the training approach is not the main goal of our study.

Interestingly, we achieve a superior performance according to the global Dice coefficient in the external test dataset compared to the internal test set for the ICH class. This is a relatively uncommon occurrence in the field of machine learning and is

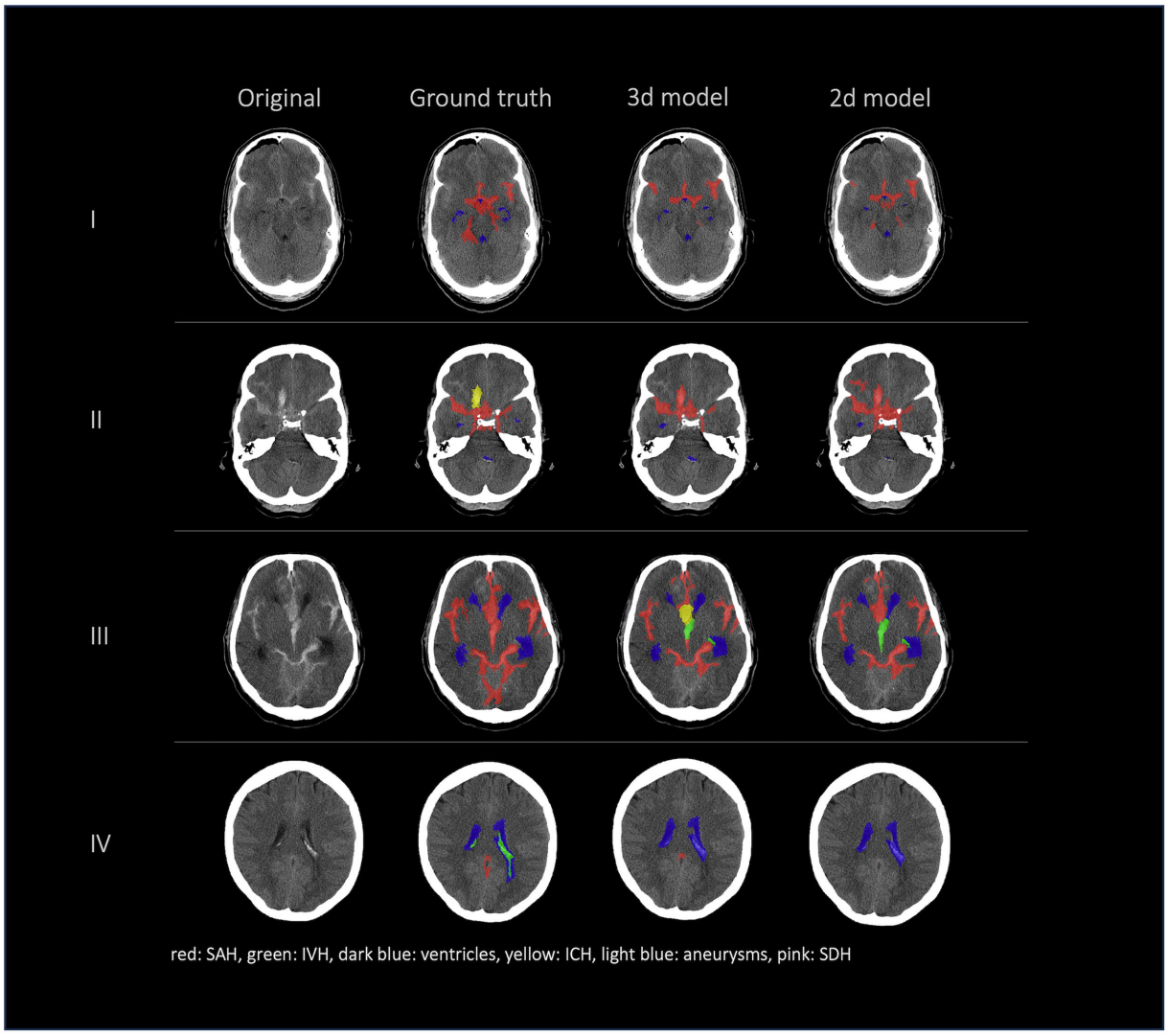

FIGURE 4
Failure cases, segmentations from left to right: first column shows original data, second column shows the ground truth (segmentation of Rater 1), third and fourth columns show the segmentations of our nnU-Net models (segmentation 3d model/segmentation 2d model). This figure showcases the most representative instances of failure we encountered. Once more, it underscores that the challenge does not primarily reside in detecting hemorrhage but rather in accurately distinguishing between various types of hemorrhages.

