Automatic Assessment of Stereotactic Radiation Therapy Outcome in Brain Metastasis Using Longitudinal Segmentation on Serial MRI

Seyed Ali Jalalifar, *Member, IEEE*, Hany Soliman, Arjun Sahgal, and Ali Sadeghi-Naini, *Senior Member, IEEE*

***Abstract*—The standard clinical approach to assess the radiotherapy outcome in brain metastasis is through monitoring the changes in tumour size on longitudinal MRI. This assessment requires contouring the tumour on many volumetric images acquired before and at several follow-up scans after the treatment that is routinely done manually by oncologists with a substantial burden on the clinical workflow. In this work, we introduce a novel system for automatic assessment of stereotactic radiation therapy (SRT) outcome in brain metastasis using standard serial MRI. At the heart of the proposed system is a deep learning-based segmentation framework to delineate tumours longitudinally on serial MRI with high precision. Longitudinal changes in tumour size are then analyzed automatically to assess the local response and detect possible adverse radiation effects (ARE) after SRT. The system was trained and optimized using the data acquired from 96 patients (130 tumours) and evaluated on an independent test set of 20 patients (22 tumours; 95 MRI scans). The comparison between automatic therapy outcome evaluation and manual assessments by expert oncologists demonstrates a good agreement with an accuracy, sensitivity, and specificity of 91%, 89%, and 92%, respectively, in detecting local control/failure and 91%, 100%, and 89% in detecting ARE on the independent test set. This study is a step forward towards automatic monitoring and evaluation of radiotherapy outcome in brain tumours that can streamline the radio-oncology workflow substantially.**

***Index Terms*—Adverse radiation effect, automatic tumour segmentation, brain metastasis, deep learning, stereotactic radiotherapy, therapy outcome assessment.**

Manuscript received 15 January 2022; revised 20 September 2022; accepted 29 December 2022. Date of publication 9 January 2023; date of current version 6 June 2023. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Grants CRDPJ507521-16 and RGPIN-2016-06472, and in part by the Lotte and John Hecht Memorial Foundation, and the Terry Fox Foundation under Grant 1083. A.S.N. holds the York Research Chair in Quantitative Imaging and Smart Biomarkers. *(Corresponding author: Ali Sadeghi-Naini.)*

Seyed Ali Jalalifar is with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, ON M3J 1P3, Canada (e-mail: alijfar@yorku.ca).

Hany Soliman and Arjun Sahgal are with the Department of Radiation Oncology, Odette Cancer Centre, Sunnybrook Health Sciences Centre, Toronto, ON M4N 3M5, Canada, also with the Physical Sciences Platform, Sunnybrook Research Institute Toronto, Toronto, ON M4N 3M5, Canada, and also with the Department of Radiation Oncology, University of Toronto, Toronto, ON M5S, Canada (e-mail: hany.soliman@sunnybrook.ca; arjun.sahgal@sunnybrook.ca).

Ali Sadeghi-Naini is with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, ON M3J 1P3, Canada, also with the Department of Radiation Oncology, Odette Cancer Centre, Toronto, ON M4N 3M5, Canada, and also with the Physical Sciences Platform, Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, Toronto, ON M4N 3M5, Canada (e-mail: asn@yorku.ca).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2023.3235304, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2023.3235304

I. Introduction

ABOUT 10% to 30% of all cancer patients develop brain metastasis [1], with a higher risk for melanoma, lung, and breast cancer patients. According to population studies, the annual incidence of brain metastases in the United States is estimated to exceed 14 persons per 100000 [2]. Metastatic brain tumours represent an important cause of morbidity and mortality in cancer patients. Whereas a significant proportion of cancer patients survive for many years if the cancer is identified at an early stage while it is still localized [3], when the tumour is metastasized to the brain, the median survival ranges from as short as 5 months to up to 4 years, based on the subgroup and origin of the cancer [4], [5], [6], [7]. Early diagnosis and precise treatment of brain metastasis may lead to the reduction of brain symptoms and may enhance the quality of life and survival of the patients [8], [9], [10].

Brain metastasis may occur as a single tumour (approximately 29% of cases), two-three tumours (35% of cases), and more than three tumours (36% of cases) [11]. Treatment planning for patients diagnosed with metastatic brain tumours depends on many factors including the origin of cancer, symptoms, number of metastases, and location of the tumour. Two main treatment modalities available for the management of metastatic brain tumours include surgery and radiation therapy. Surgery involves resection of the tumour and is often administered when the tumour is large and accessible. Other contributing factors are patient's age, presence of other extracranial diseases, and relative proximity to eloquent brain areas [12]. In whole brain radiation therapy (WBRT) the prescribed radiation dose is delivered to the whole brain in many low-dose fractions over several weeks [13]. In stereotactic radiosurgery (SRS) and hypo-fractionated stereotactic radiotherapy (SRT), high dose of radiation is delivered to a precisely targeted area to minimize injury to the neighboring regions. Whereas in SRS the prescribed radiation

dose is delivered in a single fraction, in SRT the total radiation dose is delivered in very few fractions over few days.

Magnetic resonance imaging (MRI) is the main imaging modality for diagnosis, treatment planning, and therapy outcome evaluation in brain metastasis. MRI scans are acquired before (baseline) and at multiple follow-up sessions after the radiation therapy as part of the standard treatment planning and outcome assessment procedure. The procedure requires accurate delineation of the tumour that is often performed by expert radiation oncologists and neuro-radiologists. Evaluation of radiotherapy outcome in brain metastasis on serial MRI is mainly performed based on the standard criteria presented by the response assessment in neuro-oncology–brain metastases (RANO-BM) group [14]. The RANO-BM criteria are principally based on changes in the longest diameter of the target tumour in the axial, coronal, and sagittal planes compared to baseline or nadir (smallest tumour size on the previous scans) to specify its response to therapy. The four categories of therapy response based on the RANO-BM criteria include complete response (CR; no target tumour remaining), partial response (PR; more than 30% reduction in the longest diameter compared to baseline), stable disease (SD; less than 30% decrease compared to baseline but also less than 20% increase in the longest diameter compared to nadir), or progressive disease (PD; also referred to as local failure; more than 20% increase in the longest diameter compared to nadir). Tumour enlargement on MRI after radiotherapy may also become apparent due to adverse radiation effect (ARE). Such evident tumour enlargements on MRI often become stable or followed by a decrease in tumour size on subsequent imaging follow-ups. Differentiating between tumour progression and ARE is crucial for radiotherapy response evaluation. The standard approaches to diagnose ARE include serial MRI (including the use of T1-weighted, T2-weighted, and perfusion imaging), and where applicable, histology on resected specimens [15], [16], [17].

In order to calculate the tumour size changes on serial imaging, precise delineation of tumour is required for each imaging session. Manual segmentation of tumour on volumetric images acquired at several follow-up sessions for each patient is a tedious and time-consuming job. An automatic and robust tumour segmentation framework is highly desirable in the clinic and could streamline radiation therapy outcome evaluation workflow considerably. Because of many applications of automatic tumour segmentation, intense research has been carried out on this topic [18], [19], [20]. The existing segmentation algorithms include those that apply traditional methods such as region-based [21], [22] and model-based techniques [23], with more recent methodologies based on deep neural networks [24], [25], [26]. Deep learning-based image segmentation is now very popular in the literature and has demonstrated to outperform the traditional methods [27], [28], [29]. The deep networks for image segmentation generally consist of stacked convolutional layers and occasionally fully connected layers. Among many networks introduced for the task of segmentation, 2D and 3D U-Net gained widespread popularity because of their robustness in different modalities [28], [30]. However, 2D U-Net has the drawback of extracting similar features multiple times throughout the network in addition to inefficient modeling of long-range spatial dependencies. A main limitation associated with 3D U-Net is that it often cannot handle large input sizes due to memory limitations with the complex architecture of the network.

