BMC Medical Imaging

RESEARCH

Open Access

Automated segmentation of brain metastases in T1-weighted contrast-enhanced MR images pre and post stereotactic radiosurgery

Hemalatha Kanakarajan[1*], Wouter De Baene[1*], Patrick Hanssens[2,3] and Margriet Sitskoorn[1]

Abstract

Background and purpose Accurate segmentation of brain metastases on Magnetic Resonance Imaging (MRI) is tedious and time-consuming for radiologists that could be optimized with deep learning (DL). Previous studies assessed several DL algorithms focusing only on training and testing the models on the planning MRI only. The purpose of this study is to evaluate well-known DL approaches (nnU-Net and MedNeXt) for their performance on both planning and follow-up MRI.

Materials and methods Pre-treatment brain MRIs were retrospectively collected for 255 patients at Elisabeth-TweeSteden Hospital (ETZ): 201 for training and 54 for testing, including follow-up MRIs for the test set. To increase heterogeneity, we added the publicly available MRI scans from the Mathematical oncology laboratory of 75 patients to the training data. The performance was compared between the two models, with and without the addition of the public data. To statistically compare the Dice Similarity Coefficient (DSC) of the two models trained on different datasets over multiple time points, we used Linear Mixed Models.

Results All models obtained a good DSC (DSC >= 0.93) for planning MRI. MedNeXt trained with combined data provided the best DSC for follow-ups at 6, 15, and 21 months (DSC of 0.74, 0.74, and 0.70 respectively) and jointly the best DSC for follow-ups at three months with MedNeXt trained with ETZ data only (DSC of 0.78) and 12 months with nnU-Net trained with combined data (DSC of 0.71). On the other hand, nnU-Net trained with combined data provided the best sensitivity and FNR for most follow-ups. The statistical analysis showed that MedNeXt provides higher DSC for both datasets and the addition of public data to the training dataset results in a statistically significant increase in performance in both models.

Conclusion The models achieved a good performance score for planning MRI. Though the models performed less effectively for follow-ups, the addition of public data enhanced their performance, providing a viable solution to improve their efficacy for the follow-ups. These algorithms hold promise as a valuable tool for clinicians for automated segmentation of planning and follow-up MRI scans during stereotactic radiosurgery treatment planning and response evaluations, respectively.

*Correspondence:
Hemalatha Kanakarajan
h.kanakarajan@tilburguniversity.edu
Wouter De Baene
W.DeBaene@tilburguniversity.edu

Full list of author information is available at the end of the article

Clinical trial number Not applicable.

Keywords Brain metastases, Auto-segmentation, Deep learning, nnU-Net, MedNeXt

Background

Brain metastases (BM) are the predominant intracranial tumors seen in adults [1]. It is estimated that about one-fifth of all cancer patients will ultimately develop BM [2]. Advancements in primary tumor treatments have increased life expectancy and hence the probability of developing BM [1]. The presence of BM is associated with a substantial increase in morbidity and mortality rates among cancer patients [3]. Stereotactic Radiosurgery (SRS) is a treatment option in which the BM are targeted very precisely, whereby the dose of radiation to the healthy brain tissue is limited. SRS has emerged as a well accepted treatment modality in the current standard of care for the treatment of BM [4].

For SRS treatment planning and treatment response evaluation, the physician must manually delineate numerous lesions on three-dimensional Magnetic Resonance Imaging (MRI) scans. This manual process is labor-intensive and prone to considerable variability among physicians [5]. Rudie et al. [5] compared the manual segmentations of two neuroradiologists and provided a baseline for interrater reliability. The interrater Dice Similarity Coefficient (DSC) was 0.83 ± 0.02 on their testing dataset. Introducing an automatic and reliable system for detecting and delineating BM could facilitate more precise treatment delivery in the radiotherapy clinic. Automated tools that assist radiologists and radiation oncologists can positively influence both efficiency and efficacy in detecting and delineating multiple metastases.

Deep learning (DL) models have shown great promise in medical image analysis, particularly in detection, segmentation and classification tasks with the potential to improve clinical workflow [6].

There are plenty of studies on automated segmentation of primary tumors using DL algorithms [7, 8, 9, 10, 11]. Several approaches have also been introduced for BM segmentation on MRI using DL [12]. In 2015, Losch et al. [13] produced state-of-the-art results in automated segmentation of BM on MRI using deep convolutional networks. Since then, a large variety of network architectures for DL such as Convolutional Neural Networks (CNNs) [14] and DeepMedic [15] have been tested. However, a notable limitation of these studies is their exclusive focus on training and testing the models solely on the planning MRI (e.g [14, 16, 17, 18, 19, 20, 21])., potentially overlooking variations in the performance of the DL algorithms when applied to follow-up MRI. Such discrepancies may arise due to radiation-induced shrinkage of the tumors. It is imperative to assess the performance of DL algorithms on the follow-up MRI to ascertain their utility in assisting the clinicians in the response evaluation during follow-ups. A thorough examination of the follow-up scans is essential for assessing treatment response and the current disease status in the brain. Key challenges for the clinicians during follow-up scans include classifying BMs as progressing, stable, or recurring, separating new BMs from already treated BMs, and differentiating recurrent metastases from delayed radiation necrosis and treatment effects [22]. Manual methods are time-intensive and could lead to errors if the follow-up images are not co-registered and radiation plan information is not overlayed on the images [23]. There are some commercially available software platforms to track BMs on follow-ups, but these softwares do not provide an automated solution, and they are not reliable for patients with multiple BMs [22, 23]. Evaluating the performance of the deep learning algorithms for follow-ups will help to establish whether these algorithms can be used to automate the detection and segmentation of the BMs for the follow-ups. Jalalifar et al. [22] evaluated the performance of a DL model on follow-up MRI but their study provided the performance results for only five sample patients. Similarly, the study of Lu et al. [24] for assessing the use of AI for the tracking of BMs on follow-ups was done with only three patients. Prezelski at al. [22] and Hsu et al. [22] proposed a deep learning based system for the automated detection of the BMs in the follow-ups based on a previous deep learning model developed by Hsu et al. [19]. However, these studies only evaluated the performance of their model on the longest 3D diameter without evaluating the overlap of the segmentations with the manual segmentations. However, response assessment of BMs is improved when using volumetric criteria (for which the tumor is evaluated as a 3D structure) compared to relying on a simple diameter estimate [26]. For this, the overlap of the complete 3D segmentation is crucial.

One of the popular DL network architectures is the so-called nnU-Net [27]. Isensee et al. [28] demonstrated how this architecture achieved state of the art performance on different challenges in medical image segmentation by applying it to 10 international biomedical image segmentation challenges comprising 19 different datasets and 49 segmentation tasks across a variety of organs, organ substructures, tumors, lesions and cellular structures in MRI, CT and electron microscopy images. Ziyaee et al. [29] evaluated the effectiveness of this algorithm specifically for segmentation of BM by training and testing it with planning MRI only. The model achieved an overall DSC of 82.2%, which shows good segmentation performance. Recently, the transformer technique

[30] has emerged as a noteworthy alternative to traditional CNNs in the medical domain, being employed for various tasks like classification, detection, and segmentation [31, 32, 33]. This has posed a significant challenge to existing CNN-based solutions [34] like nnU-Net.

