Automatic Detection and Multi-Component Segmentation of Brain Metastases in Longitudinal MRI

Vincent Andrearczyk, PhD[1,2], Luis Schiappacasse, MD[3,*], Daniel Abler PhD[1,4], Marek Wodzinski, PhD[1], Andreas Hottinger, MD, PhD[4], Matthieu Raccaud[5], Jean Bourhis, MD, PhD[3], John O. Prior, MD, PhD[2], Vincent Dunet, MD, PhD[5], and Adrien Depeursinge, PhD[1,2]

[1] *Institute of Informatics, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland*

[2] *Department of Nuclear Medicine and Molecular Imaging, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Switzerland*

[3] *Department of Radiation Oncology, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland*

[4] *Department of Oncology, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland*

[5] *Department of Medical Radiology, Service of Diagnostic and Interventional Radiology, Neuroradiology Unit, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland*

[*] *Corresponding author: Luis Schiappacasse, Rue du Bugnon 46, 1011 Lausanne, Switzerland, +41 21 314 12 19, luis.schiappacasse@chuv.ch*

Abstract

Manual segmentation of lesions, required for radiotherapy planning and follow-up, is time-consuming and error-prone. Automatic detection and segmentation can assist radiologists in these tasks. This work explores the automated detection and segmentation of brain metastases (BMs) in longitudinal MRIs. It focuses on several important aspects: identifying and segmenting new lesions for screening and treatment planning, re-segmenting lesions in successive images using prior lesion locations as an additional input channel, and performing multi-component segmentation to distinguish between enhancing tissue, edema, and necrosis. The key component of the proposed approach is to propagate the lesion mask from the previous time point to improve the detection performance, which we refer to as "re-segmentation". The retrospective data includes 518 metastases in 184 contrast-enhanced T1-weighted MRIs originating from 49 patients (63% male, 37% female). 131 time-points (36 patients, 418 BMs) are used for cross-validation, the remaining 53 time-points (13 patients, 100 BMs) are used for testing. The lesions were manually delineated with label 1: enhancing lesion, label 2: edema, and label 3: necrosis. One-tailed t-tests are used to compare model performance including multiple segmentation and detection metrics. Significance is considered as p<0.05. A Dice Similarity Coefficient (DSC) of 0.79 and F_1-score of 0.80 are obtained for the segmentation of new lesions. On follow-up scans, the re-segmentation model significantly outperforms the segmentation model (DSC and F_1 0.78 and 0.88 vs 0.56 and 0.60). The re-segmentation model also significantly outperforms the simple segmentation model on the enhancing lesion (DSC 0.76 vs 0.53) and edema (0.52 vs 0.47) components, while similar scores are obtained on the necrosis component (0.62 vs 0.63). Additionally, we analyze the correlation between lesion size and segmentation performance, as demonstrated in various studies that highlight the challenges in segmenting small lesions. Our findings indicate that this correlation disappears when utilizing the re-segmentation approach and evaluating with the unbiased normalized DSC. In conclusion, the automated segmentation of new lesions and subsequent re-segmentation in follow-up images was achievable, with high level of performance obtained for single- and multiple-component segmentation tasks.

1   Introduction

Brain metastases   Brain metastases (BMs) originate from cancer cells that spread to the brain from other sites, frequently breast, lung, kidney, and melanoma [1]. Despite recent advances in screening and care, BMs remain a major cause of morbidity and mortality. Treatments include one or a combination of medication, surgery, stereotactic radiosurgery (SRS) and whole-brain radiation. Contrast-enhanced T1-weighted magnetic resonance imaging (MRI), recently using 3D magnetisation-prepared rapid gradient echo (MPRAGE), is commonly used for diagnosis, treatment planning and follow-up. Manual detection and segmentation of lesions, required for radiotherapy planning and follow-up, is time-consuming and error-prone. Automatic detection and segmentation can assist radiologists in these tasks.

Automated brain metastases segmentation   Several works have reported promising results for the automatic detection and segmentation of BMs from T1 MRI [2, 3], T1 and CT [4], multiple T1s and FLAIR [5], using standard 2D, 2.5D and 3D Deep Learning (DL) models.

1

Charron et al. [6] were among the first to apply DL to BM segmentation, using adapted 3D and 2D models on T1 and FLAIR images. Huang et al. [7] proposed a DL model utilizing the contrast difference between consecutive images as a temporal prior, with the emergence or growth of high contrast an indicator for BMs. Zhang et al. [2] used a region-based model (Faster R-CNN) for the detection of BMs in T1 images. In a recent survey, Wang et al. [8] underlined the superiority of U-Net and its variants in BMs segmentation accuracy. Across all referenced studies, the lesion-wise Dice Similarity Coefficient (DSC) ranged from 0.55 to 0.915 and the lesion-wise sensitivity from 0.58 to 0.98.

Grøvik et al. [9] proposed a method with input dropout to handle missing MRI sequences in multimodal DL models for BM segmentation. Zhou et al. [10] proposed a 2-stage algorithm for BM segmentation, with a single-shot detector to first detect regions containing metastases, followed by a DL model to segment the metastases from those regions. Three-dimensional U-Net convolutional neural network for detection and segmentation of intracranial metastases [11] 3D UNet, T1w images. Most studies agree that small lesions pose a challenge in detection, leading to a high false negative rate. In particular, [3, 10] observe a DSC of 0.31/0.17 for lesions <3mm and 0.87/0.87 for $\geq$ 6mm. [9, 12, 13] observe a large performance drop for lesions <10mm^3, <15mm and <0.06mL, respectively. Dikici et al. [12] focused on detecting small lesions (<15mm) in T1 images by selecting candidates and using DL classification on cropped regions around them. Bousabarah et al. [13] trained a model exclusively on small lesions to achieve sensitive predictions and ensembled them with predictions from other models trained on all lesions. Finally, BM segmentation on pre-treatment images was the main task of the Brain Tumor Segmentation - Metastases (BraTS-METS 2023) challenge [14]. The top-performing algorithm reached an average lesion-wise DSC of 0.65±0.25 across the three component enhancing tumor, tumor core and whole tumor.