Deep learning-based techniques have demonstrated promising performance in brain tumour segmentation [31], [32], [33]. Despite previous research on the application of these techniques in feature extraction frameworks for classifying brain tumour subtypes and predicting clinical outcomes such as survival, their clinical efficacy in longitudinal monitoring of changes in tumour physical dimensions has not been investigated thoroughly. Cabezas et al. proposed an ensemble of 3D U-Nets to segment different sub-regions of gliomas on the BraTS dataset [34] to extract quantitative features for predicting the overall survival of patients. Gates et al. [35] proposed a multi-scale convolutional neural network based on the DeepMedic to segment glioma sub-volumes on MRI, and applied the features extracted from the segmented images and clinical data for predicting the overall survival. Pei et al. [36] proposed a context-aware deep learning model for brain tumour segmentation on MRI, followed by deep learning models for subtype classification and survival prediction using the tumour segments. Zhu et al. [37] developed a semi-automatic segmentation software for quantitative clinical evaluation of glioblastoma multiforme on MRI. While their results demonstrate a good correlation between the manual and semi-automatic segmentation, the developed method was not evaluated on serial MRI to quantify changes in tumour size.

In this work, a novel deep-learning-based system is introduced for automatic radiotherapy outcome assessment in brain metastases. A multi-step framework is proposed for automatic brain tumour segmentation that is applied for delineation of tumours before and at multiple imaging follow-ups after the radiotherapy to assess the therapy outcome automatically based on the RANO-BM criteria. To the best of our knowledge, this is the first time that a deep-learning-based segmentation framework is adapted and investigated comprehensively for automatic radiotherapy outcome assessment in brain malignancies.

II. Methods

A. Data Acquisition and Pre-Processing

This study was conducted in accordance with institutional research ethics approval from Sunnybrook Health Sciences Centre (SHSC), Toronto, Canada (project identification number: 2175, 2020/08/11). The imaging and clinical data were collected from 116 patients (152 tumours; average size at baseline: 2.4 $\pm$ 1.0 cm, range: 0.5-7 cm) diagnosed with brain metastasis and treated with hypo-fractionated SRT between March 2011 and December 2014 at SHSC. The patients (40.2% male, 59.8% female) were aged between 29 and 91 years (average age: 62 $\pm$ 15 years). Among the 116 patients, 86 patients had one, 24 patients had two, and 6 patients had three or more brain metastasis tumours. The primary tumour histology included lung cancer (76 tumours, 50%), breast cancer (36 tumours, 23.7%), melanoma (15 tumours, 9.9%), colorectal cancer (7 tumours, 4.6%), renal cell carcinoma (6 tumours, 3.9%), and other cancers (12 tumours, 7.9%). Lesions with prior resection were excluded. Any salvage

therapy was administrated after identifying tumour progression clinically that was the endpoint of this study. The imaging data included gadolinium-contrast-enhanced T1-weighted and T2-weighted-fluid-attenuation-inversion-recovery (T2-FLAIR) images acquired, as part of standard of care, before (baseline) and at up to 9 follow-ups after the treatment (average number of imaging follow-ups: 4). All available follow-up imaging data were used for post-treatment monitoring in this study. The dataset also included treatment-planning gross tumour volume (GTV) contours for each patient. All GTVs were contoured by an expert CNS radiation oncologist and reviewed by at least one other CNS radiation oncologist and a neuroradiologist. The GTVs were used to generate ground truth tumour masks for the baseline and follow-up scans under the supervision of expert oncologists. the MRI scans were acquired using a 1.5 T Ingenia system (Philips Healthcare, Best, Netherlands) and a 1.5 T Signa HDxt system (GE Healthcare, Milwaukee, WI, USA). The scan sequences were 3D T1w TFE (repetition time: 9.4 ms, echo time: 2.3 ms, imaging frequency: 127.77 MHz) and 3D T1w FSPGR (repetition time: 8.548 ms, echo time: 4.2 ms, imaging frequency: 63.86 MHz) for the T1-weighted images and T2 FLAIR CLEAR (repetition time: 9000 ms, echo time: 125 ms, imaging frequency: 127.77 MHz) and T2 FLAIR PROPELLER (repetition time: 8600 ms, echo time: 117 ms, imaging frequency: 63.86 MHz) for the T2-FLAIR images. The in-plane image resolution and the slice thickness were 0.5 and 1.5 mm for T1-weighted and 0.5 and 5 mm for T2-FLAIR images, respectively. All images were resampled with a voxel size of $0.5 \times 0.5 \times 1$ mm^3. The voxel intensities in each image were normalized to be between 0 and 1. The normalization was done on voxel level (*vox_intst*) using the following formula:

$$(voxel\_intst - min\_intst) \,/\, (max\_intst - min\_intst)$$

where the *min_intst* and *max_intst* are the minimum and maximum intensity values in the corresponding 3D image. The T2-FLAIR images were co-registered on their corresponding T1-weighted images using an affine registration. Among the 116 patients, 96 patients (130 tumours) were randomly selected for training the models, and the remaining 20 patients (22 tumours) were kept as an unseen test set for independent evaluation.

The tumours were monitored longitudinally on MRI after SRT and the pattern of changes in tumour size as well as the ground truth local control/failure (LC/LF) outcome for each tumour was determined by a radiation oncologist using the follow-up imaging data. The follow-up scans were performed every 2-3 months for all patients until they transitioned to palliative care or passed away. The ground truth tumour size status (decrease/stable/increase) was determined for each follow-up scan. Specifically, the tumour size status was determined as decrease/increase if a measurable ($\geq$2 mm) decrease/increase was evident in the longest diameter of the tumour in the axial plane compared to the previous scan, otherwise, it was determined as stable. The RANO-BM criteria were used to determine an outcome of LC (complete response, partial response, or stable disease) or LF (progressive disease) for each tumour separately [14]. Adverse radiation effect (ARE) was diagnosed and differentiated from local progression based on the report by Sneed et al. [15]. The ARE cases were diagnosed clinicoradiologically based on serial imaging, including the use of perfusion MRI (rCBV cut-off = 2) and chemical exchange saturation transfer (CEST) imaging, and/or through histological confirmation (available for 50% of tumours diagnosed with ARE) [16], [38].