Both approaches, however, have their own advantages and limitations. CNNs can accurately segment tumors by analyzing local details in the images [35]. But, CNN-based approaches generally exhibit limitations for modeling long-range dependencies. On the other hand, transformers are effective in considering the broader context of the entire images, but can result in limited localization abilities due to lack of detailed localization information [36]. Some network architectures now incorporate both convolutional layers and transformers to leverage the strengths of both approaches, aiming for improved performance and overcoming the limitations of each individual architecture [37]. An example is ConvNeXt [38] which combines the strenghts of both approaches. Building upon this, Roy et al. [39] introduced MedNeXt, a modernised and scalable convolutional architecture customised to challenges of data-scarce medical settings.

Compared to other algorithms, the nnU-Net and MedNeXt algorithms achieved better segmentation performance [39]. MedNeXt achieved state-of-the-art performance benefits on segmentation tasks of varying modality and sizes and hence Roy et al. [39] proposed MedNeXt as a strong and modernized alternative to standard ConvNets like nnU-Net for building deep networks for medical image segmentation. MedNeXt achieved this performance against baselines consisting of Transformer-based, convolutional and large kernel networks. However, the effectiveness of the nnU-Net and MedNeXT algorithms specifically for the segmentation of the BM follow-up images has not yet been evaluated.

The present study aims to bridge this gap by assessing the applicability of these state-of-the-art algorithms for automated segmentation of both planning and follow-up BM images. Typically, in most hospitals, segmentation of BM is performed solely on the planning MRI scans and not on the follow-up scans. This absence of segmented images poses a challenge for training DL algorithms with follow-up scans. In this research, we evaluated the performance of the nnU-Net and MedNeXt algorithms by training them with planning images and publicly available segmented images, then testing them on both planning and follow-up images. We used publicly available BM images and added these images to the training data to increase the heterogeneity of the training data. This evaluation will help to understand whether these state-of-the-art DL algorithms can assist the clinicians in detection and segmentation of BM images for treatment planning and treatment response evaluation during follow-ups.

Method

Data collection

This study was approved by the Elisabeth-TweeSteden Hospital (ETZ) science office and by the Ethics Review Board at Tilburg University (Reference number: RP548). Pre-treatment contrast-enhanced (with triple dose gadolinium) T1-weighted brain MRIs of 255 BM patients were used. Scans were made as part of clinical care at the Gamma Knife Center of the ETZ between 2015 and 2021 at Tilburg, The Netherlands. These planning MRI scans were collected using a 1.5T Philips Ingenia scanner (Philips Healthcare, Best, The Netherlands) with a contrast-enhanced T1-weighted sequence (TR/TE: 25/1.86 ms, FOV: $210 \times 210 \times 150$, flip angle: 30°, transverse slice orientation, voxel size: $0.82 \times 0.82 \times 1.5$ mm). For contrast enhancement, a total of 45 ml of a 0.5 mmol/ml gadolinium solution was administered, regardless of the weight of the patient. The contrast agent was given in three separate doses, with an interval of 4 to 5 min between each administration. Additionally, there was a 4 to 5 min delay between the administration of the final contrast dose and the acquisition of the contrast-enhanced T1w scan. The total of 255 patients were split into 201 patients for model training and 54 patients for testing. All the patients underwent Gamma Knife Radiosurgery (GKRS) at the Gamma Knife Center. The 54 patients who were part of the testing set are from the set of patients included in the Cognition And Radiation Study A (CAR-Study A) at ETZ [40]. Our test set is a random subset of patients included in this CAR-Study A. Patients with other brain tumor types (e.g. meningioma) in addition to BM were excluded from the training and test data sets. For all patients in the training and test data set, the baseline segmentations and the regions of interest around each enhancing metastatic lesion were manually delineated by an expert neuroradiologist at ETZ and were cross-checked by a second neuroradiologist. The tumors were outlined on each slice on the contrast-enhanced 3D T1-weighted sequence, using the Leksell GammaPlan software. We refer to these manually delineated segmentations as reference segmentations. The reference segmentations for follow-up scans were only available for the patients who were part of the CAR-Study A.

For the 54 patients used for testing, the post-treatment contrast-enhanced (with single dose gadolinium) T1-weighted follow-up MRI scans were also retrospectively collected using a slightly different scanning protocol (TR/TE: 25/4.6 ms, FOV: $230 \times 220 \times 168$, flip angle: 30°, transverse slice orientation, voxel size: $0.79 \times 0.79 \times 0.8$ mm). For a single dose, patients weighing between 70 and 100 kg received 5 ml of a solution of the

0.5 mmol/ml gadolinium solution. For patients weighing more than 100 kg, the dose was increased to 20 ml of the same solution.

The images from 6 follow-up (FU) sessions were available. The FU scans were made 3, 6, 9, 12, 15, and 21 months after treatment. For these follow-ups, scans of 54 (FU1), 41 (FU2), 32 (FU3), 27 (FU4), 19 (FU5) and 14 (FU6) patients were available.

Treatment data

GKRS was performed with a Leksell Gamma Knife (Elekta AB). All patients received a dose of 18–25 Gy with 99–100% coverage of the target. Dose limits for organs at risk were 18 Gy for the brainstem and 8 Gy for the optic chiasm and optic nerves.

Preprocessing

As a first preprocessing step, all the MRI scans were registered to standard MNI space using Dartel in SPM12 (Wellcome Trust Center for Neuroimaging, London, UK), implemented in Python (version 3.11) using the Nipype (Neuroimaging in Python–Pipelines and Interfaces) software package (version 1.8.6) [41]. The voxel size of the normalized image was set to $1 \times 1 \times 1$ mm. For all other normalization configurations, the default values offered by SPM12 were used. One other preprocessing step was to combine the multiple labels for patients with more than one BM in one single mask. FSL library (Release 6.0) was used for this integration [42].

Public data

We also used the publicly available BM images from the Mathematical oncology laboratory provided by Ocaña-Tienda et al. [43] and added these images to the training data for models trained with the combination of ETZ and public data. To the best of our knowledge, this is the only publicly available dataset that also includes follow-up MRI scans and segmentations. The other publicly available BM datasets contain only the planning MRI. Hence, we added only this public dataset to our training data. This data set contained 355 contrast-enhanced (with a single dose of contrast) T1-weighted planning and follow-up MRIs acquired using either General Electric, Philips or Siemens scanner for 75 patients. The voxel size for all scans for the x- and y-dimensions ranged from 0.39 mm to 1.01 mm. The median slice thickness was 1.30 mm. Similar to the scans from ETZ, all the MRI scans from this public data set were also registered to standard MNI space with a voxel size of $1 \times 1 \times 1$ mm. All the scans in the ETZ were made using a Philips Ingenia scanner. Also, the voxel size and the slice thickness of the scans in this public data set were different from the scans at ETZ. Moreover, the public scans were contrast enhanced with single dose while the planning scans at ETZ were contrast enhanced with triple dose. All these differences between the datasets increase the heterogeneity of our combined dataset.