Several gaps remain to address in the existing literature, in particular multi-component segmentation and re-segmentation in follow-up images.

Longitudinal (re-)segmentation Studies on longitudinal data lesion segmentation primarily address the detection of new lesions (e.g. in multiple sclerosis lesions [15]). Examples include incorporating auxiliary tasks like image registration [16], and utilizing multiple time-points as inputs [17, 7]. These works do not address the re-segmentation of brain lesions in follow-up images. This problem is tackled for instance in whole-body CT scans for tracking soft-tissue lesions [18], by inputting a region around the lesion from the previous time-point after registration.

In this work, we explore multiple scenarios of DL for automated detection and segmentation of BMs in pre- and post-treatment T1 MRI images motivated by clinical applications and research. An overview of the study, put in a clinical context, is illustrated in Figure 1. This includes the automatic detection and segmentation of BMs prior to treatment, and the re-segmentation on consecutive post-treatment images. We also evaluate the benefit of adding the T2 sequence for the segmentation of edema. The research questions are (i) Can prior knowledge on previously contoured lesions improve re-segmentation of BMs in follow-up images ? (ii) Can a DL model segment separately the enhancing lesion, edema and necroses ? The novelty is therefore two-fold. We incorporate a BM location prior as input to the DL model to re-segment previously contoured lesions, and we propose a model for 3-label segmentation of relevant BM components including enhancing lesion, edema and necrosis, using a dataset specifically annotated for this purpose.

2 Materials and Methods

2.1 Tasks

We define different tasks for corresponding clinical scenarios. Task 1 is the detection and segmentation of lesions, particularly focused on the initial appearance of lesions, imaged before treatment. Task 2 is the re-segmentation of lesions on consecutive images for patient follow-up. During follow-ups, these two tasks are run in parallel, as shown in Figure 1,
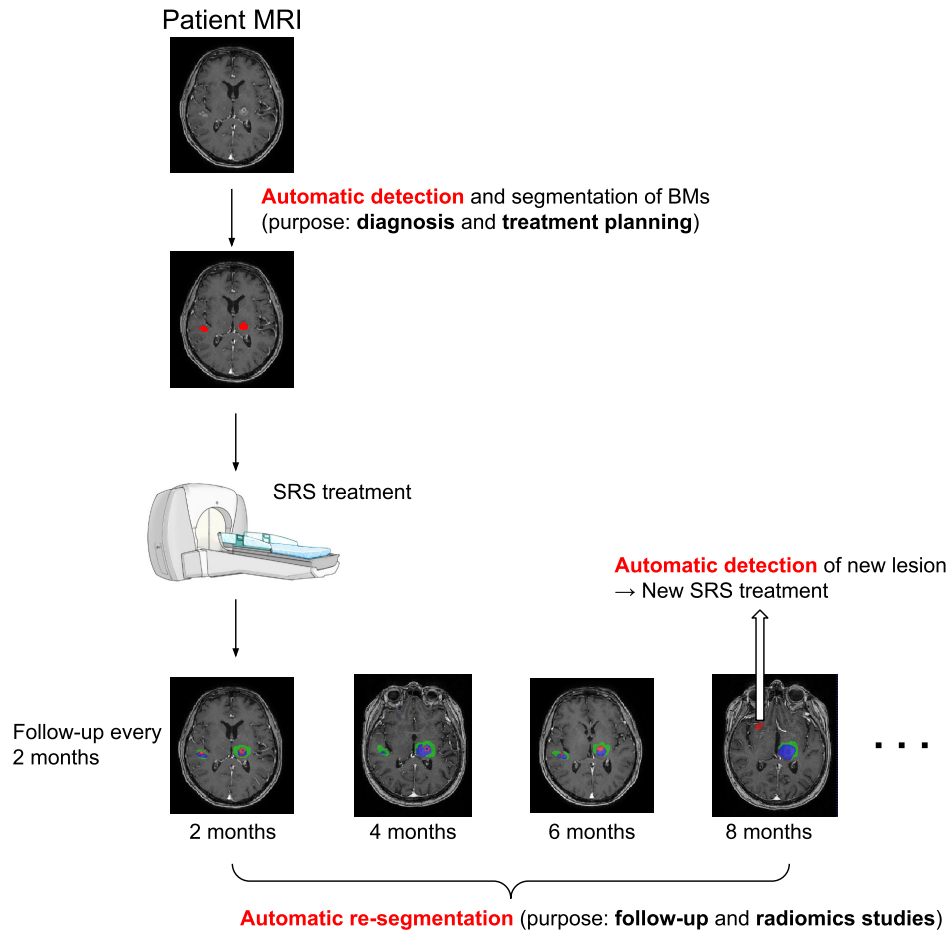
Figure 1: Overview of the scenarios for different clinical and research applications. In this example, new lesions are detected as a single label (Task 1.1) and are then re-segmented in follow-up images with three labels (Task 2.2, enhancing lesion, edema and necrosis).

to re-segment previously existing lesions, and detect the appearance of new ones. Both tasks are divided into sub-tasks for the segmentation of a single label and multiple labels. Three labels are available from the annotations; label 1: enhancing lesion, label 2: edema, and label 3: necrotic part of the lesion (not all lesions present edema or necrosis).

- Task 1.1: detection and segmentation of whole lesions (union of labels 1 and 3).

- Task 1.2: multi-label detection and segmentation of lesions (labels 1, 2 and 3 separately).