B. Tumour Segmentation Framework

Fig. 1 presents a scheme of the proposed framework for automatic segmentation of brain tumours on MRI. The framework consists of two cascaded 2D U-Nets to find the approximate position of the tumour. Once the approximate tumour position is found, the image is cropped around the tumour to make the size of input image smaller for the next network. Specifically, the size of input T1-weighted images for the first and second 2D U-Nets is 512×512 and 256×256 pixels, respectively. The need for cropping images stems from the fact that both the 3D U-Net and multi-scale self-guided attention (MSGA) network [39] adapted in the framework have memory limitation which makes their training process challenging. If the input size for the 3D U-Net is the original image size ($512 \times 512 \times 128$ voxels) without cropping, one needs to patch or resize the input volume to meet the memory limitations of the network. Patching the volume leads to losing contextual information (e.g., tumour tears apart in different patches) while resizing it results in losing detailed local information. Similarly, and due to its complex architecture, training the MSGA network on the original 2D images (512×512 pixels) with two channels of T1-weighted and T2-FLAIR requires limiting the batch size. With cropping, it would be possible to preserve both local and contextual information using the approximate position of the tumour estimated with the cascaded 2D U-Nets. The output of each 2D U-Net for a patient is a set of 128 2D masks with size of 512×512 pixels for the first and 256×256 pixels for the second 2D U-Net. To find the approximate position of the tumour from these masks, a logical OR operation is applied on all the 2D masks to create a single mask presenting an upper-bound of the tumour areas in different slices. Subsequently, the connected components are identified in the single mask and the center of each connected component is regarded as the approximate center of the corresponding tumour. The approximated centers are used to crop the image around the tumour region. In cases where there is more than one tumour in an MRI volume, the tumours are treated separately, and the final masks are fused at the end. At the core of the framework there are two segmentation networks including a 3D U-Net and a MSGA network. The 3D U-Net is fed with the cropped T1-weighted volumetric images ($128 \times 128 \times 128$ voxels). The MSGA network is fed with cropped two-channel T1-weighted and T2-FLAIR co-registered image slices (128×128 pixels each). The output of these two networks is fused at the end through slice-wise averaging over their output probability maps. The final output masks are generated by thresholding the averaged probability maps with a threshold level of 0.5.

The choice of a combination of 2D U-Net, 3D U-Net, and MSGA network is to take advantage of their features, while

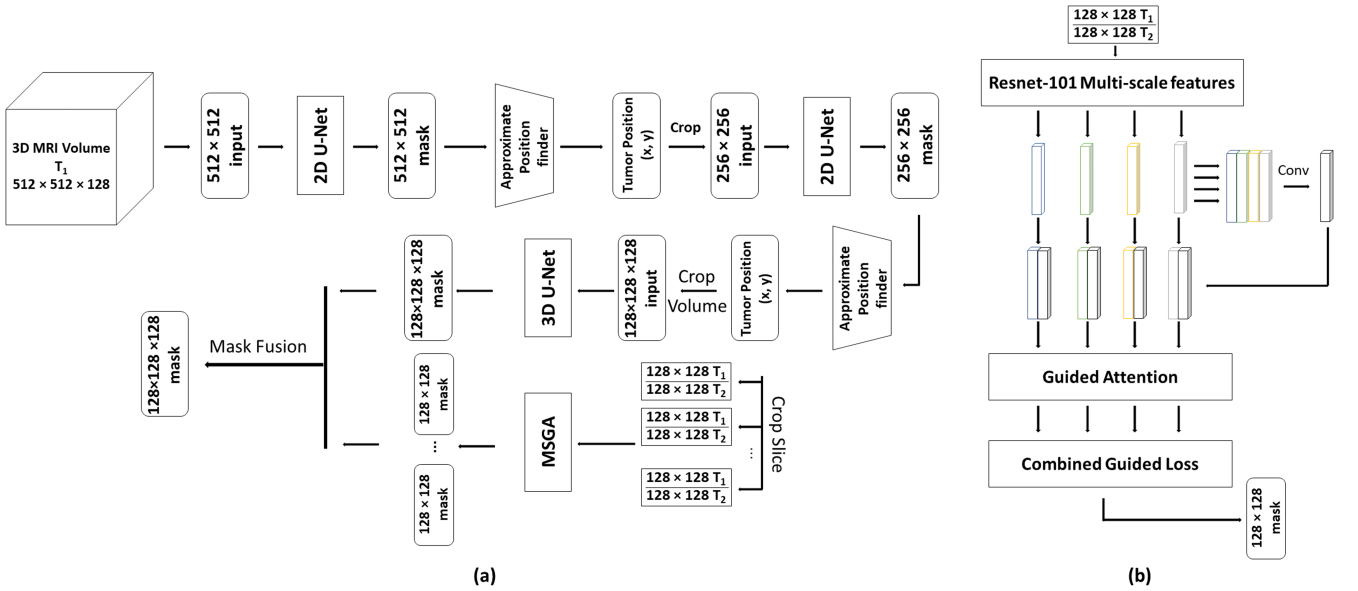

Fig. 1. (a) Overview of the proposed segmentation framework. For a volumetric input image (contrast-enhanced T1-weighted, 512×512×128 voxels), first all slices are fed to a 2D U-Net one by one. The generated masks from the 2D U-Net are used to finds an approximate tumour position *(x, y)*. The volumetric input image is then cropped around *(x, y)* into a 256×256×128 voxel volume. A similar procedure is performed to reduce the size of volumetric image containing the tumour to 128×128×128 voxels. This volume is then fed into a 3D U-Net for segmentation with no patching. The slices of this volume are also fed to MSGA, after concatenation with the co-registered T2-FLAIR image, and the segmentation masks of MSGA are then fused with those of 3D U-Net. (b) The MSGA network structure. Features extracted at different scales from Resnet-101 are concatenated and convolved and then self-concatenated and fed into guided attention module. The resulting self-guided features are fed into the guided loss.

simultaneously mitigating their limitations. More specifically, whereas the 2D U-Nets can effectively localize the region of interest even for smaller tumours to crop the large input image, it can not generate precise segmentation masks for all tumours. On the other hand, a localized input for the 3D U-Net and MSGA network reduces irrelevant information and enhances the model focus on the region of interest, leading to considerable improvements in their performance in generating precise segmentation masks. The good performance of the 2D U-Net architecture in various segmentation tasks is due to its capability to capture context and enable localization, using a contracting path and a symmetric expanding path, with skip connections in-between the two paths [30]. Such architecture enables the network to share features from multiple layers and overcome the trade-off between localization accuracy and context utilization. The drawback of 2D U-Net, however, is that it does not consider the 3D spatial dependencies between the voxels, and consequently, loses a considerable amount of useful information for segmentation. To overcome this, Çiçek et al. proposed the 3D U-Net as a volumetric image segmentation network [28], which maintains the benefits of the 2D U-Net architecture but also considers the voxel dependencies. Considering 3D spatial dependencies comes at the cost of high memory consumption because of the huge input size. A cascaded network of 2D U-Net and 3D U-Net could benefit from the advantages of 3D U-Net while the redundant information could be filtered out using 2D U-Net to meet the memory limitations of the 3D U-Net. The two main drawbacks of the encoder-decoder architectures such as 2D and 3D U-Net include deriving redundant information, and more importantly, inefficient modeling of long-range feature dependencies in these networks. Sinha et al. [39] proposed a multi-scale self-guided attention network to overcome these limitations. The MSGA network enables capturing richer contextual dependencies and neglecting irrelevant information by using an attention mechanism. Also, the utilization of interdependent channel maps which enables the network to integrate local features with their corresponding global dependencies makes it efficient in our application, where the network is fed with two channels of T1-weighted and T2-FLAIR images.