Deep learning models

The nnU-Net algorithm, a framework built on top of the U-Net [27], makes key design decisions regarding preprocessing, post-processing, data augmentation, network architecture, training scheme, and inference, all tailored to the specific properties of the dataset at hand [27]. It analyzes the provided training cases and automatically configures a matching U-Net-based segmentation pipeline. These automatic design choices allow nnU-Net to perform well on many medical segmentation tasks. nnU-Net readily executes systematic rules to generate DL methods for previously unseen datasets without the need for further optimization [27]. The nnU-Net model was trained in 3d full resolution mode.

On the other hand, MedNeXt represents a novel approach to medical image segmentation, drawing inspiration from transformers. The architecture of MedNeXt includes ConvNeXt blocks, which are used for processing the image data. These blocks help in efficient sampling of the image [39]. MedNeXt also uses a novel technique to adjust the size of the processing units (kernels). MedNeXt is also customized to the challenges of sparsely annotated medical image segmentation datasets and is an effective modernization of standard convolution blocks for building deep networks for medical image segmentation [39]. MedNeXt offers four predefined architecture sizes (Small, Base, Medium, and Large) and two predefined kernel sizes (3*3*3, 5*5*5). As per the performance comparison by Roy et al. [39], the larger kernel sizes of MedNeXt comprehensively outperform its smaller kernel sizes for organ segmentations but in a more limited fashion for tumor segmentations. Also, considering that the larger kernel sizes of MedNeXt consume higher training time and our pilot testing with different sizes didn't show systematic differences, the combination we used for training the model was Small with 3*3*3 kernel size.

The different models that we created were.

1. nnU-Net trained with ETZ planning BM data only ($n = 201$).
2. nnU-Net trained with ETZ planning BM data and BM public data ($n = 556$).
3. MedNeXt trained with ETZ planning BM data only ($n = 201$).
4. MedNeXt trained with ETZ planning BM data and BM public data ($n = 556$).

Evaluation of models

We evaluated the performance of these models on both planning and follow-up MRI. To assess the quality of

the resulting segmentations, multiple metrics were employed. The DSC measures the overlap with the reference segmentations (ranging from 0 for no overlap to 1 for perfect overlap) per patient. It is calculated by dividing the double of the area of overlap by the sum of the areas of the predicted and the reference segmentation. The algorithm's performance in detecting individual metastases was measured by sensitivity (number of voxels in the detected metastases divided by the number of voxels in all metastases contained in the reference segmentation) and by the False Negative Rate (FNR). The FNR is the probability that a true metastasis will be missed by the model. In addition to DSC, we also report Intersection over Union (IoU) as a complementary metric for segmentation performance. IoU is calculated by dividing the area of overlap by the union of the areas of the predicted and the reference segmentation. Compared to DSC, IoU applies a stronger penalty to both under- and over-segmentation, making it particularly relevant for small BMs where precise delineation is critical [44]. In the results section, these metrics are presented for the predictions done for baseline and for the follow up test data.

Comparison of models

To determine whether there was a significant difference in performance between the four models and to understand the practical significance of the observed differences, we statistically compared the DSC scores of the different models trained on different datasets over multiple time points. We employed a Linear Mixed Model (LMM) in MATLAB to analyze the relationship between the dependent variable, DSC, and the predictors model type and dataset type and their interaction while modeling the individual differences and variations across time by including random intercepts for time and the interaction between subject and time. Statistical significance was set to $p <= 0.05$.

Results

Patient characteristics

Table 1 shows the patient characteristics from the ETZ and public data set included in our study. Figure 1 shows the BM volume distribution for the datasets.

Segmentation performance

The mean DSC (i.e., the overlap between the reference segmentation and the predicted segmentation) obtained for the baseline and the FU tests for the four models are shown in Table 2 and visually depicted in Fig. 2. For a patient, the models detected an extra tumor in the baseline test that was not contained in the baseline reference segmentation masks but was part of the FU1 reference segmentation masks. We updated the baseline reference segmentations to include the extra tumor, and hence, the below results also take into account this extra tumor.

All four models obtained a good DSC for planning MRI. The models 1, 3 and 4 had the same DSC of 0.94 for the planning MRI. For model 2, we obtained a slightly lower DSC of 0.93.

All four models obtained a lower DSC for the segmentation of follow-up MRI when compared with the DSC of planning MRI. Model 4 (MedNeXt trained with both ETZ and public data) and Model 3 (MedNeXt trained

Table 1 Characteristics of patients from ETZ and public data set

	ETZ training set	Public data training set	Baseline test set	FU1 test set	FU2 test set	FU3 test set	FU4 test set	FU5 test set	FU6 test set
Number of patients	201	75	54	54	41	32	27	19	14
Gender									
Male	97	28	26	26	19	15	12	7	5
Female	104	47	28	28	22	17	15	12	9
Age									
Average	63	57	62	62	63	63	63	62	62
Min	34	27	32	32	44	44	44	49	49
Max	85	77	81	81	81	81	81	81	75
Total number of brain lesion segmentations	1173	367	211	218	212	167	97	73	55
Number of BM per patient									
Average	6	5	4	4	4	4	4	3	3
Min	1	2	1	2	1	1	1	1	1
Max	154	16	10	10	10	10	10	10	10
Metastases volume (mm^3)									
Average	2551	7621	2739	1145	1219	900	1444	1257	1885
Min	1	2	1	1	1	1	1	1	2
Max	80,074	60,174	29,094	19,961	26,093	14,308	19,083	17,964	35,799
Median tumor volume (mm^3)	102	73	331	75	46	44	63	44	25

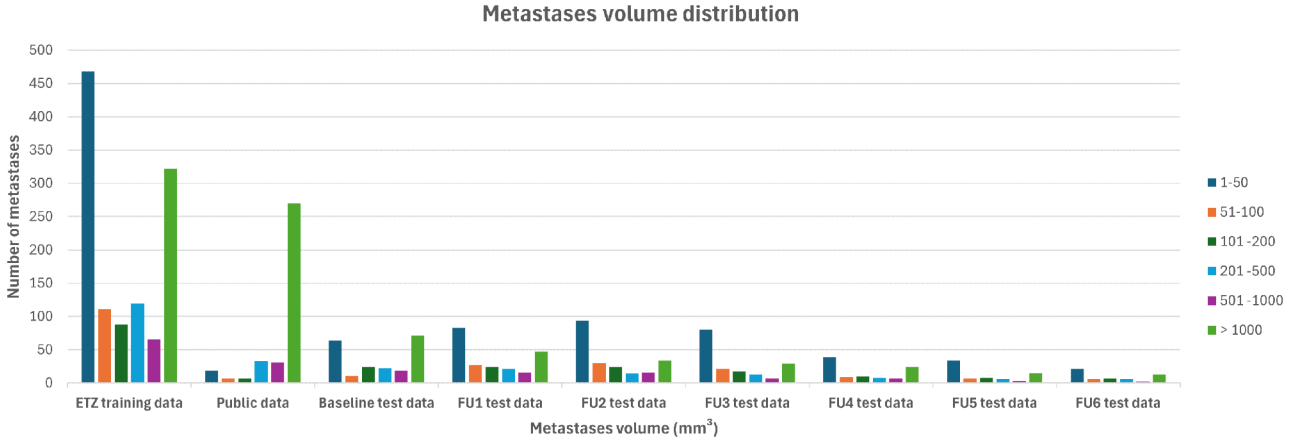

Fig. 1 Brain metastases volume distribution

Table 2 DSC of the segmentation models along with the standard deviation (SD)