- Task 2.1: re-segmentation of whole lesions (union of labels 1 and 3).

- Task 2.2: multi-label re-segmentation of lesions (labels 1, 2 and 3 separately).

The targeted clinical and research goals include detection and segmentation for new SRS treatment, follow-up for automatic RANO-BM [19] or volume assessment, and radiomics studies for prediction of response to treatment.

2.2 Data

The anonymized dataset originates from a retrospective, single-center, longitudinal study at CHUV [20], in accordance with the Declaration of Helsinki, the Swiss legal requirements and the principles of Good Clinical Practice.The protocol, including the requirement for informed consent from all patients whose data was utilized in the study, was approved by the Research Ethics Committee of Vaud Canton, Switzerland (No. 2019-00448).

The dataset comprises 184 time-points from 49 patients and a total of 518 BMs. The inclusion criteria require patients diagnosed with BMs originating from a melanoma primary cancer, treated with SRS, and imaged with a post-contrast MPRAGE T1-weighted MRI. Patients with meningeal metastases were excluded. Patients and treatments characteristics are summarized in Table 1.

After training with a senior neuroradiologist (14 years experience), a master student delineated the lesions on the post-contrast T1 with label 1: enhancing lesion, label 2: edema, and label 3: necrosis, using ITK-SNAP[21]. The enhancing lesion and necrosis were delineated on the T1 sequence, the edema on the T2 sequence, superimposed with the T1. The delineations were verified by the senior neuroradiologist. Examples of manually delineated labels are illustrated in Figure 2, 4th column.

The average number of time-points per patient is 3.7 ± 3.0 (median 2). There is an average of 2.8 ± 2.6 (median 2) lesions across all patients and time-points. The percentage of lesions with edema and necrosis is 57% and 23%, respectively. The average volume of all enhancing lesions is 890 ± 2342 mm^3 (median 171). That of edemas is 7908 ± 17556 mm^3 (median 1367), and necrosis 1089 ± 2546 mm^3 (median 158). Examples of lesions are illustrated in Figure 2, 1st column, showing the heterogeneity of the data in terms of lesion size, location in the brain, structure (enhancing, edema, necrosis), and appearance.

2.3 Pre-processing

For the re-segmentation models, pairs of consecutive images are co-registered using the ANTS toolbox [22], specifically an affine transformation followed by deformable transformation, with cross-correlation optimization metric. Brain masks, obtained with HD-BET [23], restrict the registration within the brain. The labels from the previous time-point are aligned using the resulting transform to use as additional input alongside the MRI.

The registration is also used as a potential re-segmentation method itself, directly using the aligned labels as predictions.

For all tasks, the images are pre-processed following the nnUNet pipeline [24], including z-score normalization, 1mm^3 resampling of the images (3rd order spline) and ground truth labels (nearest-neighbor).

2.4 Models and Training

We use the publicly available 3D nnUNet framework [24], a commonly used semantic segmentation

Demographics	
Gender	
Females	18 (36.7%)
Males	31 (63.3%)
Age (years)	
Average	65.78
Median	66
Standard deviation	11.96
Diagnosis	
Primary site of melanoma	
Trunk	15 (30.6%)
Lower limb	9 (18.4%)
Head & neck	7 (14.3%)
Upper limb	6 (12.2%)
Mucosal	2 (4.1%)
Choroid	1 (2%)
Unknown	9 (18.4%)
Treatments	
Technique of radiosurgery (number of treatments) [a]	
CyberKnife	48
Gamma Knife	26
Systemic treatments - Number (%) of patients receiving	
Checkpoint inhibitors	
Ipilimumab (anti-CTLA-4)	27 (55.1%)
Nivolumab (anti-PD1)	21 (42.8%)
Relatlimab (LAG-3 inhibitor)	4 (8.2%)
Oncolytic viral immunotherapy	
Talimogene laherparepvec (T-VEC)	2 (4.1%)
BRaf- and MEK-selective inhibitors	9 (18.4%)
BRAF inhibitors	
Vemurafenib	14 (28.6%)
Dabrafenib	12 (24.5%)
MEK inhibitors	
Trametinib	15 (30.6%)
Cobimetinib	3 (6.1%)
Tyrosine kinase inhibitors	
Sorafenib	3 (6.1%)
Lapatinib	1 (2%)
Pazopanib	1 (2%)
Chemotherapies	
Temozolomide	11 (22.4%)
Dacarbazine	9 (18.4%)
Carboplatin-Taxol	5 (10.2%)
Nab-Paclitaxel	3 (6.1%)
Fotemustine	2 (4.1%)

Table 1: Characteristics of patients and treatments. [a] 19 (38.8%) patients received more than one radiosurgery treatment.

model developed to adapt to a given dataset. The model contains an encoder path, consisting of a standard convolutional network (including convolutions, activations, max pooling etc.), and a decoder path.

Various methods are evaluated and compared including (i) nnUNet given the image at a single time-point as input (single input channel); (ii) labels propagated from a previous time-point via registration (as described in Section 2.3, only for Task 2); (iii) re-segmentation using nnUNet with labels propagated from the previous time-point as an additional input channel (only for Task 2). Besides, we are interested in the segmentation of each lesion as a single structure and three separate components. Accordingly, we use models with a single output label or three output labels. Methods (ii) and (iii) constitute important novelty when compared to previously proposed approaches, which did not considered the propagation of the lesion masks from the previous time points.

We follow the standard nnUNet training, with random initial weights, involving a 5-fold cross-validation trained with Dice and cross-entropy loss for a total of 1000 epochs, a stochastic gradient descent optimizer with Nesterov momentum. Sliding windows (160x128x112) are used with patch overlap and standard data augmentation (e.g. rotation, scaling, noise). As implemented in the nnUNet, ensembling is performed by choosing the best combination of models on the cross-validation for final prediction on the test set.