C. Training and Evaluation of the System

In order to train and evaluate the tumour segmentation framework, the data associated with samples of the training and test sets were completely separated at patient level. The networks in the framework were trained independently using the data acquired from the training samples. The second 2D U-Net, the 3D U-Net, and the MSGA network were trained using the manually cropped data from the training set. The networks were only trained on the images acquired at the baseline. This was done to permit evaluating the framework's performance on the training set at the first follow-up and compare it with the performance on the independent test set. The framework was initially evaluated in terms of segmentation accuracy, using the images of the independent test set acquired at the baseline and follow-up scans. The Dice similarity coefficient, Hausdorff distance, and the tumour volume estimation error were used for this evaluation. The performance of the system was subsequently evaluated in monitoring the tumour size status after SRT and automatic assessment of therapy outcome using the imaging data

of the independent test set acquired at the baseline and all follow-ups available for each patient. For comparison, experiments were conducted using seven different models following a similar training and evaluation procedure. The first model included two cascaded 2D U-Nets, the second model consisted of a 3D U-Net, and the third model included a 3D U-Net along with an MSGA network. For training and testing the standalone 3D U-Net and 3D U-Net + MSGA in the second and third models, each 512 × 512 × 128 voxel volume was patched into 16 input patches of 128 × 128 × 128 voxels and the associated masks were concatenated together at the end. The fourth model included two cascaded 2D U-Nets followed by a 3D U-Net, the fifth model utilized the framework proposed in this study inputting the T1-weighted image only, and the sixth model incorporated the complete framework proposed (Fig. 1). The seventh model utilized the well-recognized nnU-Net framework [40] for further comparison. The nnU-Net framework input the co-registered T1-weighted and T2-Flair images (512 × 512 × 128 voxels) as two channels, where each image was down-sampled to 128 × 128 × 128 voxels for the first 3D U-Net in the framework. Pre-training of the networks for weight initialization was performed using the data from the brain tumour segmentation (BraTS) dataset [34]. A set of 9 tumours from the training samples was used as the validation set for tuning the network hyperparameters in the training phase. A batch size of 4 and 2 was used for training the 2D U-Nets and nnU-Net, respectively. The batch size for the 3D U-Net and MSGA network was tuned to one. The training was performed with a learning rate of 0.0001 for all networks. Experimental results with different hyperparameters have been presented in Table S1 of the Supplementary Materials. A dice and a cross-entropy based loss function was used for the 2D and 3D U-Nets, respectively. The loss function for the nnU-Net was defined as the sum of the dice and cross-entropy losses. The dice loss function was defined as *(1 – dice coefficient)*, where *dice coef.* $= 2TP/(2TP + FN + FP + smooth)$. A smoothing term was added in the dice coefficient to prevent division by zero. Instead of setting Boolean intensity values for the ground truth and the automatically generated masks and performing Boolean operations, the mask intensities were defined as continuous values to make the dice loss differentiable. The cross-entropy loss was defined as $-1/N \sum_i^N \sum_j^M (y_{ij}.log(p_{ij}))$ where N and M are the number of pixels and classes (in our case two, tumour vs normal tissue), respectively. The loss function for the MSGA network was defined as the summation of three terms: L_{seg_total}, L_{G_total}, and L_{rec_total}. L_{rec_total} is the mean squared error between the original input and output features of the encoder-decoder network in the attention module. L_{G_total} is the mean squared error between the encoded representation of features in the encoder-decoder network inside the attention module. Finally, L_{seg_total} is the cross-entropy between the ground-truth and network output masks. The training and validation loss for the 3D U-Net and MSGA networks over the training epochs are presented in Fig. S1 of the Supplementary Materials. The framework was developed on an Nvidia GeForce RTX 2080 Ti with 12 GB of Memory. All models were developed in Python and trained and tested using Keras with TensorFlow backend.

D. Procedure and Criteria for Automatic Assessment of Tumour Size Status, Local Response, and ARE Outcome

The segmentation masks generated by the deep learning models were used to estimate the size of tumour in each scan and, subsequently, the tumour size changes after SRT. The tumour size status, local response, and the ARE outcome were then assessed automatically based on the estimated changes in tumour size using the procedure and criteria described below.

A typical SRT outcome evaluation workflow in the clinic consists of determining the tumour size status at each follow-up scan compared to the previous scan. For automatic assessment of tumour size status, following the protocol applied in clinic, the longest diameter of tumour in the axial plane was calculated for all scans using the automatic segmentation masks. Tumour size status at each follow-up scan was labeled as increase or decrease if a measurable increase or decrease ($\geq$ 2 mm) was estimated, respectively, in the tumour's longest diameter compared to the previous scan. Otherwise, it was labeled as stable. The tumour size status labels identified automatically were compared with the ground truth labels to evaluate the performance of automatic labeling in terms of accuracy, precision, and recall. It should be noted that this step was only to evaluate the performance of the network in automatic labeling of tumour size status and not the local response (discussed below).

The SRT outcome in terms of LC/LF and ARE was evaluated for each tumour automatically based on the RANO-BM criteria. Using the automatic segmentation masks, the longest diameter of tumour in the axial, coronal, and sagittal planes was estimated for the baseline and all follow-ups. The relative change in the longest diameter of tumour was calculated at each follow-up compared to the baseline and nadir. The change in the tumour diameter at each follow-up was categorized into three categories of shrinkage, steady, and enlargement when more than 30% decrease compared to baseline, less than 30% decrease compared to baseline but also less than 20% increase compared to nadir, and more than 20% increase compared to nadir was detected in the tumour longest diameter, respectively [14]. Further, the relative change in tumour volume was calculated at each follow-up compared to the baseline and nadir. The change in the tumour volume at each follow-up scan was categorized into three categories of shrinkage, steady, and enlargement based on the volumetric response assessment criteria proposed by Oft et al. [41] which is an extension of the RANO-BM guideline recommendations for volumetric response assessment. Specifically, shrinkage at a follow-up scan was defined as more than 65% reduction in tumour volume compared to baseline, steady as less than 65% reduction compared to baseline but also less than 72.8% increase compared to nadir, and enlargement as more than 72.8% increase in tumour volume compared to nadir. The categories detected at each follow-up scan using the automatic segmentation models were compared to those identified from the ground-truth segmentation masks to evaluate the performance of automatic response categorization at individual follow-ups in terms of accuracy, precision, and recall.