DSC	Baseline	FU1	FU2	FU3	FU4	FU5	FU6
Model 1 – nnU-Net trained with ETZ data only	**0.94** ± 0.04	0.74 ± 0.31	0.63 ± 0.38	0.58 ± 0.36	0.53 ± 0.42	0.57 ± 0.36	0.47 ± 0.36
Model 2 - nnU-Net trained with ETZ and public data	0.93 ± 0.11	0.77 ± 0.20	0.73 ± 0.26	**0.73 ± 0.25**	**0.71 ± 0.29**	0.72 ± 0.28	0.69 ± 0.28
Model 3 - MedNeXt trained with ETZ data only	**0.94** ± 0.04	**0.78** ± 0.26	0.66 ± 0.37	0.65 ± 0.37	0.57 ± 0.42	0.67 ± 0.32	0.62 ± 0.34
Model 4 - MedNeXt trained with ETZ and public data	**0.94 ± 0.03**	**0.78 ± 0.19**	**0.74 ± 0.26**	0.72 ± 0.27	**0.71** ± 0.29	**0.74 ± 0.24**	**0.70 ± 0.28**

In bold is the value of the best DSC for the corresponding test set

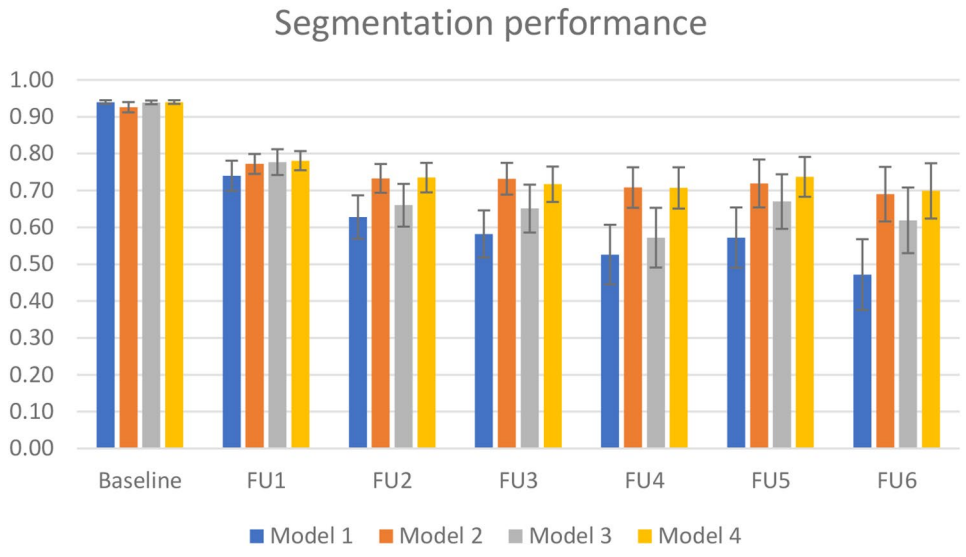

Fig. 2 DSC of the segmentation models along with Standard Error (SE)

with ETZ data only) had a DSC of 0.78 for FU1. This is higher than the DSC of other models for FU1. Similarly, model 4 had the highest DSC for FU2, FU5 and FU6. Model 2 (nnU-Net trained with both ETZ and public data) had a DSC of 0.73 for FU3. This is higher than the DSC of other models for FU3. Model 2 and Model 4 had a DSC of 0.71 for FU4. This is higher than the DSC of other models for FU4.

In general, the models which included the public data also in the training data set (model 2 and 4) performed better for the follow-ups when compared to the models which were trained with ETZ data only (model 1 and 3). For FU2, FU5 and FU6, the model 4 had a DSC which is 0.08, 0.07 and 0.08 respectively higher than the DSC of model 3. For FU3 and FU4, model 2 had a DSC of 0.15 and 0.18 respectively higher than the DSC of model 1 (nnU-Net trained with ETZ data only).

The FNR, Sensitivity and IOU obtained for the baseline and the FU tests for the four models are shown in Tables 3, 4 and 5 respectively.

These results show that Model 2 had a superior FNR and sensivity for the most follow-ups when compared to

Table 3 FNR of the segmentation models

FNR	Baseline	FU1	FU2	FU3	FU4	FU5	FU6
Model 1 – nnU-Net trained with ETZ data only	0.07	0.29	0.37	0.47	0.47	0.45	0.49
Model 2 - nnU-Net trained with ETZ and public data	**0.06**	**0.08**	**0.16**	0.17	**0.18**	**0.17**	**0.17**
Model 3 - MedNeXt trained with ETZ data only	**0.06**	0.21	0.35	0.34	0.42	0.36	0.35
Model 4 - MedNeXt trained with ETZ and public data	**0.06**	0.09	**0.16**	**0.16**	0.20	0.19	0.18

In bold is the value of the best FNR for the corresponding test set

Table 4 Sensitivity of the segmentation models

Sensitivity	Baseline	FU1	FU2	FU3	FU4	FU5	FU6
Model 1 – nnU-Net trained with ETZ data only	0.93	0.71	0.63	0.53	0.53	0.55	0.51
Model 2 - nnU-Net trained with ETZ and public data	**0.94**	**0.92**	**0.84**	0.83	**0.82**	**0.83**	**0.83**
Model 3 - MedNeXt trained with ETZ data only	**0.94**	0.79	0.65	0.66	0.58	0.64	0.65
Model 4 - MedNeXt trained with ETZ and public data	**0.94**	0.91	**0.84**	**0.84**	0.80	0.81	0.82

In bold is the value of the best sensitivity for the corresponding test set

Table 5 Intersection over Union (IoU) of the segmentation models

IoU	Baseline	FU1	FU2	FU3	FU4	FU5	FU6
Model 1 – nnU-Net trained with ETZ data only	**0.89**	0.65	0.55	0.50	0.46	0.48	0.38
Model 2 - nnU-Net trained with ETZ and public data	0.88	0.66	**0.63**	**0.62**	**0.61**	0.62	0.58
Model 3 - MedNeXt trained with ETZ data only	**0.89**	**0.69**	0.59	0.58	0.52	0.58	0.52
Model 4 - MedNeXt trained with ETZ and public data	**0.89**	0.67	**0.63**	0.61	**0.61**	**0.63**	**0.59**