most likely due to the presence of SAH in our internal dataset making the labeling of the ICH more challenging. This phenomenon was also reported in the INSTANCE Grand challenge where SAH achieved the worst results compared to all the other intracranial hemorrhages that were evaluated (SDH, epidural hematoma (EDH), IPH, IVH) (20). Wu et al. also reported the best results for multiclass segmentation for IVH and ICH and an inferior performance for SAH, SDH and EDH (25). This highlights the challenging nature of distinguishing between different hemorrhage subtypes in real world cases of SAH.

Hemorrhage segmentation can also become more difficult in patients with shunts and metal artifacts (Appendix Figure 7) due to previous bleeding or intensive-care-unit (ICU)-monitoring. Shunt artifacts could be misinterpreted as ventricles and metal artifacts might resemble SAH (Appendix Figure 7). Being aware of potentially inferior performance in scans with artifacts, we include those patients to improve model generalizability and robustness in cases where a metal artifact or shunt was present.

The ventricle class achieves the best results out of all classes most likely because of its consistent shape and location and because it is the most represented class in the dataset. It is important to consider the size and shape of ventricles in SAH patients because they can be altered due to increased intracranial pressure and provide valuable information concerning the patient's outcome, e.g., enlarged ventricles can be a sign of increased intracranial pressure which can lead to decrease of cerebral perfusion pressure and therefore an unfavorable outcome (23). Ventricles as one of the most significant anatomical structures in SAH can serve as an anatomical orientation for downstream outcome prediction models. Therefore, the segmentation of the ventricles is relevant and can be considered an advantage of our model compared to already published automated multiclass segmentation models (20, 25).