The shrinkage/steady/enlargement patterns determined based on the longest diameter for each tumour at the follow-up scans

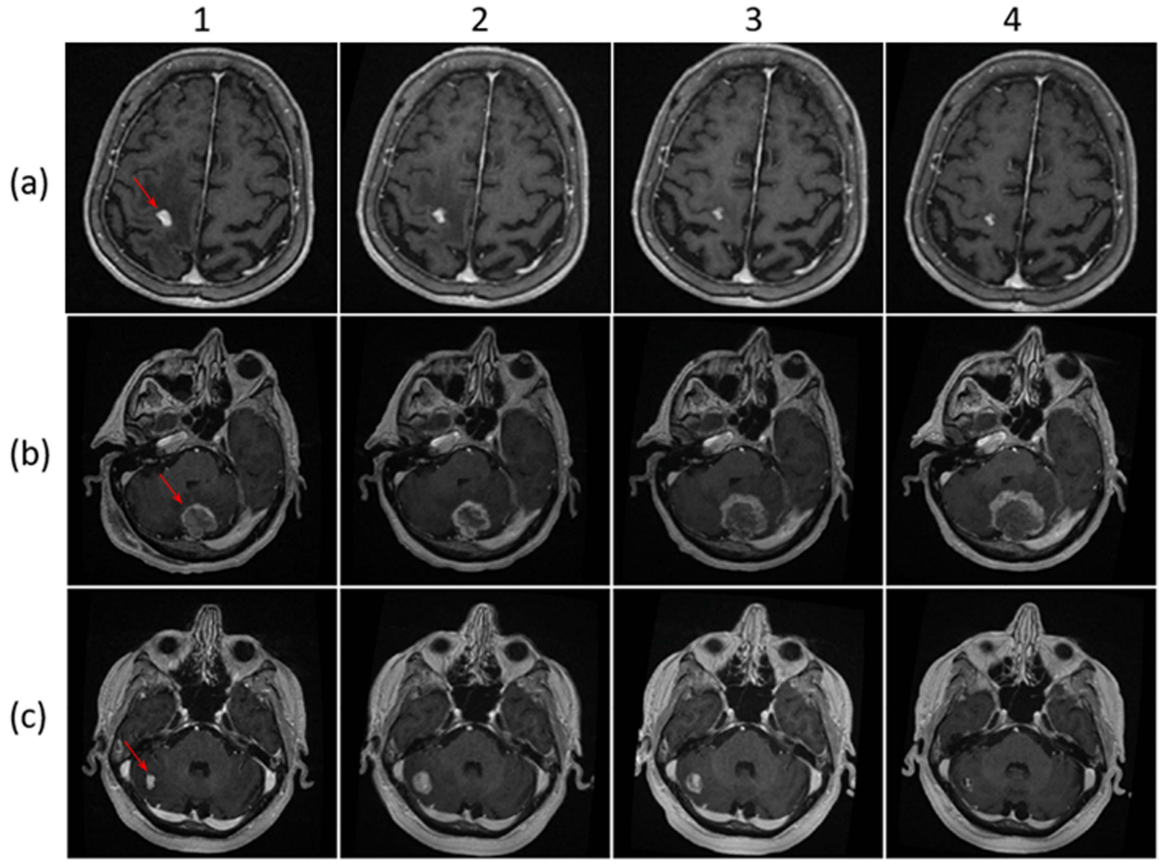

Fig. 2. Contrast-enhanced T1-weighted images acquired at the baseline (1), and the first (2), second (3), and third (4) follow-ups after SRT from three representative patients with brain metastasis demonstrating local control (a), local failure (b), and ARE (c) after treatment. The arrow in the baseline image shows the location of brain metastasis. LC/LF/ARE is evaluated based on the changes in longest diameter. In (c) an initial growth in first follow-up is followed by a decrease in the second, and then third follow-ups.

were used for automatic detection of LC/LF and ARE outcome. Any tumour demonstrating a sequence of steady or shrinkage patterns at follow-ups with no enlargement was classified with an LC outcome. When an enlargement was detected in the pattern of size changes, the change in the tumour longest diameter at the next follow-up was calculated compared to the scan in which the enlargement was detected. The tumour was classified with an LF outcome if its size increased again (more than 2 mm to account for measurement errors) compared to the previous scan. If the tumour size decreased or remained stable after the initial enlargement, the tumour was classified as LC but with ARE. As a tumour with ARE could possibly progress later and be classified as LF, detection of LC/LF and ARE outcome was performed and evaluated independently for each tumour. The outcomes identified automatically were compared with the ground truth outcome for each tumour to evaluate the performance of the automatic outcome assessment in terms of accuracy, sensitivity, and specificity.

III. Results

Fig. 2 demonstrates contrast-enhanced T1-weighted images acquired from three representative brain metastasis patients with an outcome of LC, LF, and ARE after SRT, respectively. In Fig. 2(a), the tumour has consistently shrunk after SRT (follow-ups 1-3), demonstrating an LC outcome. In Fig. 2(b) the tumour has continued to grow after the first follow-up, showing an LF outcome. In Fig. 2(c), initial growth in the first follow-up stopped immediately in the second follow-up, followed by further shrinkage in the third follow-up, that is evidence for ARE.

Fig. 3 shows the ground truth and automatic tumour segmentation masks generated by different deep learning models for five representative patients of the test set. The images show a step-by-step improvement in the automatic segmentation masks generated by the cascaded 2D U-Nets, 3D U-Net, cascaded 2D & 3D U-Nets, and the complete segmentation framework proposed in this paper (cascaded 2D & 3D U-Nets + MSGA). Specifically, the proposed frameworks could achieve a close to perfect segmentation for cases (a) and (c), while for cases (b) and (d) it slightly under-segmented the tumour, and for case (e) the results indicate over-segmentation. In general, the results demonstrate that the model is not biased towards under- or over-segmentation. A detailed comparison between the segmentation results of different networks at the baseline and follow-up sessions is given in Table I in terms of dice similarity coefficient, Hausdorff distance, and tumour volume estimation error. A consistent step-by-step improvement is observed in different criteria of segmentation accuracy, with the best results associated with

TABLE I
DICE SIMILARITY COEFFICIENT (DSC), HAUSDORFF DISTANCE (HD), AND VOLUME ESTIMATION ERROR (VEE) FOR SEGMENTATION OF BRAIN METASTASIS AT THE BASELINE AND FOLLOW-UP SCANS USING DIFFERENT NETWORK ARCHITECTURES