In bold is the value of the best IoU for the corresponding test set

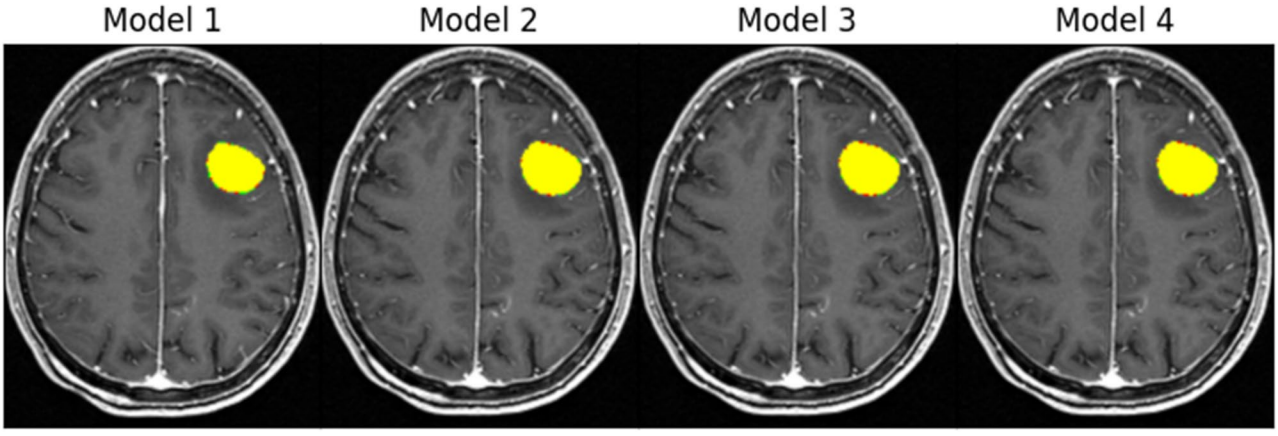

Fig. 3 Results of the 4 models for a sample patient in baseline test set

the other models. Similar to DSC, the models 2 and 4 had a superior IoU for most follow-ups.

Figure 3 shows the segmented output of the four models for a sample patient with a good segmentation performance in the baseline test set. The reference segmentations are marked in green, the segmentations from the models are marked in red and the overlapping region is marked in yellow. The DSC of all four models for this patient is 0.98. The images show that there is an almost complete overlap between the reference segmentations and the model output and provide a sample illustration of the good performance of all four models for the baseline test set.

Figure 4 shows the segmented output of the four models for a sample patient in the FU5 test set. The reference segmentations are marked in green, the segmentations from the models are marked in red and the overlapping region is marked in yellow. Model 4 is the best performaning model with a DSC of 0.85, followed by model 2 with a DSC of 0.83. The DSC of model 1 and model 3 are 0.63 and 0.60 respectively. The images show that model 4 and model 2 has a higher overlap of the segmented output and reference segmentation when compared to model 1 and model 3. The figure provides a sample illustration of better performance of the models trained with the combined data when compared with the models trained with ETZ data only for follow-up images.

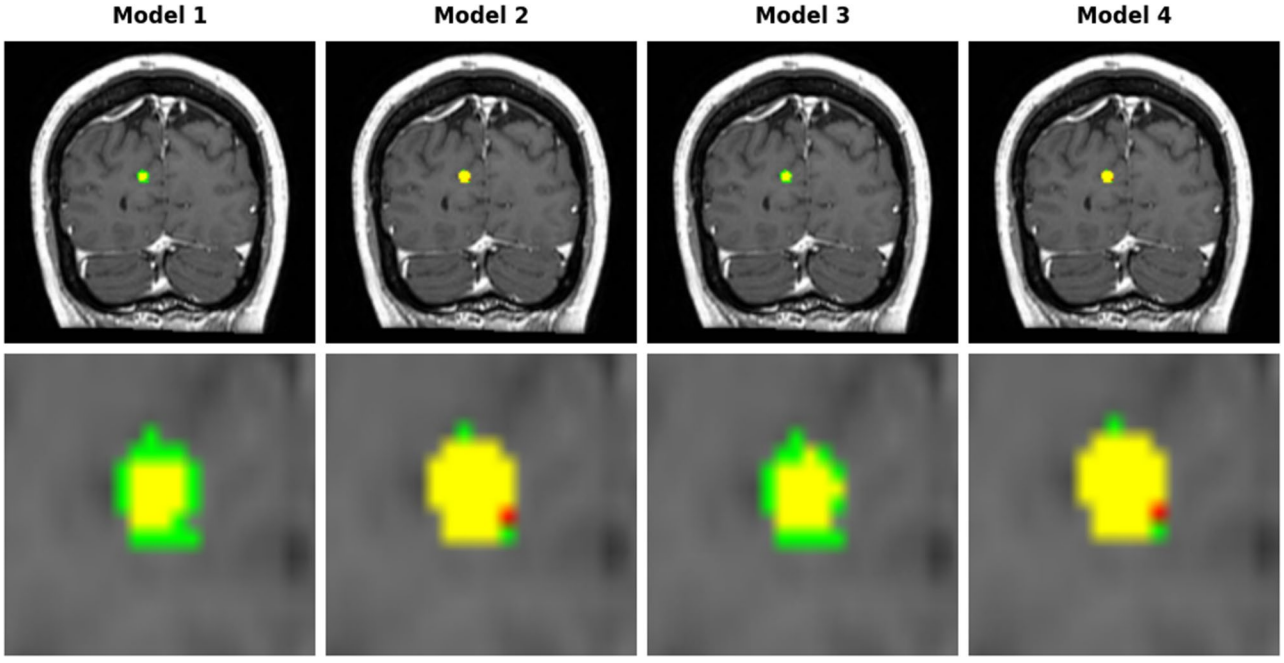

Fig. 4 Results of the 4 models for a sample patient in a follow-up test set (FU5). The second row is a zoomed-in version of the segmentations in the first row. Models 2 and Model 4 segmented larger tumor region when compared to the other models. The reference segmentations are marked in green, the segmentations from the models are marked in red and the overlapping region is marked in yellow

Table 6 LMM – fixed effects

Name	Estimate (Std. error)	t-statistic	*p*
(Intercept)	0.46 (0.06)	7.40	< 0.001
model	0.11 (0.03)	3.56	< 0.001
data	0.14 (0.03)	4.54	< 0.001
model: data	-0.05 (0.02)	-2.68	0.008

Table 7 LMM – fixed effects for ETZ data

Name	Estimate (Std. error)	t-statistic	*p*
(Intercept)	0.59 (0.05)	11.05	< 0.001
model	0.06 (0.01)	4.45	< 0.001

Table 8 LMM – fixed effects for combined data

Name	Estimate (Std. error)	t-statistic	*p*
(Intercept)	0.67 (0.04)	15.79	< 0.001
model	0.03 (0.01)	3.13	0.002

Table 9 LMM – fixed effects for nnU-Net model

Name	Estimate (Std. error)	t-statistic	*p*
(Intercept)	0.57(0.05)	11.45	< 0.001
model	0.09 (0.01)	5.11	< 0.001

Table 10 LMM – fixed effects for MedNeXt model

Name	Estimate (Std. error)	t-statistic	*p*
(Intercept)	0.69(0.04)	16.26	< 0.001
model	0.04 (0.01)	2.74	0.006

Outcome of statistical analysis

We developed an LMM to compare the DSC of different models trained on different datasets over multiple time points. The outcome of the LMM is shown in Table 6.

Both the main effect of model type (B = 0.11, t(960) = 3.56, $p < 0.001$) and data type (B = 0.14, t(960) = 4.54, $p < 0.001$) as well as their interaction (B = -0.05, t(960)=-2.68, $p = 0.008$) reached significance. To understand the interaction effect, we ran additional analyses, separately for each data type and model. The outcomes of these analyses are shown in Tables 7, 8, 9 and 10. The effect of the model reached significance for the ETZ data (B = 0.06, t(960) = 4.45, $p < 0.001$) and also for the combined data (B = 0.03, t(960) = 3.13, $p = 0.002$). Similarly, the effect of the data also reached significance for nnU-Net model (B = 0.09, t(960) = 5.11, $p < 0.001$) and MedNeXt model (B = 0.04, t(960) = 2.74, $p = 0.006$).