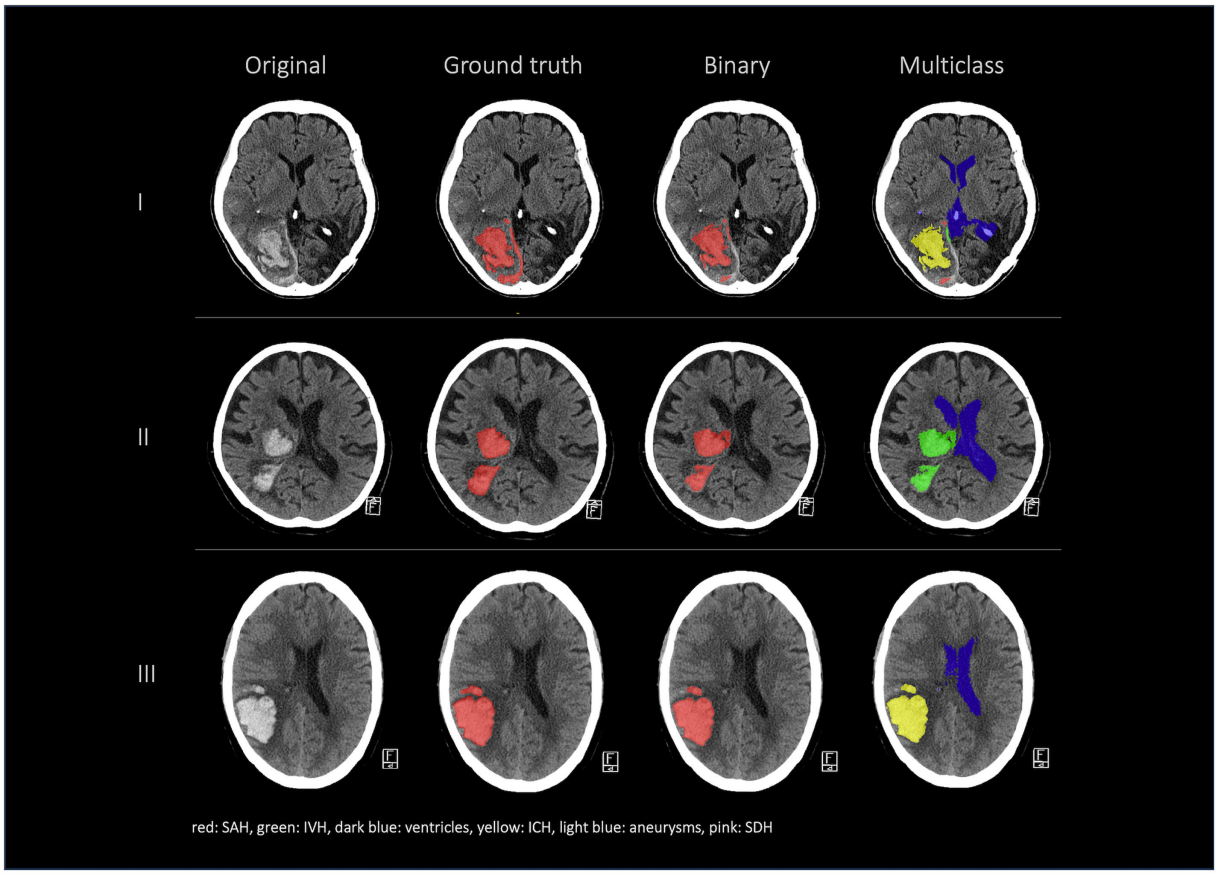

FIGURE 5
External validation. Comparison of segmentations in the following order from left to right: first column shows the original data, the second column shows the hemorrhage segmentation that was done at an external institution by a human rater (ground truth), the third column shows our 3d model binary segmentation and the fourth column shows our 3d model multiclass segmentation.

TABLE 5 Agreement/disagreement.

	Agreement		Disagreement		Class not present	
	SAH	IVH//ICH// Aneurysm//SDH	SAH	IVH//ICH// Aneurysm//SDH	SAH	IVH//ICH// Aneurysm//SDH
3d_fullres	20	8//5//3//1	0	3//4//1//1	0	9//11//16//18
2d	20	8//5//2//1	0	4//4//2//1	0	8//11//16//18
Rater 2	18	9//4//3//2	2	2//3//2//1	0	9//13//15//17

Agreement/disagreement 3d model/2d model/Rater 2. The first column illustrates the agreements in labeling the different classes between the ground truth (Rater 1) and the 3d model (first row), 2d model (second row), Rater 2 (third row). The second column shows the disagreements in labeling the classes and the third column shows the number of NCCT where a certain class was not present, therefore we cannot assess the agreement/disagreement.

4.2 2d/3d model

The 2d and 3d model architectures performed similarly based on quantitative assessments by global Dice coefficient and volumetric similarity. However, there were some locations (Figure 2: aneurysm, SDH) where it is difficult for the model to distinguish the type of bleeding looking only at one axial slice. The use of 2d models can be more advantageous for generalization to volumes with various slice thicknesses because 3d models either require resampling or might show poor performance when the voxel spacings are significantly different on the test sets. However, as stated in the INSTANCE Grand challenge, directly utilizing 2d networks would lead to a loss of significant context information among slices (20). Additionally, the spatial distribution of SAH in three dimensions is not considered in most of the known radiological scores but can be important for outcome prediction.

4.3 Current use of deep learning-based segmentation

Radiological features of hemorrhages in aSAH patients provide valuable outcome relevant information that has not been exploited to date. Despite the general acceptance that the volume of blood after