Segmentation Model	Metric	Baseline		1st Follow-up		2nd Follow-up	3rd Follow-up	4th Follow-up	5th Follow-up
		Training Set	Test Set	Training Set	Test Set	Test Set	Test Set	Test Set	Test Set
Cascaded 2D U-Nets	DSC	86.5 ± 5.8	85.4 ± 7	82.8 ± 6	81.3 ± 5.9	79.7 ± 11	77.2± 9.7	76.1 ± 10.6	74.3 ± 11.2
	HD (mm)	2.8 ± 0.4	3 ± 0.6	3.2 ± 0.6	3.7 ± 0.5	4.6 ± 0.7	4.4 ± 0.6	4.3 ± 0.7	4.5 ± 0.9
	VEE (cc)	0.55 ± 0.5	0.58 ± 0.5	0.64 ± 0.5	0.67 ± 0.5	0.82 ± 0.7	0.78 ± 0.7	0.75 ± 0.6	0.75 ± 0.7
	VEE (%)	15.8 + 7	16.4 ± 9	18.3 ± 8.4	19.2 ± 8.3	23.7 ± 11	24.5 ± 12.5	26.1 ± 11.9	30.2 ± 14
3D U-Net	DSC	88.8 ± 4.5	87.2 ± 5.4	84.8 ± 5.5	83.7 ± 6.4	83 ± 8	81.6 ± 7.7	80 ± 8	79.8 ± 8
	HD (mm)	2.4 ± 0.7	2.6 ± 0.7	2.6 ± 0.6	3.3 ± 0.6	4.2 ± 0.7	4.3 ± 0.7	4.5 ± 0.6	4.2 ± 0.6
	VEE (cc)	0.50 ± 0.4	0.52 ± 0.5	0.56 ± 0.5	0.60 ± 0.5	0.79 ± 0.7	0.74 ± 0.6	0.73 ± 0.7	0.73 ± 0.6
	VEE (%)	14.9 ± 5.2	15.3 ± 6.8	17.5 ± 6.3	17.8 ± 9.1	21.3 ± 9.4	23.3 ± 12	24.1 ± 9.4	24.7 ± 9.4
3D U-Net + MSGA	DSC	89.7 ± 5	88.9 ± 5.3	85.6 ± 5.1	84.8 ± 6.2	84.1 ± 7.7	83.5 ± 8.3	82.4 ± 8.8	82.1 ± 9.6
	HD (mm)	2.3 ± 0.5	2.5 ± 0.7	2.6 ± 0.5	3.2 ± 0.8	4 ± 0.7	3.5 ± 0.7	3.6 ± 0.7	3.6 ± 0.7
	VEE (cc)	0.51 ± 0.4	0.52 ± 0.4	0.55 ± 0.4	0.59 ± 0.5	0.74 ± 0.7	0.74 ± 0.7	0.72 ± 0.7	0.71 ± 0.7
	VEE (%)	12.3 ± 4.3	13.6 ± 6	16.9 ± 8.5	18 + 8.8	19.3 ± 7.8	21.5 ± 6.8	22.5 ± 9	23.3 ± 12
Cascaded 2D & 3D U-Nets	DSC	90.1 ± 4.4	89.6 ± 4.6	86.2 ± 4.6	85.1 ± 5	84.3 ± 7.2	83.4 ± 7	82.9 ± 7	82.8 ± 7
	HD (mm)	2.3 ± 0.2	2.4 ± 0.4	2.6 ± 0.5	3.1 ± 0.8	3.8 ± 0.7	3.2 ± 0.6	3.5 ± 0.6	3.4 ± 0.6
	VEE (cc)	0.5 ± 0.3	0.51 ± 0.4	0.55 ± 0.4	0.57 ± 0.5	0.73 ± 0.6	0.71 ± 0.6	0.72 ± 0.6	0.71 ± 0.7
	VEE (%)	11.1 ± 4.2	12.5 ± 5.3	16.7 ± 8.3	18.2 ± 8	18.7 ± 7.5	20 ± 7.5	21.7 ± 9	22.6 ± 10
Cascaded 2D & 3D U-Nets + MSGA (with T1-weighted only)	DSC	91.1 ± 3.8	90.3 ± 4.1	87.1 ± 3.9	86.4 ± 4.9	85.3 ± 7.2	83.7 ± 8.4	83.2 ± 8.5	82.8 ± 9.2
	HD (mm)	2 ± 0.4	2.3 ± 0.5	2.43 ± 0.5	2.95 ± 0.7	3.7 ± 0.7	3.4 ± 0.7	3.5 ± 0.7	3.4 ± 0.6
	VEE (cc)	0.42 ± 0.3	0.49 ± 0.4	0.55 ± 0.4	0.58 ± 0.5	0.72 ± 0.6	0.68 ± 0.6	0.68 ± 0.6	0.69 ± 0.6
	VEE (%)	10.5 ± 4.8	11 ± 5	14.6 ± 5.4	16.4 ± 5.4	17.9% ± 6.8%	18.7% ± 7%	19.3% ± 8.2%	23.8 ± 10.1%
Cascaded 2D & 3D U-Nets + MSGA	DSC	**92.3 ± 3.1**	**91.5 ± 3.7**	**88.7 ± 3.7**	**87.4 ± 5.2**	**86.7 ± 5.5**	85.1 ± 6.1	**84.1 ± 6.8**	**84.5 ± 7**
	HD (mm)	**1.84 ± 0.4**	**2.1 ± 0.6**	2.21 ± 0.5	**2.84 ± 0.7**	**2.98 ± 0.7**	3 ± 0.6	**2.89 ± 0.6**	**2.9 ± 0.58**
	VEE (cc)	**0.39 ± 0.3**	**0.44 ± 0.4**	**0.52 ± 0.4**	**0.57 ± 0.5**	**0.59 ± 0.5**	**0.61 ± 0.5**	**0.62 ± 0.6**	**0.6 ± 0.56**
	VEE (%)	**9.2 ± 4.6**	**10.2 ± 5.3**	**12.5 ± 4**	**13.4 ± 5.1**	**15.7% ± 6.3%**	16.5% ± 7%	**17.3% ± 8.3%**	**19.7 ± 8.5%**
nnU-Net	DSC	91.7 ± 3.3	90.9 ± 3.1	88.5 ± 3.2	87.1 ± 5.9	86.6 ± 6	**85.3 ± 6.5**	84 ± 7	84.2 ± 8.1
	HD (mm)	1.88 ± 0.4	2.3 ± 0.57	**2.18 ± 0.5**	2.95 ± 0.8	3 ± 0.8	**2.9 ± 0.7**	2.93 ± 0.61	2.9 ± 0.66
	VEE (cc)	0.43 ± 0.4	0.5 ± 0.3	**0.52 ± 0.4**	0.63 ± 0.5	0.67 ± 0.5	0.61 ± 0.6	0.63 ± 0.5	0.61 ± 0.6
	VEE (%)	9.8 ± 4.2	11 ± 5.5	12.5 ± 4.1	15.1 ± 5.7	16.3% ± 6.8	**16.4% ± 7**	17.5% ± 8.2	20.2 ± 8.8

the cascaded 2D & 3D U-Nets + MSGA architecture inputting both the T1-weighted and T2-FLAIR images. The networks demonstrate a similar performance of the training and test sets, implying very good generalizability for tumour segmentation of new unseen cases. Further, the segmentation results of the proposed framework are comparable between the baseline and follow-up scans. It should be noted that in the experiments conducted in this study, no data from the follow-ups were used for training the networks. Specifically, the networks were solely trained using the data of the training set patients acquired at the baseline, but subsequently evaluated using the follow-up data from the patients of the training and test sets, separately. The segmentation results for separate categories of baseline tumour size are presented in Table S2 of the Supplementary Materials. The results demonstrate that the proposed framework outperformed the other segmentation models in most cases, with few cases of performance in par with the nnU-Net model. Results of statistical comparisons on percental tumour size changes at each follow-up session relative to the baseline are presented in Table S3 of the Supplementary Materials. The table includes the results of Pearson correlation analyses and paired t-tests (two-sided, $\alpha = 0.05$) performed on the tumour size changes obtained using the automatically generated segmentation masks compared to those based on the ground-truth masks. The table demonstrates good correlations between the percentual tumour size changes estimated automatically compared to the ground truth at different follow-ups, with no statistically significant difference between them, where the proposed segmentation framework and the nnU-Net model demonstrate better results compared to the other models.

Table II presents the results of detecting tumour size status at the imaging follow-ups after SRT for patients of the test set using the five different segmentation models. The cascaded 2D & 3D U-Nets + MSGA architecture demonstrated the best performance with an accuracy of 85.9%, while the nnU-Net model resulted in an accuracy of 84.4%. The results of detecting the shrinkage/steady/enlargement categories at individual follow-up scans are presented in Table III. Here, the proposed framework demonstrated a similar performance to that of nnU-Net in terms of accuracy when the tumour size changes were categorized based on the longest diameter of tumour, but it outperformed the nnU-Net model when change in the tumour volume was used as the measurement method.