Figure 5 shows the interaction plot for the performance of the nnU-Net and MedNeXt algorithms for ETZ data and for the combination of ETZ and public data. MedNeXt provides superior DSC when compared with nnUNet for both data sets although the difference is smaller (but still statistically significant) for the combined dataset.

Discussion

In the present work we assessed the effectiveness of the nnU-Net and MedNeXt algorithms for automated segmentation of both planning and follow-up MRI for BM patients. We conducted experiments by training four distinct models using these algorithms. Specifically, two

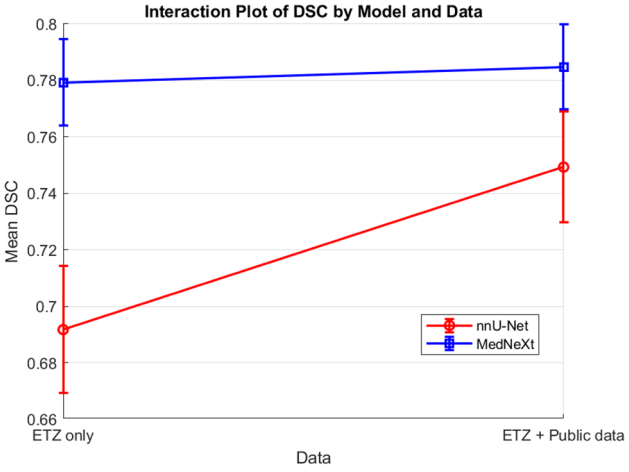

Fig. 5 Effect of model and dataset on DSC

nnU-Net models were trained – one solely with ETZ data and the other with the combination of ETZ and public data. Similarly, two MedNeXt models were trained – one exclusively with ETZ data and the other with the combination of ETZ and public data.

As shown in Table 2, three of the models achieved a DSC of 0.94 on planning MRI while Model 2 (nnU-Net trained with ETZ and public data) achieved a DSC of 0.93. These results indicate a good performance of the models in segmenting BM on the planning MRI, with minimal variation in performance across the different models. Notably, the performance of our models for the planning MRI surpassed that reported in similar studies. For example, Hsu et al. [19] expounded a fully 3D DL approach capable of automatically detecting and segmenting BM on MRI and on CT scans. The DSC of this algorithm was found to be 0.76. Grøvik et al. [21] observed a DSC of 0.79 while evaluating a DL algorithm for detection and segmentation of BM on multisequence MRI. Ziyaee et al. [29] evaluated the effectiveness of the nnU-Net al.gorithm specifically for segmentation of BM by training and testing it with planning MRI only and achieved an overall DSC of 0.82.

However, during the evaluation on the planning MRI, our models exhibited some limitations. During the visual inspection of the segmentation outputs with our radiologists, we observed that the models tended to miss metastases situated near a blood vessel or the tentorium and occassionally produced false positive segmentations by misidentifying blood vessels as BM. These observations highlight the need to further investigate location-specific performance variations in future studies. Location-specific improvements to the models such as location-based post-processing can enhance sensitivity and specificity, leading to more reliable detection of BM in clinically challenging regions. However, the models did detect and segment a single brain metastasis that was accidentally missed in the reference segmentation (see Fig. 6 in the supplementary section). This tumor that was detected by the models was included in the reference segmentation of the subsequent follow-up scans. While this was an isolated case, it underscores the potential utility of these models in assisting clinicians with the early detection and segmentation of the tumors.

In contrast to the robust performance on planning MRI, the models showed lower efficacy for the follow-up images. This decline could be attributed to several factors, including the radiation effect which causes the tumors to shrink over time. Table 1 shows that the median tumor volume of the follow-up images is less than the planning MRI and also the median tumor volume decreased on subsequent follow-ups. The detection and segmentation performances of the DL algorithms tend to decrease for smaller lesions [20]. Hence, the shrinkage of the tumors over time due to the radiation effect could be a reason for the lower performance for successive follow-up scans. The decrease in performance for the follow-ups may also be due to the different dose of contrast administered during follow-up scans compared to planning scans. The planning MRI were contrast-enhanced with triple-dose gadolinium and the follow-up images were contrast-enhanced with single-dose gadolinium. Additionally, changes in tumor texture over successive follow-up scans, after multiple sessions of treatment might contribute to the diminished performance of the models on the follow-up MRI. The experiments of You et al. [45] confirmed that the performance of DL models change due to the contrast and texture modifications employed during training and/or testing time.

In most hospitals, the segmentations are done only for the baseline scans and not for the follow-up scans. This lack of segmented follow-up images creates a limitation for training the DL algorithms with follow-up scans. Since the performance of the models trained with ETZ only data is lower for the follow-up MRI compared to the performance for the planning MRI, we added the public data set to the training data and then evaluated the models trained with both ETZ and public data. The performance of the models on follow-up MRI images improved (when compared with the models trained with ETZ data only) by the addition of the public data to the training set and this performance is comparable with other studies on BM segmentation on pre-treatment images [19, 21], suggesting this to be a viable solution to improve the model efficacy for the follow-ups in scenarios where the reference segmentations are lacking for follow-up MRI. This improvement in performance could be attributed to the increased heterogeneity introduced by the additional data in the training set, as well as the similarity in contrast enhancement (both single dose) between the public data in the training data set and the follow-up test

data. We used the publicly available BM images from the Mathematical oncology laboratory provided by Ocaña-Tienda et al. [43]. To our knowledge, this is the only publicly available dataset that includes follow-up MRI scans and segmentations. Making follow-up scans publicly available can further improve the segmentation performance of the algorithms.

An interesting observation from this evaluation of nnU-Net and MedNeXt algorithms is that there is a statistically significant difference in the DSC of the two algorithms when trained with the combination of ETZ and public data. Also, when the algorithms are trained with ETZ data only, the MedNeXt algorithm provides a statistically significant increase in DSC compared to nnU-Net. This shows that MedNeXt provides better DSC when there is limited training data, and the difference becomes smaller (but is still statistically significant) when there is heterogeneous training data. The performance comparison done by Roy et al. [39] shows that MedNeXt provides higher DSC when compared to other algorithms for organ and tumor segmentations. MedNeXt provides higher DSC for both datasets, but the nnU-Net model trained with the combined data provides the best sensitivity and FNR for most follow-ups. If detecting all tumors and minimizing the risk of missing tumors is the highest priority, then nnU-Net might be preferable due to its higher sensitivity and lower false negative rate. On the other hand, if accurate segmentation of the tumor's shape and size is more important (for precise treatment planning, for example), then MedNeXt might be better due to its higher DSC. Hence, MedNeXt emerges as the favorable option for the segmentation of planning MRI and nnU-Net emerges as the favorable option for the deduction of new or recurrent tumors in follow-up MRI.