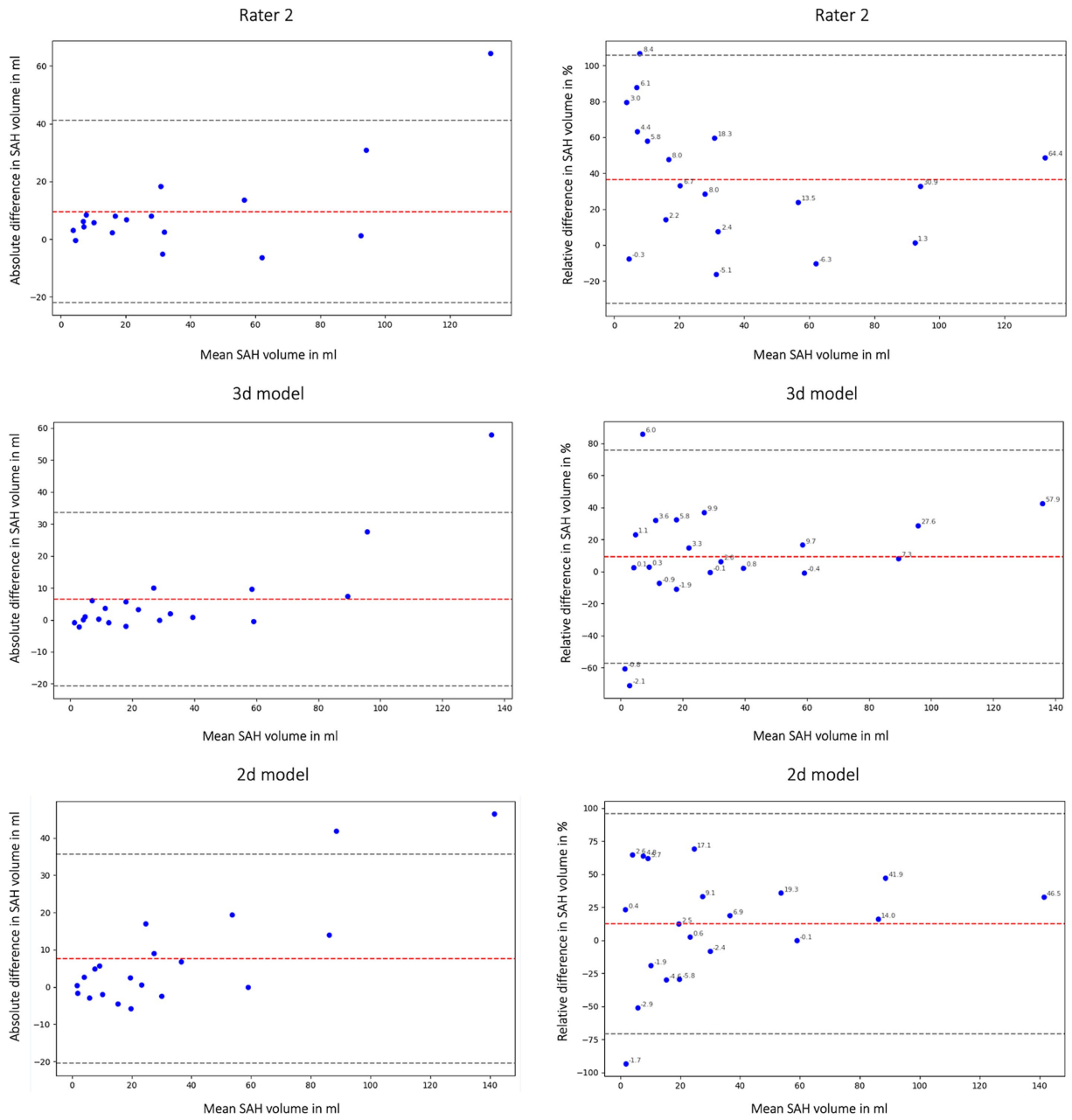

FIGURE 6
Bland Altman Plot absolute (first column) and relative (second column) difference in segmented subarachnoid hemorrhage (SAH) volume. Comparison of segmented subarachnoid hemorrhage volume in mL and % of Rater 2, 3d model and 2d model. Absolute difference in SAH volume: total bias Rater 2: 9.54 mL; total bias 3d model: 6.45 mL; total bias 2d model: 7.59 mL. The average total bias for SAH volume is 7.86 mL which means that Rater 1 (ground truth) segmented 7.86 mL more SAH volume on average than Rater 2, the 3d model and the 2d model. This phenomenon is also shown in Figures 2, 4.

SAH is prognostic for outcome and provides guidance for treatment decisions, no method to estimate the volume and distribution patterns of subarachnoid blood has been successfully implemented in clinical practice so far.

For a successful translation of artificial intelligence (AI) models into clinical practice there is a need to not only assess the stand-alone performance of models but more so focus on the outcomes when these algorithms are used as assistive tools in clinical practice (32). In our study, achieving performance levels comparable to human assessment based on quantitative metrics is feasible across most classes, despite the constraints of limited data. However, what is interesting for future investigations is assessing the model's efficacy when utilizing segmentation outputs in a downstream task aimed at predicting outcomes in patients with SAH. There are a few examples for radiomics-based classification of intracranial pathologies, e.g., intracranial aneurysm rupture or SAH prognosis prediction (33–35). In future works, the extraction of radiomics features based on segmentation masks might allow the extraction of imaging biomarkers predictive of SAH outcome. Deep learning-based segmentation models presented in this study would not only allow

to distinguish between different types of hemorrhages and healthy brain tissue, but also give precise information about the exact volumes of different structures and the impact on the patient's outcome.