Images from 131 time-points (36 patients, 418 BMs) are used for the cross-validation, the remaining 53 time-points (13 patients, 100 BMs) are used for testing. All time-points of a patient are in the same split. The splits are designed to maintain a relatively constant average number of time-points.

For Task 1, we train and evaluate using the entire dataset. Since this model is specifically developed for the segmentation of new lesions, we also evaluate it on a subset containing pre-treatment lesions only which are larger and easier to detect than treated lesions. For Task 2, the goal is to re-segment BMs already contoured at a previous time-point. Thus, we consider for this task only pairs of consecutive images where there is no appearance of new lesion

(36 test pairs). To use all possible data for training (131 time-points), we "augment" the remaining images by artificially creating fictitious "previous" labels for new BMs using random dilation, erosion and translation in all three axes (ranges [0,6] voxels, [0,3] voxels and [-6,+6] voxels, respectively). We do not augment the test set to evaluate only on real consecutive follow-up images. While future work could explore learning the distribution of transformations reflecting lesion evolution, using a simple heuristic method yielded satisfactory results.

2.5   Evaluation

We employ multiple detection and segmentation metrics for the evaluation of the predictions.

Overall segmentation metrics   The DSC is computed as follows.

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{1}$$

where TP, FP and FN are true positive, false positive and false negative voxels, respectively.

DSC is biased to yield greater values for larger lesions [25]. The normalized DSC (nDSC) is introduced in [26] to de-correlate the DSC with the lesion load and obtain an unbiased metric of performance. The nDSC is particularly relevant for comparing the segmentation of lesions of different sizes, and the correlation of lesion size with segmentation performance without bias. Let $p = \frac{\text{FP}}{\text{TP}}$ and $n = \frac{\text{FN}}{\text{TP}}$.

$$\text{nDSC} = 2(2 + \kappa p + n)^{-1}, \kappa = h(r^{-1} - 1), \tag{2}$$

where h represents the ratio between the positive and negative predicted classes, and $0 < r < 1$ is the occurrence rate of the positive class averaged across all samples (i.e. total number of lesion voxels across divided by total number of voxels).

The DSC and nDSC are computed for all images (possibly containing multiple lesions) and averaged. The two metrics are also computed per ground-truth lesion (only in Section 3.3) to evaluate the correlation with lesion size.

Detection metrics The F_1-score evaluates the detection performance, i.e. at the lesion level.

$$F_1 = \frac{\text{TP}_l}{\text{TP}_l + 0.5(\text{FP}_l + \text{FN}_l)}. \quad (3)$$

A predicted lesion that overlaps (intersection over union) for more than 25% with a ground-truth lesion is considered a match to calculate the true positive, false positive and false negative lesions (TP_l, FP_l and FN_l).

True positive rate (TPR) and false discovery rate (FDR) are defined as follows.

$$\text{TPR} = \frac{\text{TP}_l}{\text{TP}_l + \text{FN}_l}, \quad (4)$$

$$\text{FDR} = \frac{\text{FP}_l}{\text{TP}_l + \text{FP}_l}. \quad (5)$$

2.6 Statistical tests

Metrics means are reported with standard deviation and statistical tests are performed with a one-tailed t-test to compare models performance. Correlation between lesion volume and model performance is performed with a Spearman correlation. Significance is considered as p<0.05.

The SciPy library is employed to conduct the statistical tests and correlation analyses.

3 Results

3.1 Detection and Segmentation

The performance of the models evaluated on Task 1, detection and segmentation on a single image, are reported in Table 2 for single-label and multi-label. As illustrated in Figure 2, some very small lesions are missed by the model. Small lesions are often the result of effective treatment. Since the analysis performed here is dedicated to newly appeared lesions, we also evaluate on the more clinically relevant first appearance of each lesion, resulting in a higher DSC of 0.79 and F_1-score of 0.80 (vs. 0.64 and 0.68, respectively on the full test set). These results are reported as "full" and "pre-treatment" test sets in Table 2. Other lesions that can be tracked in follow-up images, often shrunk after treatment, are more accurately segmented using the re-segmentation algorithm in Section 3.2). These two models can be run in parallel on follow-up images to optimize both detection of new lesions and re-segmentation.

To provide a reference, we present the mean Dice Similarity Coefficient DSC and F1-score achieved on the single-label training dataset, yielding values of 0.85±0.13 and 0.92±0.15, respectively.

Qualitative results with single label (Task 1.1) are reported in Figure 2, columns 2 and 3.

3.2 Re-segmentation in Follow-ups

The results of the re-segmentation (Task 2) with location prior are reported in Table 3 and compared with the segmentation results without prior. Note that the results of the latter are different from those in Table 2 because the test set does not contain exactly the same cases, see Section 2.2. Additionally, we report results obtained solely via registration (see Section 2.3), without re-segmentation. These results are only reported for single-label Task 2.1 because registration cannot predict the appearance of necrosis or edema.

For comparison, the DSCs and F1-scores on the entire 3-labels training set are label 1: 0.92± 0.06, 0.98 ± 0.10; label 2: 0.65± 0.38, 0.51± 0.37; and label 3: 0.96± 0.10, 0.97± 0.11.

The lowest results are obtained for the edema, difficult to delineate using only T1 sequences. Training another re-segmentation model with T1 and T2 inputs, despite the missing T2 data (10% of the entire data, all comprised in the training set, T2 is available for all test cases) replaced by zero values, leads to a significant increase in performance for edema segmentation (DSC = 0.62* (p<0.05) and F_1-score = 0.42, vs DSC = 0.52 and F_1-score = 0.38 with the single T1 sequence).