A reliable automated detection and segmentation process for the follow-up scans has great clinical value by assisting the radiologists in the detection of new lesions in the follow-up scans that were not part of the initial planning MRI and may not be immediately apparent through the manual comparison of the follow-up MRI with the planning MRI. Additionally, automated segmentation enhances the assessment of subtle tumor changes, improving treatment response evaluation. Follow-up scans can differ from initial diagnostic MRIs in imaging protocol, making direct comparisons complex. The limited availability of labelled follow-up MRI datasets further challenges the training of deep learning models specifically for follow-ups. Leveraging models trained on initial diagnostic MRIs is a necessary and practical approach, ensuring accurate and consistent segmentation across different imaging conditions. Accurate segmentation of BMs in the follow-ups can not only improve the clinical workflow and the reading confidence of the radiologists but also reduce workload, making it a valuable tool for clinicians. This study also showed that nnU-Net performs best for the detection of the tumors while MedNeXt is the best for the segmentation of the tumors.

This study also demostrated a solution for the development for DL models in situations where the training data is sparse or not available. Also when there is large training data set, the addition of public data set could increase the heterogeneity of the training data and hence improve the model performance. Since, both nnU-Net and MedNeXt algorithms achieved state of the art performance when compared to other algorithms for medical image segmentations, the aim of this study was to assess their applicability for automated segmentation of both planning and follow-up BM images without modifying the underlying algorithms. While no technical improvements were made, future research could explore enhancements to the algorithms tailored specifically to automated BM segmentation.

Abbreviations

SRS	Stereotactic radiosurgery
BM	Brain metastases
MRI	Magnetic Resonance Imaging
CAR-Study A	Cognition and radiation study a
FU	Follow-up
DSC	Dice similarity coefficient
FNR	False Negative Rate
WBRT	Whole brain radiotherapy
DL	Deep learning
CNNs	Convolutional neural networks
CT	Computed tomography scans
ETZ	Elisabeth-tweeSteden hospital
GKRS	Gamma knife radiosurgery
IoU	Intersection over union

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-025-01643-y.

Supplementary Material 1

Acknowledgements

We would like to acknowledge the support provided by Eline Verhaak for this research and thank her for helping us with the manually delineated ground truth for follow-up scans from the CAR study.

Author contributions

Conceptualization, H.K., W.d.B., P.H., and M.S.; Methodology, H.K. and W.d.B.; Formal Analysis, H.K.; Writing–Review & Editing, H.K., W.d.B., P.H., and M.S.; Supervision, W.d.B., and M.S.

Funding

This research is supported by KWF Kankerbestrijding and NWO Domain AES, as part of their joint strategic research programme: Technology for Oncology IL. The collaboration project is co-funded by the PPP Allowance made available by Health Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

Data availability

The data used for this study is available at ETZ and is accessible after approval from the ETZ Science office.

Declarations

Ethical approval

This study is part of the AI in Medical Imaging for novel Cancer User Support (AMICUS) project at Tilburg University. This project is approved by the Ethics Review Board at the Tilburg University. This study adhered to the ethical principles outlined in the Declaration of Helsinki.

Consent to participate

The data did not contain any identifiable personal information, therefore the need for consent to participate was waived by the Institutional Review Board Elisabeth-TweeSteden Hospital (ETZ), Tilburg, The Netherlands (Study number: L1267.2021 - AMICUS).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Cognitive Neuropsychology, Tilburg University, Tilburg, The Netherlands
[2]Gamma Knife Center, Elisabeth-TweeSteden Hospital, Tilburg, The Netherlands
[3]Department of Neurosurgery, Elisabeth-TweeSteden Hospital, Tilburg, The Netherlands

Received: 25 October 2024 / Accepted: 17 March 2025
Published online: 26 March 2025

References

1. Donofrio CA, Cavalli A, Gemma M et al. Cumulative intracranial tumour volume prognostic assessment: a new predicting score index for patients with brain metastases treated by stereotactic radiosurgery. Clinical & Experimental Metastasis. 2020;37(4):499–508. https://doi.org/10.1007/s10585-020-10037-z.
2. Barnholtz-Sloan JS, Sloan AE, Davis FG, Vigneau FD, Lai P, Sawaya RE. Incidence proportions of brain metastases in patients diagnosed (1973 to 2001) in the metropolitan detroit cancer surveillance system. J Clin Oncol. 2004;22(14):2865–72. https://doi.org/10.1200/jco.2004.12.149.
3. Gavrilovic IT, Posner JB. Brain metastases: epidemiology and pathophysiology. J Neurooncol. 2005;75(1):5–14. https://doi.org/10.1007/s11060-004-8093-6.
4. Soliman H, Das S, Larson DA, Sahgal A. Stereotactic radiosurgery (SRS) in the modern management of patients with brain metastases. Oncotarget. 2016;7(11):12318–30. https://doi.org/10.18632/oncotarget.7131.
5. Rudie JD, Weiss DJ, Colby JB, et al. Three-dimensional U-Net Convolutional Neural Network for Detection and Segmentation of Intracranial Metastases. Radiology. 2021;3(3):e200204–200204. https://doi.org/10.1148/ryai.2021200204.
6. Hwang EJ, Park CM. clinical implementation of deep learning in thoracic radiology: Potential applications and challenges. Korean J Radiol. 2020;21(5):511. https://doi.org/10.3348/kjr.2019.0821.
7. Roque Rodríguez, Outeiral P, van González, Schaake EE, Simões R. Deep learning for segmentation of the cervical cancer gross tumor volume on magnetic resonance imaging for brachytherapy. Radiat Oncol. 2023;18(1). https://doi.org/10.1186/s13014-023-02283-8.
8. Schouten JPE, Noteboom S, Martens RM, et al. Automatic segmentation of head and neck primary tumors on MRI using a multi-view CNN. Cancer Imaging. 2022;22(1). https://doi.org/10.1186/s40644-022-00445-7.
9. Li W, Jia F, Hu Q. Automatic Segmentation of Liver Tumor in CT Images with Deep Convolutional Neural Networks. J Comput Commun. 2015;03(11):146–51. https://doi.org/10.4236/jcc.2015.311023.
10. Yue W, Zhang H, Zhou J, et al. Deep learning-based automatic segmentation for size and volumetric measurement of breast cancer on magnetic resonance imaging. Front Oncol. 2022;12. https://doi.org/10.3389/fonc.2022.984626.
11. Zhang J, Cui Z, Shi Z, et al. A robust and efficient AI assistant for breast tumor segmentation from DCE-MRI via a spatial-temporal framework. Patterns. 2023;4(9):100826–100826. https://doi.org/10.1016/j.patter.2023.100826.
12. Nayak L, Lee EQ, Wen PY. Epidemiology of Brain Metastases. Curr Oncol Rep. 2011;14(1):48–54. https://doi.org/10.1007/s11912-011-0203-y.
13. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A853460. DIVA. Accessed February 21, 2024. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A853460&dswid=8349
14. Bousabarah K, Ruge M, Brand JS, et al. Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. Radiat Oncol. 2020;15(1). https://doi.org/10.1186/s13014-020-01514-6.
15. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. Comput Biol Med. 2018;95:43–54. https://doi.org/10.1016/j.compbiomed.2018.02.004.
16. Lenhard Pennig, Shahzad R, Luciana Albuquerque, Caldeira, et al. Automated Detection and Segmentation of Brain Metastases in Malignant Melanoma: Evaluation of a Dedicated Deep Learning Model. Am J Neuroradiol. 2021;42(4):655–62. https://doi.org/10.3174/ajnr.a6982.
17. Liang Y, Lee K, Bovi JA, et al. Deep Learning-Based Automatic Detection of Brain Metastases in Heterogenous Multi-Institutional Magnetic Resonance Imaging Sets: An Exploratory Analysis of NRG-CC001. Int J Radiation Oncology*Biology*Physics. 2022;114(3):529–36. https://doi.org/10.1016/j.ijrobp.2022.06.081.
18. Jünger ST, Hoyer UCI, Schaufler D, et al. Fully Automated MR Detection and Segmentation of Brain Metastases in Non-small Cell Lung Cancer Using Deep Learning. J Magn Reson imaging: JMRI. 2021;54(5):1608–22. https://doi.org/10.1002/jmri.27741.
19. Hsu DG, Ballangrud Å, Shamseddine A, et al. Automatic segmentation of brain metastases using T1 magnetic resonance and computed tomography images. Phys Med Biol. 2021;66(17):175014. https://doi.org/10.1088/1361-6560/ac1835.
20. Yoo Y, Ceccaldi P, Liu S, et al. Evaluating deep learning methods in detecting and segmenting different sizes of brain metastases on 3D post-contrast T1-weighted images. J Med Imaging. 2021;8(03). https://doi.org/10.1117/1.jmi.8.3.037001.
21. Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. J Magn Reson Imaging. 2019. https://doi.org/10.1002/jmri.26766
22. Hsu DG, Åse Ballangrud, Prezelski K, Swinburne NC, Young R, Beal K, Deasy JO, Cerviño L, Michalis Aristophanous. Automatically tracking brain metastases after stereotactic radiosurgery. Phys Imaging Radiation Oncol [online]. 2023;27:100452–100452. https://doi.org/10.1016/j.phro.2023.100452.
23. Prezelski K, Hsu DG, Balzo L, del, Heller E, Ma J, Pike LRG. Åse Ballangrud and Michalis Aristophanous (2024). Artificial intelligence-driven measurements of brain metastases' response to SRS compare favorably with current manual standards of assessment. Neuro-Oncology Adv [online] 6(1). https://doi.org/10.1093/noajnl/vdae015
24. Jalalifar A, Soliman H, Sahgal A, Sadeghi-Naini AA. Cascaded Deep-Learning Framework for Segmentation of Metastatic Brain Tumors Before and After Stereotactic Radiation Therapy. Annu Int Conf IEEE Eng Med Biol Soc. 2020. https://doi.org/10.1109/embc44109.2020.9175489
25. Lu SL, Yang WC, Chang YC, Chao CC, Liang CH, Chiang PL, Lin V, LU JT, Hsu FM. Automated Detection, Segmentation, and Tracking of Brain Metastases in Repeated Courses of Stereotactic Radiosurgery Using Integrated Artificial Intelligence. Int J Radiation Oncology*Biology*Physics. 2023;117(2):e476–476. https://doi.org/10.1016/j.ijrobp.2023.06.1690.
26. Ocaña-Tienda B, Julián Pérez-Beteta JoséA, Romero-Rosales, Asenjo B, Ortiz A, Alberto L, David J, Nagib F, Denis MV, Luque B, Arana E, Pérez-García VM. Volumetric Analysis: Rethinking Brain Metastases Response Assessment. Neuro-Oncology Adv. 2023;6(1). https://doi.org/10.1093/noajnl/vdad161.
27. 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. 2020;18(2):203–11. https://doi.org/10.1038/s41592-020-01008-z.
28. Isensee F, Jager PF, Simon, Petersen J, Klaus Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. arXiv:1904.08128 [preprint]: Computer Vision and Pattern Recognition.
29. Hamidreza Ziyaee, Ibbott GS, Debra N, Yeboa, et al. Automated Brain Metastases Segmentation With a Deep Dive Into False-positive Detection. Adv radiation Oncol. 2022;8(1):101085–101085. https://doi.org/10.1016/j.adro.2022.101085.