There are a few examples of automated segmentation in NCCT in patients with SAH or ICH as leading pathology (16, 20, 36). The different models achieved Dice scores that were close to human level performance via volume quantification for SAH based on density quantification (16) or intracerebral hemorrhage segmentation via viola-U-Net ("Voxels-Intersecting Along Orthogonal Levels Attention U-Net") (36) in the INSTANCE Grand Challenge 2022. Even though the studies included patients with different subtypes of intracranial bleeding, only a recent multiclass model from Wu et al. was able to distinguish between different subtypes of intracranial hemorrhage (25). However, since anatomical changes, e.g., altered shape of ventricles, can be a negative predictor for the patient's outcome, it is important to assess the shape and volume of the ventricles as well and not only report the different types of hemorrhages.

Co-occurring pathologies can change the outcome of SAH-patients drastically for the worse, creating the need for a tool that is able to distinguish between the different subtypes of hemorrhage and can segment anatomical structures like ventricles (23, 24). Automated multiclass segmentation in NCCT was only tested on patients with ICH and intracranial hemorrhage in general so far (25, 37, 38). In contrast, our proposed model is developed specifically for patients with aneurysmal SAH, highlighting a key distinction of our work.

4.4 Study limitations

Our study has several limitations. First, the number of patients in our dataset was limited due to the time-consuming manual segmentation step and resulted in the underrepresentation of some classes in the training set as well as in the validation and test set (e.g., aneurysm, SDH), which could lead to an inferior performance of the models for these classes. Additionally, to validate the performances of the nnU-Net models, further training with a larger dataset may be beneficial. Second, the model is trained mostly on patients with subarachnoid hemorrhage to quantify the volume and distribution of hemorrhage subtypes. Hence, the models are not directly applicable for classifying CTs whether they contain hemorrhage or healthy brain tissue.

Finally, our external validation set does not include primarily SAH patients. Despite being tested on a different patient population our model demonstrated considerable generalization. Future works should test the performance of our model in segmentation challenges and public multiclass cerebral hemorrhage segmentation datasets. Validating the model across more diverse patient populations, particularly those with varying distributions of hemorrhage subtypes, could help address generalizability concerns. Additional external validation is essential before clinical adaptation of models from our exploratory study.

In this study we aim to compare our models' performance to human benchmarks which we created by assessing the performance of a different human rater (Rater 2) other than the ground truth (Rater 1).

The comparison to human level performance is likely more relevant in the clinical context but needs presumably more than 2 different human raters and a consensus-based ground truth creation to validate the comparison of the performances of humans and our nnU-Net models.

4.5 Future work

Our work has implications for future works in SAH segmentation and outcome prediction. Traditional scores lack interrater and intrarater reliability and often create disagreements between different reviewers. All known radiological scores calculate the amount of blood via hemorrhage thickness estimations in different areas in NCCT which is either roughly precise [modified Fisher scale, Hijdra Scale (7, 9)] or only applicable for symmetrical hemorrhages [ABC/2 Scale (39)], which leads to a low interrater reliability. A high interrater reliability ensures consistency in the segmentations irrespective of different raters, it provides a measure of agreement among human raters and is serving as a benchmark for evaluating the performance of automated methods. In clinical context, it ensures quality control, reproducibility and reduces bias. Because in clinical context the exact determination of the hemorrhage volume and severity of the hemorrhage is crucial for the treatment decision and outcome of the patient (23), there is a strong clinical need for a score or model that is reliable and shows consistent performance regardless of qualification or experience of the clinical personnel.

The prognostic value of different traditional radiological scores can be increased by an automated calculation that is supported by deep learning methods like our model or other radiomics approaches (35, 40). Additionally, the multiclass nature of our proposed model can be utilized in future studies to analyze the correlation of outcomes with co-occurrence of hemorrhage classes, their spatial distribution, hemorrhage volume and radiomics signatures. Our framework can be used to segment a large number of patients, enabling the processing of extensive patient data while being as precise as manual segmentation without being time-consuming.