Qualitative results for Task 2.2 are illustrated in Figure 2. The first row depicts a small lesion missed by the segmentation model and correctly segmented by the re-segmentation model. Other examples of le-

labels	test set	mean DSC ↑	mean nDSC ↑	F_1-score ↑	TPR ↑	FDR ↓
1+3	full	0.64 ± 0.41	0.64 ± 0.41	0.68 ± 0.44	0.66 ± 0.44	0.01 ± 0.05
	pre-treat.	0.79 ± 0.31	0.78 ± 0.31	0.80 ± 0.33	0.78 ± 0.34	0.02 ± 0.05
1	full	0.62 ± 0.38	0.60 ± 0.39	0.66 ± 0.43	0.66 ± 0.44	0.05 ± 0.16
	pre-treat.	0.81 ± 0.22	0.79 ± 0.23	0.86 ± 0.26	0.83 ± 0.27	0.03 ± 0.10
2	full	0.52 ± 0.29	0.51 ± 0.26	0.28 ± 0.32	0.23 ± 0.30	0.42 ± 0.44
	pre-treat.	0.55 ± 0.29	0.52± 0.26	0.39 ± 0.40	0.38 ± 0.40	0.57 ± 0.38
3	full	0.67 ± 0.25	0.66 ± 0.24	0.78 ± 0.34	0.85 ± 0.36	0.20 ± 0.35
	pre-treat.	0.80 ± 0.12	0.77 ± 0.12	0.93 ± 0.12	1.00 ± 0.00	0.11 ± 0.18

Table 2: Results on Task 1, detection and segmentation. Results of two sub-tasks are reported, separated by a double horizontal line, for the segmentation of a single label (whole lesion, Task 1.1) and three separate labels (Task 1.2). Results are reported separately for all test cases (full, 53 images), and for pre-treatment images only (15 images).

labels	model	mean DSC ↑	mean nDSC ↑	F_1-score ↑	TPR ↑	FDR ↓
1+3	seg.	0.56 ± 0.43	0.57 ± 0.44	0.60 ± 0.47	0.60 ± 0.47	0.01 ± 0.06
	regis.	0.74^* ± 0.18	0.82^* ± 0.14	0.93^* ± 0.23	0.94^* ± 0.23	0.07 ± 0.23
	re-seg.	0.78^* ± 0.26	0.81^* ± 0.27	0.88^* ± 0.32	0.88^* ± 0.32	0.07 ± 0.23
1	seg.	0.53 ± 0.41	0.52 ± 0.42	0.56 ± 0.47	0.56 ± 0.48	0.05 ± 0.19
	re-seg.	0.76^* ± 0.25	0.77^* ± 0.27	0.86^* ± 0.32	0.87^* ± 0.32	0.11 ± 0.29
2	seg.	0.47 ± 0.28	0.47 ± 0.25	0.22 ± 0.25	0.15 ± 0.18	0.35 ± 0.45
	re-seg.	0.52^* ± 0.26	0.51 ± 0.23	0.38^* ± 0.32	0.38^* ± 0.37	0.45 ± 0.37
3	seg.	0.63 ± 0.25	0.62 ± 0.24	0.73 ± 0.40	0.82 ± 0.39	0.21 ± 0.36
	re-seg.	0.62 ± 0.25	0.60 ± 0.24	0.79 ± 0.38	0.82 ± 0.39	0.14 ± 0.31