Table IV reports the results of automatic outcome assessment for the test set patients using five segmentation models. The results demonstrate that the proposed framework and the nnU-Net

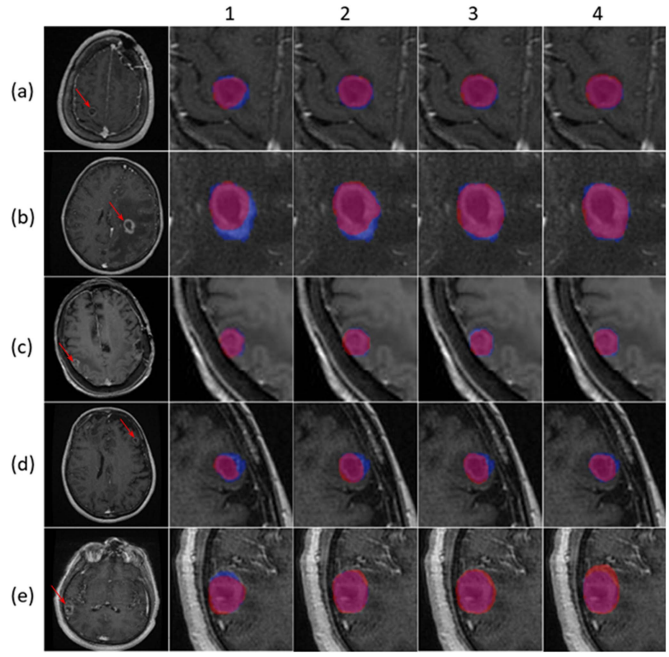

Fig. 3. Tumour segmentation masks generated by the cascaded 2D U-Nets (1), 3D U-Net (2), cascaded 2D & 3D U-Nets (3), and Cascaded 2D & 3D U-Nets + MSGA (4) for five representative patients (a-e) in the test set. The images are acquired at the baseline. The arrow in the first image of each row shows the location of brain metastasis. The ground truth and automatic segmentation masks (middle slide) are shown in blue and red, respectively, with the purple area showing the overlap region in each case.

TABLE II
RESULTS OF DETECTING TUMOUR SIZE STATUS AT FOLLOW-UP SESSIONS AFTER SRT FOR THE PATIENTS OF TEST SET USING DIFFERENT SEGMENTATION MODELS

Segmentation Model	Tumour Size Status	Accuracy	Precision	Recall
Cascaded 2D U-Nets	Increase	71.8%	82.3%	70%
	Stable		57.6%	82.6%
	Decrease		92.8%	62.9%
3D U-Net	Increase	79.6%	84%	80%
	Stable		67.8%	82.6%
	Decrease		94%	76.4%
Cascaded 2D & 3D U-Nets	Increase	82.8%	**90%**	**90%**
	Stable		70%	**91.3%**
	Decrease		**100%**	66.7%
Cascaded 2D & 3D U-Nets + MSGA	Increase	**85.9%**	**90%**	**90%**
	Stable		75%	**91.3%**
	Decrease		**100%**	76.2%
nnU-Net	Increase	84.4%	81.8%	**90%**
	Stable		**76%**	82.6%
	Decrease		**100%**	**81%**

model resulted in the best performance with a sensitivity and specificity of 88.9% and 92.3%, respectively, for detecting the LC/LF, and 100% and 89.2% for detecting the ARE outcome. Kaplan-Meier analyses were conducted to compare the time to detected event for LF and ARE based on the clinical radiotherapy outcome assessment and the assessment performed by the proposed automatic system. A log-rank test was applied to evaluate

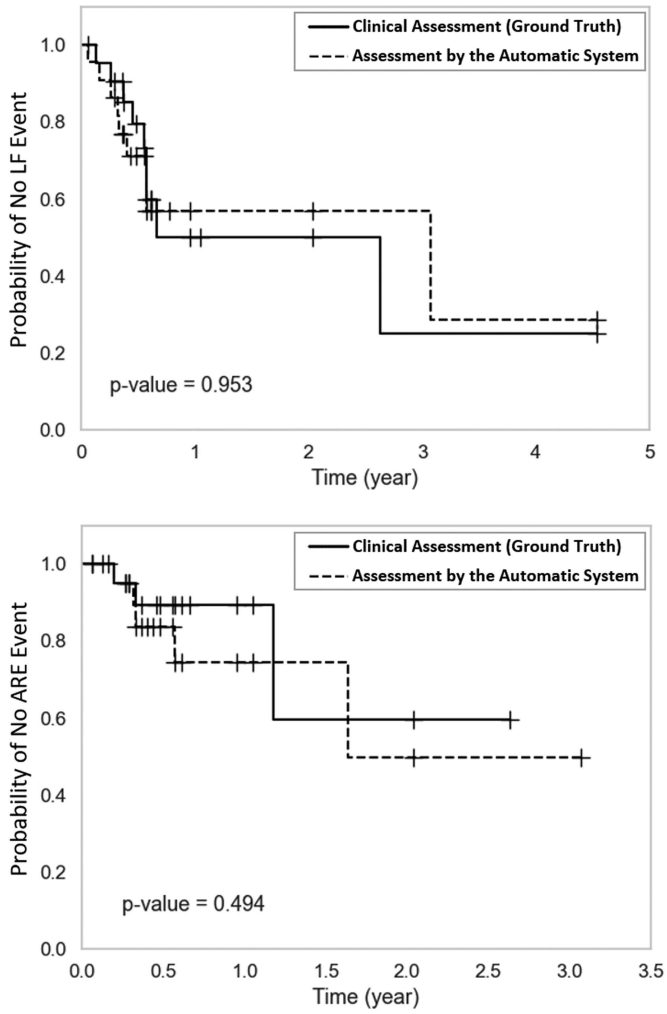

Fig. 4. Kaplan-Meier curves for comparative time to event analysis between the clinical radiotherapy outcome assessment and the assessment performed by the proposed automatic system on the test set. The plots have been shown for the LF (top) and ARE (bottom) events. The time to event for each tumour was calculated from the date of radiotherapy to the date an LF/ARE was detected clinically or by the automatic system using the proposed segmentation framework.

for any statistically significant difference between the curves for each event. Fig. 4 demonstrates the Kaplan-Meier curves for the LF and ARE events. The curves obtained for the automatic system are similar to their clinically assessed counterparts. No significant difference was observed between the curves for the LF (p-value = 0.95) or ARE (p-value = 0.49) event.