With the help of deep learning there is the possibility to develop novel scores and biomarkers based on the exact imaging information. Deep learning-based automated semantic segmentation can facilitate the extraction of volumetric information from CT scans and enable an objective assessment of SAH severity. Development of novel prognostic approaches that overcome limitations of traditional scores—pending further validation—can enable rater-independent outcome prediction. These approaches can result in the extraction and utilization of more radiological data from CT scans, providing additional information that enables outcome prediction and facilitates decision support for personalized treatment decisions to improve patient outcomes.

5 Conclusion

Our deep learning-based nnU-Net-model demonstrated a performance close to the human benchmark and achieved accurate segmentation of SAH and SAH-related pathologies. This can be the starting point for automation of traditional radiological scores, correlation analysis of outcomes with co-occurrence of hemorrhage

classes and development of novel, prognostic scores for predicting outcomes in SAH. Deep learning can overcome the significant limitations of interrater and intrarater variability and provide an efficient solution to better exploit outcome-relevant image information than traditional scores. Our open-source model can enable analyses of large multicentric datasets to further improve performance and explore generalization.

Taken together, our study demonstrates the potential of deep learning to improve patient outcomes by advancing radiological examination of SAH and other intracranial hemorrhages.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

The studies involving humans were approved by the Ethics Committees of Charité© University Hospital and Kaisei Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin due to the retrospective nature of the study.

Author contributions

JK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. OA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing – review & editing. AH: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Software, Supervision, Validation, Writing – review & editing. MG: Formal analysis, Validation, Writing – review & editing. AN: Formal analysis, Software, Writing – review & editing. AK: Formal analysis, Supervision, Validation, Writing – review & editing. PV: Resources, Writing – review & editing. ST: Data curation, Formal analysis, Writing – review & editing. FI: Data curation, Writing – review & editing. ND: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Supervision, Writing – review & editing. DF: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work has received funding from the European Commission through Horizon Europe grant VALIDATE (Grant No. 101057263, coordinator: DF). This funding had no role in study design, data collection and analysis, decision to publication, or preparation of the manuscript.

Acknowledgments

Computation has been performed on the HPC for Research cluster of the Berlin Institute of Health.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2024.1490216/full#supplementary-material

References

1. Macdonald RL, Schweizer TA. Spontaneous subarachnoid haemorrhage. *Lancet*. (2017) 389:655–66. doi: 10.1016/S0140-6736(16)30668-7

2. Claassen J, Park S. Spontaneous subarachnoid haemorrhage. *Lancet Lond Engl*. (2022) 400:846–62. doi: 10.1016/S0140-6736(22)00938-2

3. Passier PECA, Visser-Meily JMA, van Zandvoort MJE, Post MWM, Rinkel GJE, van Heugten C. Prevalence and determinants of cognitive complaints after aneurysmal subarachnoid hemorrhage. *Cerebrovasc Dis Basel Switz*. (2010) 29:557–63. doi: 10.1159/000306642

4. Rinkel GJE. Management of patients with aneurysmal subarachnoid haemorrhage. *Curr Opin Neurol*. (2016) 29:37–41. doi: 10.1097/WCO.0000000000000282

5. Ahn S-H, Savarraj JP, Pervez M, Jones W, Park J, Jeon S-B, et al. The subarachnoid hemorrhage early brain edema score predicts delayed cerebral ischemia and clinical outcomes. *Neurosurgery*. (2018) 83:137–45. doi: 10.1093/neuros/nyx364

6. Graeb DA, Robertson WD, Lapointe JS, Nugent RA, Harrison PB. Computed tomographic diagnosis of intraventricular hemorrhage. Etiology and prognosis. *Radiology*. (1982) 143:91–6. doi: 10.1148/radiology.143.1.6977795

7. Hijdra A, Brouwers PJ, Vermeulen M, van Gijn J. Grading the amount of blood on computed tomograms after subarachnoid hemorrhage. *Stroke*. (1990) 21:1156–61. doi: 10.1161/01.str.21.8.1156

8. Wilson DA, Nakaji P, Abla AA, Uschold TD, Fusco DJ, Oppenlander ME, et al. A simple and quantitative method to predict symptomatic vasospasm after subarachnoid hemorrhage based on computed tomography: beyond the fisher scale. *Neurosurgery*. (2012) 71:869–76. doi: 10.1227/NEU.0b013e318267360f