Table 3: Results on Task 2, the re-segmentation of previously existing lesions. Results of two sub-tasks are reported, separated by a double horizontal line, for the segmentation of a single label (whole lesion,Task 2.1) and three labels (Task 2.2). The re-segmentation model with prior (re-seg.) is compared to the segmentation model without prior developed for Task 1 (seg.). For the single label segmentation, the results obtained with a simple non-rigid registration are also reported (regis.). For the multi-label segmentation, the average results on labels 2 and 3 are computed for images containing a ground truth volume of the corresponding label (36 and 20 images, respectively) to avoid boosting the results with perfect predictions resulting from true negatives. The asterisk (*) denotes statistical significance, investigating the superiority of performance of the re-segmentation (and registration-only approach) over segmentation without prior.
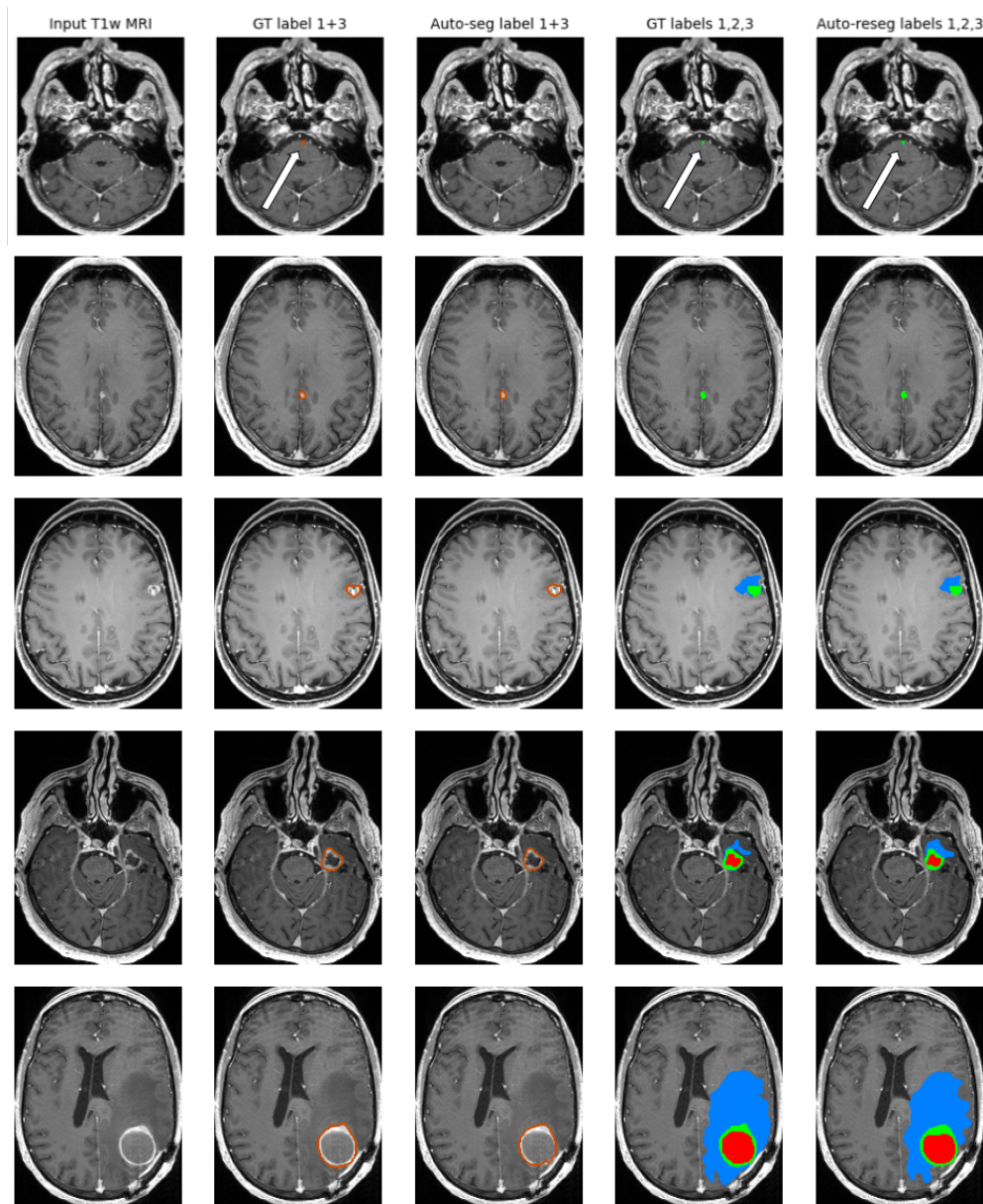
Figure 2: Qualitative results (2D axial views) of the automatic BM segmentation in T1 images. The second column shows the ground truth for a single whole lesion label (labels 1+3). The third column is the contoured prediction of the detection and segmentation model (Task 1.1). The fourth column shows the ground truth with the three individual labels, namely 1: enhancing lesion (green), 2: edema (blue), 3: necrosis (red). The last column illustrates the prediction of the re-segmentation model with these three labels (Task 2.2). The first row illustrate a very small lesion missed by the segmentation model without prior information on previous time-point.

9

sions of different sizes and with different components are also illustrated.

3.3 Correlation Lesion Volume and Performance

For single-label segmentation (Task 1.1), the Spearman correlation between lesion volume and DSC is 0.73 (significant) when considering all lesions, new and previously treated. Note that Pearson correlation is lower than Spearman on these evaluations. The correlation decreases to 0.31 with the nDSC (unbiased towards lesion size), yet remains statistically significant due to false negative small lesions. Notably, when focusing on the initial appearance of lesions in baseline images, which is more relevant to Task 1 and encompasses larger lesions than the treated ones, no correlation is found using the nDSC ($p=0.19$).

With the resegmentation model, no significant correlation (Pearson or Spearman) is observed between the nDSC and the lesion volume (coefficients -0.17 and -0.07, respectively).

4 Discussion

4.1 Detection and Re-Segmentation

Previous studies [6, 12, 5, 3] reported the challenge in detecting small lesions (e.g., <15mm diameter) with standard models. Our experimental results confirm this observation through the strong correlation between lesion volume and segmentation performance. This correlation, however, vanished when using the prior for re-segmentation and evaluating with the nDSC. The proposed re-segmentation model with location prior significantly improves the segmentation of small treated lesions, as depicted in Figure 2 (first row) and reported in the overall segmentation results (Task 2.1, Table 3, with DSC and F_1-score of 0.78 and 0.88, respectively, vs 0.56 and 0.60 with the simple segmentation model). These results show that an excellent detection of new lesions and re-segmentation in consecutive follow-up images can be obtained by coupling two models specifically trained on Task 1 and Task 2 respectively.

4.2 Comparison with Registration

Employing registration alone for re-segmenting lesions in consecutive time-points leads to seemingly good results (Table 3 3rd row). The F_1-score is high (0.93) because all lesions are present in the previous time-point and matched after registration (overlap>25%). However, the boundary often largely deviates from the actual lesion border due to the variation in size and shape across post-treatment images. Moreover, this method cannot handle multiple labels and fails to address lesions that have completely disappeared, resulting in false positives with important consequences in patient follow-up. These drawbacks limit the utility of registration-based re-segmentation. However, local registration and rules for detecting complete responses could be used in the future to improve results using registration only.

4.3 Multi-Label Segmentation

Depending on the clinical application (lesion detection, treatment planning, patient follow-up, radiomics studies), different sublesional components may be needed. The proposed models can be used with good performance to segment the whole lesions as a single label, as performed in other studies, or three separate labels (enhancing lesion, edema and necrosis) which is particularly relevant for the follow-up of patients via a quantitative treatment response assessment. The boundary of the edema is difficult to locate on T1 images. Including the T2 sequence as additional input, despite missing data, significantly improves its segmentation. However, a drop of performance is observed for the other labels. As the lesion is more important clinically than the edema, we primarily reported the results without T2. More sophisticated handling of the missing data, or an ensembling of predictions may be used in the future for an optimal compromise.

4.4 Limitations

Interestingly, carefully inspecting the predictions in the cross-validation allowed us to detect some BMs that were not manually contoured (similarly to [3]) and to correctly re-delineate them. Limitations of segmentation metrics [27] also emerged from the evaluation and visualization of predictions and manual delineations. These observations motivate a comprehensive user study of segmentation quality for better evaluation of clinical readiness.