30. Vaswani A, Shazeer N, Parmar N et al. Attention Is All You Need. arXiv:1706.03762v7 [preprint].2023 [cited 2023 Apr 2]. Available from:https://doi.org/10.48550/arXiv.1706.03762
31. Dai Y, Gao Y, Liu F, TransMed. Transformers Advance Multi-Modal Medical Image Classification. Diagnostics. 2021;11(8):1384. https://doi.org/10.3390/diagnostics11081384.
32. Behnaz Gheflati, Rivaz H, vision transformers for classification of breast ultrasound images. Annu Int Conf IEEE Eng Med Biol Soc. 2020. https://doi.org/10.1109/embc48229.2022.9871809
33. Karimi D, Vasylechko S, Gholipour A. Convolution-free medical image segmentation using transformers. arXiv:2102.13645v2 [preprint]. 2022 [cited 2022 Apr 3]. Available from: https://doi.org/10.48550/arxiv.2102.13645
34. O'Shea K, Nash R. An Introduction to convolutional neural networks. arXiv:1511.08458v2 [preprint]. 2015 [cited 2015 Dec 2]. Available from: https://arxiv.org/abs/1511.08458
35. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with Deep Neural Networks. Med Image Anal. 2017;35:18–31. https://doi.org/10.1016/j.media.2016.05.004.
36. Chen J, Lu Y, Yu Q et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. https://www.cs.jhu.edu/~alanlab/Pubs21/chen2021transunet.pdf
37. Park N, Kim S. How Do Vision Transformers Work? Published online February 14, 2022. https://doi.org/10.48550/arxiv.2202.06709
38. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the IEEE, /CVF Conference on Computer Vision and Pattern Recognition, Roy S, Koehler G, Ulrich C et al. MedNeXt: Transformer-driven Scaling of ConvNets for medical image segmentation. arXiv:2303.09975v5 [preprint]. 2024 [cited 2024Jun 2]. Available from:https://doi.org/10.48550/arXiv.2303.09975.
39. Roy S, Koehler G, Ulrich C, et al. MedNeXt: Transformer-driven Scaling of ConvNets for Medical Image Segmentation. arXiv (Cornell University). Published online March 17, 2023. doi:https://doi.org/10.48550/arxiv.2303.09975
40. Verhaak E, Gehring K, Hanssens PEJ, Sitskoorn MM. Health-related quality of life of patients with brain metastases selected for stereotactic radiosurgery. J Neurooncol. 2019;143(3):537–46. https://doi.org/10.1007/s11060-019-03186-z.
41. Gorgolewski K, Burns CD, Madison C, et al. Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. Front Neuroinformatics. 2011;5. https://doi.org/10.3389/fninf.2011.00013.
42. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL NeuroImage. 2012;62(2):782–90. https://doi.org/10.1016/j.neuroimage.2011.09.015.
43. Ocaña-Tienda B, Pérez-Beteta J, Villanueva-García JD, et al. A comprehensive dataset of annotated brain metastasis MR images with clinical and radiomic data. Sci Data. 2023;10(1):208. https://doi.org/10.1038/s41597-023-02123-0.
44. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res Notes. 2022;15:210. https://doi.org/10.1186/s13104-022-06096-y.
45. You S, Reyes M. Influence of contrast and texture based image modifications on the performance and attention shift of U-Net models for brain tissue segmentation. Front Neuroimaging. 2022;1:1012639. https://doi.org/10.3389/fnimg.2022.1012639.

Publisher's note