9. Frontera JA, Claassen J, Schmidt JM, Wartenberg KE, Temes R, Connolly ES, et al. Prediction of symptomatic vasospasm after subarachnoid hemorrhage: the modified fisher scale. *Neurosurgery*. (2006) 59:21–27; discussion 21–27. doi: 10.1227/01.neu.0000243277.86222.6c

10. Woo PYM, Tse TPK, Chan RSK, Leung LNY, Liu SKK, Leung AYT, et al. Computed tomography interobserver agreement in the assessment of aneurysmal subarachnoid hemorrhage and predictors for clinical outcome. *J Neurointerventional Surg*. (2017) 9:1118–24. doi: 10.1136/neurintsurg-2016-012576

11. Ibrahim GM, Weidauer S, Macdonald RL. Interobserver variability in the interpretation of computed tomography following aneurysmal subarachnoid hemorrhage. *J Neurosurg*. (2011) 115:1191–6. doi: 10.3171/2011.7.JNS11725

12. Kramer AH, Hehir M, Nathan B, Gress D, Dumont AS, Kassell NF, et al. A comparison of 3 radiographic scales for the prediction of delayed ischemia and prognosis following subarachnoid hemorrhage. *J Neurosurg*. (2008) 109:199–207. doi: 10.3171/JNS/2008/109/8/0199

13. Degen LAR, Dorhout Mees SM, Algra A, Rinkel GJE. Interobserver variability of grading scales for aneurysmal subarachnoid hemorrhage. *Stroke*. (2011) 42:1546–9. doi: 10.1161/STROKEAHA.110.601211

14. Melinosky C, Kincaid H, Claassen J, Parikh G, Badjatia N, Morris NA. The modified fisher scale lacks interrater reliability. *Neurocrit Care*. (2021) 35:72–8. doi: 10.1007/s12028-020-01142-8

15. Barros RS, Van Der SWE, Boers AMM, IjJ Z, Van Den BR, Youssoufi WE, et al. Automated segmentation of subarachnoid hemorrhages with convolutional neural networks. *Inform Med Unlocked*. (2020) 19:100321. doi: 10.1016/j.imu.2020.100321

16. Boers AM, Zijlstra IA, Gathier CS, van den Berg R, Slump CH, Marquering HA, et al. Automatic quantification of subarachnoid hemorrhage on noncontrast CT. *AJNR Am J Neuroradiol*. (2014) 35:2279–86. doi: 10.3174/ajnr.A4042

17. Street JS, Pandit AS, Toma AK. Predicting vasospasm risk using first presentation aneurysmal subarachnoid hemorrhage volume: a semi-automated CT image segmentation analysis using ITK-SNAP. *PLoS One*. (2023) 18:e0286485. doi: 10.1371/journal.pone.0286485

18. van der Steen WE, Marquering HA, Ramos LA, van den Berg R, Coert BA, Boers AMM, et al. Prediction of outcome using quantified blood volume in aneurysmal SAH. *AJNR Am J Neuroradiol*. (2020) 41:1015–21. doi: 10.3174/ajnr.A6575

19. Yuan JY, Chen Y, Jayaraman K, Kumar A, Zlepper Z, Allen ML, et al. Automated quantification of compartmental blood volumes enables prediction of delayed cerebral ischemia and outcomes after aneurysmal subarachnoid hemorrhage. *World Neurosurg*. (2023) 170:e214–22. doi: 10.1016/j.wneu.2022.10.105

20. Li X, Luo G, Wang K, Wang H, Liu J, Liang X, et al. The state-of-the-art 3D anisotropic intracranial hemorrhage segmentation on non-contrast head CT: the INSTANCE challenge. (2023). Available at: http://arxiv.org/abs/2301.03281 (Accessed May 26, 2024).