The study's limitations include potential dataset biases, which may affect the model's generalization to other populations or lesion types (e.g. generalization to other centers and non-melanoma primary cancers). Finally, the segmentation of edema was the most challenging. While this performance can be improved with additional sequences (T2, as reported in Section 3.2), it is also less clinically relevant than the enhancing lesion and necrosis.

4.5 Clinical Significance

The improved segmentation accuracy, particularly for small lesions, has a significant impact on clinical decision-making. Precise lesion delineation enhances treatment planning by ensuring accurate radiation targeting, which may reduce side effects and improve patient outcomes. Additionally, it enables reliable tracking of lesion progression, facilitating timely treatment adjustments. Automatic segmentation, in particular, supports the use of volumetric measurements, which may better represent lesion size and progression than the standard 2D axial measurements. Additionally, the reduced need for manual corrections by radiologists improves workflow efficiency and minimizes human error, leading to a more streamlined and effective clinical workflow.

4.6 Integration with Radiomics and Predictive Analytics

Automatic segmentation enables the extraction of radiomic features essential for training predictive models, such as 12-month response, radionecrosis, or brain disease-free survival (DFS). This also supports the modeling of volume trajectories to identify response populations, allowing for personalized patient management by adjusting follow-up frequency and SRS treatment plans to optimize the balance between tumor response and radionecrosis risk. For instance, Peng et al. [28] demonstrated that radiomic features from segmented lesions improve the distinction between progression and radionecrosis after SRS. Accurate delineation by our model could further enhance such predictions and support precision medicine. In future work, we plan to use these models to automatically segment the BMs in a larger cohort to conduct a large-scale radiomics study for automatic response assessment and outcome prediction, without the need for manual annotation.

5 Conclusion

In this study, we presented a deep learning model for the detection and segmentation of brain metastases in longitudinal MRI, demonstrating significant improvements in accuracy, particularly in handling small lesions and complex multi-component segmentation tasks. The main novelty of our approach is to propagate the segmentation from the previous timepoint, referred to as "re-segmentation", allowing to significantly improve the detection performance. Our model's precise delineation of lesions enhances clinical decision-making by enabling more accurate treatment planning, reliable tracking of lesion progression, and efficient workflow integration. The reduced need for manual corrections not only minimizes human error but also results in substantial time savings, which is critical in high-volume clinical environments.

Despite the promising results, we acknowledge certain limitations, such as dataset biases, which may affect the model's generalizability. Future research should focus on addressing these limitations by expanding the dataset. Additionally, the potential integration of our segmentation model with radiomics and predictive analytics tools holds promise for advancing personalized medicine, offering automated and more accurate assessments of treatment response

and patient outcomes.

Our findings underscore the potential of advanced segmentation models to not only improve clinical workflows but also to contribute to more effective and personalized patient care.

6 Data availability

The datasets generated and/or analyzed during the current study are not publicly available as not permitted by the ethics agreement. However, the corresponding author can be contacted for any inquiries or requests related to the data. Interested researchers are encouraged to reach out to the corresponding author for further information.

References

[1] L. Nayak, E. Q. Lee, and P. Y. Wen, "Epidemiology of brain metastases," *Current oncology reports*, vol. 14, pp. 48–54, 2012.

[2] M. Zhang, G. S. Young, H. Chen, J. Li, L. Qin, J. R. McFaline-Figueroa, D. A. Reardon, X. Cao, X. Wu, and X. Xu, "Deep-learning detection of cancer metastases to the brain on MRI," *Journal of Magnetic Resonance Imaging*, vol. 52, no. 4, pp. 1227–1236, 2020.

[3] H. Ziyaee, C. E. Cardenas, D. N. Yeboa, J. Li, S. D. Ferguson, J. Johnson, Z. Zhou, J. Sanders, R. Mumme, L. Court, *et al.*, "Automated brain metastases segmentation with a deep dive into false-positive detection," *Advances in Radiation Oncology*, vol. 8, no. 1, p. 101085, 2023.

[4] D. G. Hsu, Å. Ballangrud, A. Shamseddine, J. O. Deasy, H. Veeraraghavan, L. Cervino, K. Beal, and M. Aristophanous, "Automatic segmentation of brain metastases using t1 magnetic resonance and computed tomography images," *Physics in Medicine & Biology*, vol. 66, no. 17, p. 175014, 2021.

[5] J. A. Ottesen, D. Yi, E. Tong, M. Iv, A. Latysheva, C. Saxhaug, K. D. Jacobsen, Å. Helland, K. E. Emblem, D. L. Rubin, *et al.*, "2.5 D and 3D segmentation of brain metastases with deep learning on multinational MRI data," *Frontiers in Neuroinformatics*, 2023.

[6] O. Charron, A. Lallement, D. Jarnet, V. Noblet, J.-B. Clavier, and P. Meyer, "Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network," *Computers in biology and medicine*, vol. 95, pp. 43–54, 2018.

[7] Y. Huang, C. Bert, P. Sommer, B. Frey, U. Gaipl, L. V. Distel, T. Weissmann, M. Uder, M. A. Schmidt, A. Dörfler, *et al.*, "Deep learning for brain metastasis detection and segmentation in longitudinal MRI data," *Medical Physics*, vol. 49, no. 9, pp. 5773–5786, 2022.

[8] T.-W. Wang, M.-S. Hsu, W.-K. Lee, H.-C. Pan, H.-C. Yang, C.-C. Lee, and Y.-T. Wu, "Brain metastasis tumor segmentation and detection using deep learning algorithms: a systematic review and meta-analysis," *Radiotherapy and Oncology*, p. 110007, 2023.

[9] E. Grøvik, D. Yi, M. Iv, E. Tong, L. B. Nilsen, A. Latysheva, C. Saxhaug, K. D. Jacobsen, Å. Helland, K. E. Emblem, *et al.*, "Handling missing MRI sequences in deep learning segmentation of brain metastases: a multicenter study," *NPJ digital medicine*, vol. 4, no. 1, p. 33, 2021.

[10] Z. Zhou, J. W. Sanders, J. M. Johnson, M. Gule-Monroe, M. Chen, T. M. Briere, Y. Wang, J. B. Son, M. D. Pagel, J. Ma, *et al.*, "MetNet: Computer-aided segmentation of brain metastases in post-contrast T1-weighted magnetic resonance imaging," *Radiotherapy and Oncology*, vol. 153, pp. 189–196, 2020.

[11] J. D. Rudie, D. A. Weiss, J. B. Colby, A. M. Rauschecker, B. Laguna, S. Braunstein, L. P. Sugrue, C. P. Hess, and J. E. Villanueva-Meyer, "Three-dimensional U-Net convolutional neural network for detection and segmentation of intracranial metastases," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e200204, 2021.

[12] E. Dikici, J. L. Ryu, M. Demirer, M. Bigelow, R. D. White, W. Slone, B. S. Erdal, and L. M. Prevedello, "Automated brain metastases detection framework for T1-weighted contrast-enhanced 3D MRI," *IEEE journal of biomedical and health informatics*, vol. 24, no. 10, pp. 2883–2893, 2020.

[13] K. Bousabarah, M. Ruge, J.-S. Brand, M. Hoevels, D. Rueß, J. Borggrefe, N. Große Hokamp, V. Visser-Vandewalle, D. Maintz, H. Treuer, *et al.*, "Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data," *Radiation Oncology*, vol. 15, no. 1, pp. 1–9, 2020.

[14] A. W. Moawad, A. Janas, U. Baid, D. Ramakrishnan, L. Jekel, K. Krantchev, H. Moy, R. Saluja, K. Osenberg, K. Wilms, *et al.*, "The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri," *arXiv preprint arXiv:2306.00838*, 2023.

[15] M. Diaz-Hurtado, E. Martínez-Heras, E. Solana, J. Casas-Roma, S. Llufriu, B. Kanber, and F. Prados, "Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review," *Neuroradiology*, vol. 64, no. 11, pp. 2103–2117, 2022.

[16] S. Denner, A. Khakzar, M. Sajid, M. Saleh, Z. Spiclin, S. T. Kim, and N. Navab, "Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation," *arXiv preprint arXiv:2004.03675*, 2020.

[17] J. Krüger, R. Opfer, N. Gessert, A. Ostwaldt, C. Walker-Egger, P. Manogaran, A. Schlaefer, and S. Schippling, "Fully automated longitudinal segmentation of new or enlarging Multiple Scleroses (MS) lesions using 3D convolution neural networks," in *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 192, © Georg Thieme Verlag KG, 2020.

[18] A. Hering, F. Peisen, T. Amaral, S. Gatidis, T. Eigentler, A. Othman, and J. H. Moltz, "Whole-body soft-tissue lesion tracking and segmentation in longitudinal ct imaging studies," in *Medical Imaging with Deep Learning*, pp. 312–326, PMLR, 2021.

[19] N. U. Lin, E. Q. Lee, H. Aoyama, I. J. Barani, D. P. Barboriak, B. G. Baumert, M. Bendszus, P. D. Brown, D. R. Camidge, S. M. Chang, *et al.*, "Response assessment criteria for brain metastases: proposal from the rano group," *The lancet oncology*, vol. 16, no. 6, pp. e270–e278, 2015.

[20] F. Martins, L. Schiappacasse, M. Levivier, C. Tuleasca, M. A. Cuendet, V. Aedo-Lopez, B. Gautron Moura, K. Homicsko, A. Bettini, G. Berthod, *et al.*, "The combination of stereotactic radiosurgery with immune checkpoint in-

hibition or targeted therapy in melanoma patients with brain metastases: a retrospective study," *Journal of neuro-oncology*, vol. 146, pp. 181–193, 2020.

[21] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[22] B. B. Avants, N. Tustison, G. Song, *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[23] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, *et al.*, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human brain mapping*, vol. 40, no. 17, pp. 4952–4964, 2019.

[24] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[25] L. Maier-Hein, B. Menze, *et al.*, "Metrics reloaded: Pitfalls and recommendations for image analysis validation," *arXiv. org*, no. 2206.01653, 2022.

[26] V. Raina, N. Molchanova, M. Graziani, A. Malinin, H. Muller, M. B. Cuadra, and M. Gales, "Tackling bias in the dice similarity coefficient: Introducing nDSC for white matter lesion segmentation," *arXiv preprint arXiv:2302.05432*, 2023.

[27] F. Kofler, I. Ezhov, F. Isensee, F. Balsiger, C. Berger, M. Koerner, J. Paetzold, H. Li, S. Shit, R. McKinley, *et al.*, "Are we using appropriate segmentation metrics," *Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. arXiv*, vol. 2103, 2021.

[28] L. Peng, V. Parekh, P. Huang, D. D. Lin, K. Sheikh, B. Baker, T. Kirschbaum, F. Silvestri, J. Son, A. Robinson, *et al.*, "Distinguishing true progression from radionecrosis after stereotactic radiation therapy for brain metastases with machine learning and radiomics," *International Journal of Radiation Oncology* Biology* Physics*, vol. 102, no. 4, pp. 1236–1243, 2018.

Acknowledgments

Author Contributions

VA, LS, DA, MW, VD, AD designed this work and analyzed the data. LS, AH, MR, JB, JOP, VD collected, curated and interpreted the data. All authors have read and approved the final manuscript.