21. Li X, Luo G, Wang W, Wang K, Gao Y, Li S. Hematoma expansion context guided intracranial hemorrhage segmentation and uncertainty estimation. *IEEE J Biomed Health Inform*. (2022) 26:1140–51. doi: 10.1109/JBHI.2021.3103850

22. Xie Y, To M-S, Wang C, Wang D, Wu Q, Xia Y, et al. Multi-class brain hemorrhage segmentation in non-contrast computed tomography under limited annotations. (2024). doi: 10.5281/zenodo.10979177

23. Venti M. Subarachnoid and intraventricular hemorrhage. *Front Neurol Neurosci*. (2012) 30:149–53. doi: 10.1159/000333625

24. Wan A, Jaja BNR, Schweizer TA, Macdonald RL. On behalf of the SAHIT collaboration. Clinical characteristics and outcome of aneurysmal subarachnoid hemorrhage with intracerebral hematoma. *J Neurosurg*. (2016) 125:1344–51. doi: 10.3171/2015.10.JNS151036

25. Wu B, Xie Y, Zhang Z, Ge J, Yaxley K, Bahadir S, et al. BHSD: a 3D multi-class brain hemorrhage segmentation dataset. (2023). Available at: http://arxiv.org/abs/2308.11298 (Accessed May 26, 2024).

26. Neuroimaging in Python Team. Neuroimaging in Python – Pipelines and interfaces — nipy pipeline and interfaces package. Available at: https://nipype.readthedocs.io/en/latest/api/generated/nipype.interfaces.dcm2nii.html (Accessed May 26, 2024).

27. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. (2012) 30:1323–41. doi: 10.1016/j.mri.2012.05.001

28. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. (2006) 31:1116–28. doi: 10.1016/j.neuroimage.2006.01.015

29. 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) 18:203–11. doi: 10.1038/s41592-020-01008-z

30. Hilbert A, Madai VI, Akay EM, Aydin OU, Behland J, Sobesky J, et al. BRAVE-NET: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease. *Front Artif Intell*. (2020) 3:552258. doi: 10.3389/frai.2020.552258

31. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. (2015) 15:29. doi: 10.1186/s12880-015-0068-x

32. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. *N Engl J Med*. (2023) 388:1981–90. doi: 10.1056/NEJMra2301725

33. Alwalid O, Long X, Xie M, Yang J, Cen C, Liu H, et al. CT angiography-based radiomics for classification of intracranial aneurysm rupture. *Front Neurol*. (2021) 12:619864. doi: 10.3389/fneur.2021.619864

34. Tong X, Feng X, Peng F, Niu H, Zhang B, Yuan F, et al. Morphology-based radiomics signature: a novel determinant to identify multiple intracranial aneurysms rupture. *Aging*. (2021) 13:13195–210. doi: 10.18632/aging.203001

35. Shan D, Wang J, Qi P, Lu J, Wang D. Non-contrasted CT Radiomics for SAH prognosis prediction. *Bioeng Basel Switz*. (2023) 10:967. doi: 10.3390/bioengineering10080967

36. Liu Q, MacIntosh BJ, Schellhorn T, Skogen K, Emblem K, Bjørnerud A. Voxels intersecting along orthogonal levels attention U-net for intracerebral haemorrhage segmentation in head CT. (2023). Available at: http://arxiv.org/abs/2208.06313 (Accessed May 26, 2024).

37. Vogt E, Vu LH, Cao H, Speth A, Desser D, Schlunk F, et al. Multilesion segmentations in patients with intracerebral hemorrhage: reliability of ICH, IVH and PHE masks. *Tomogr Ann Arbor Mich*. (2023) 9:89–97. doi: 10.3390/tomography9010008

38. Phaphuangwittayakul A, Guo Y, Ying F, Dawod AY, Angkurawaranon S, Angkurawaranon C. An optimal deep learning framework for multi-type hemorrhagic lesions detection and quantification in head CT images for traumatic brain injury. *Appl Intell Dordr Neth*. (2022) 52:7320–38. doi: 10.1007/s10489-021-02782-9

39. Webb AJS, Ullman NL, Morgan TC, Muschelli J, Kornbluth J, Awad IA, et al. Accuracy of the ABC/2 score for intracerebral hemorrhage: systematic review and analysis of MISTIE, CLEAR-IVH, and CLEAR III. *Stroke*. (2015) 46:2470–6. doi: 10.1161/STROKEAHA.114.007343

40. Huang X, Wang D, Ma Y, Zhang Q, Ren J, Zhao H, et al. Perihematomal edema-based CT-radiomics model to predict functional outcome in patients with intracerebral hemorrhage. *Diagn Interv Imaging*. (2023) 104:391–400. doi: 10.1016/j.diii.2023.04.008