IV. DISCUSSION AND CONCLUSION

In this work, a novel system was proposed for automatic assessment of therapy outcome in brain metastasis patients treated with SRT. At the heart of the proposed system is a deep learning-based segmentation framework to delineate tumours longitudinally in serial MRI with high precision. Longitudinal segmentation of tumour before and at multiple follow-up sessions after the SRT permits monitoring changes in tumour size

TABLE III
RESULTS OF DETECTING THE RANO-BM RESPONSE CATEGORIES AT INDIVIDUAL FOLLOW-UP SCANS FOR THE PATIENTS OF TEST SET USING DIFFERENT SEGMENTATION MODELS, VALIDATED BASED ON THE RESPONSE CATEGORIES IDENTIFIED FROM THE GROUND-TRUTH SEGMENTATION MASKS

Segmentation Model	Tumour Size Status (Response Category)	Measurement Method					
		Longest Diameter of Tumour			Tumour Volume		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall
Cascaded 2D U-Nets	Enlargement (PD)	72%	70%	70%	68.8%	68.4%	65%
	Steady (SD)		71.4%	74.1%		70.4%	70.4%
	Shrinkage (PR)		75%	70.6%		66.7%	**70.6%**
3D U-Net	Enlargement (PD)	78.1%	77.8%	70%	71.9%	75%	75%
	Steady (SD)		81.5%	81.5%		79.2%	70.4%
	Shrinkage (PR)		73.7%	**82.4%**		60%	**70.6%**
Cascaded 2D & 3D U-Nets	Enlargement (PD)	81.3%	**83.4%**	75%	75.4%	75%	75%
	Steady (SD)		82.1%	**85.1%**		78.6%	78.6%
	Shrinkage (PR)		77.8%	**82.4%**		70.6%	**70.6%**
Cascaded 2D & 3D U-Nets + MSGA	Enlargement (PD)	**84.4%**	78.3%	**90%**	**81.3%**	**76.2%**	**80%**
	Steady (SD)		**91.7%**	81.5%		**85.7%**	**88.9%**
	Shrinkage (PR)		82.4%	**82.4%**		**80%**	**70.6%**
nnU-Net	Enlargement (PD)	**84.4%**	78.3%	**90%**	79.7%	**76.2%**	**80%**
	Steady (SD)		88%	81.5%		85.2%	85.2%
	Shrinkage (PR)		**87.5%**	**82.4%**		75%	**70.6%**

TABLE IV
RESULTS OF DETECTING THE LC/LF AND ARE OUTCOMES FOR THE PATIENTS OF TEST SET BASED ON THE RANO-BM CRITERIA USING DIFFERENT SEGMENTATION MODELS. SENS: SENSITIVITY; SPEC: SPECIFICITY; ACC: ACCURACY

Segmentation Model	LC/LF Detection			ARE Detection		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
Cascaded 2D U-Nets	72.7%	66.7%	76.9%	77.2%	66.7%	79%
3D U-Net	81.9%	77.8%	84.6%	81.9%	66.7%	84.2%
Cascaded 2D & 3D U-Nets	86.3%	77.8%	**92.3%**	86.4%	**100%**	84.2%
Cascaded 2D & 3D U-Nets + MSGA	**90.9%**	**88.9%**	**92.3%**	**90.9%**	**100%**	**89.2%**
nnU-Net	**90.9%**	**88.9%**	**92.3%**	**90.9%**	**100%**	**89.2%**

for automatic assessment of therapy outcome based on standard clinical criteria.

The segmentation framework was designed such that it can tackle the memory limitations associated with effective training of complex deep networks by cropping the volumetric images around the tumour. Two cascaded 2D U-Nets were trained to find the approximate position of the tumour. This position is later used to crop the MRI volume around the tumour. Experimental results show that the cascaded 2D & 3D U-Net model could considerably improve the segmentation accuracy compared to the cascaded U-Nets and the 3D U-Net alone. Further, the segmentation framework proposed in this study outperformed the cascaded 2D U-Nets, the 3D U-Net, the cascaded 2D & 3D U-Net, and the nnU-Net models. By incorporating the MSGA network into the framework, the model benefits from both the cascading and ensembling mechanisms to improve the segmentation accuracy [42]. The MSGA network applies a multi-scale attention mechanism to focus on crucial regions of the images and discard redundancies in the extracted features while learning tumour segmentation. Also, complementary information is provided to the framework through MSGA by feeding T2-FLAIR images as an additional input channel to the MSGA network. As such, fusing the outcome of this network with the 3D U-Net potentially improves the overall performance of the segmentation framework, as observed in this study. Performance of the proposed system was subsequently evaluated in monitoring the tumour size status at several imaging follow-ups after SRT. Experimental results demonstrated an accuracy of 86% in detecting tumour size status (increase/stable/decrease), on the independent test. It should be noted though, that these labels were manually determined at each follow-up by only one observer, and therefore labelling error is expectable due to measurement errors, especially for smaller tumours and those lying closer to the class boundaries. Such errors may affect the reported accuracies in automatic labeling of the tumour size status. Future studies may mitigate possible errors in ground truth labeling of tumour size status using a multiple observer strategy.

The proposed system also demonstrated a promising performance in detecting tumour size status in terms of response categories at individual follow-up scans, and subsequently automatic assessment of SRT outcome (LC/LF and ARE) on the independent test set. The automatic outcome assessment system in this study evaluates the presence of ARE after radiotherapy based on the pattern of changes in tumour size on serial MRI, with acceptable accuracy. However, it should be noted that monitoring tumour size changes on serial imaging is not always enough to draw an accurate conclusion on whether an observed tumour size increase on imaging is associated with progressive disease or ARE. Along with other radiological insights such as those based on T1/T2 matching or use of perfusion MRI [17], [43], additional clinical evidence including histological confirmation is sometimes required to diagnose ARE. As such, standard serial MRI is usually used by oncologists in conjunction with other clinical criteria to detect pseudo-progression or radiation necrosis after radiotherapy. Considering the performance of the proposed system in accurate tumour segmentation, monitoring tumour size changes longitudinally, and detecting LC/LF and ARE outcomes, it can be applied as an effective decision support system for radiotherapy outcome assessment to triage complicated boundary cases that required further assessment by clinicians.

Previous studies have shown the potential of deep-learning-based methods in automatic brain tumour segmentation and assessment of tumour size changes in response to treatment. Xue et al. [44] proposed a cascade of modified 3D U-Net architecture for detection and segmentation of brain metastases on 3D T1 MPRAGE images. They proposed the utility of automatically generated segmentation masks for facilitating radiotherapy treatment planning and post-treatment monitoring of tumour size, where they demonstrated example results for one case. Cho et al. [45] developed a CAD system for automated brain metastasis detection on MRI using a U-Net based cascaded model and applied it for categorizing tumour size changes at two follow-up sessions separately, where they achieved a moderate agreement with the RANO-BM criteria. The study here features a novel deep-learning-based system for automatic assessment of radiotherapy outcome in brain metastasis using an attention-guided architecture for accurate tumour segmentation. The system was evaluated on multiple MRI scans for each patient to demonstrate its performance in precise tumour segmentation and monitoring tumour size status at individual follow-up sessions, and in detecting LC/LF and ARE outcomes after SRT using the pattern of tumour size changes on serial MRI. The system was also evaluated in terms of similarity of time to detected LF and ARE events compared to those identified clinically. To our knowledge, this is the first time a comprehensive study is performed to investigate the efficacy of deep-learning-based segmentation frameworks for automatic radiotherapy outcome assessment. The findings of this study are in agreement with observations of the previous papers where the potential of data-driven segmentation models was shown in monitoring tumour size changes after treatment, while it extends the preliminary investigations by developing a novel segmentation framework and demonstrating its promising performance for various tasks within a radiotherapy outcome assessment workflow.

Objective assessment of tumour response to therapy has been the basis for many investigations in cancer therapeutics during recent years [46]. The RANO-BM criteria and recommendations were proposed to establish a basis for standard response assessment in clinical trials for brain metastasis. The improved uniformity in response assessment following the RANO-BM criteria facilitates the interpretation of studies involving patients with brain metastasis. This is especially important as the new trend is away from automatically excluding patients with active brain metastasis from the clinical trials of novel therapies [47]. A number of previous studies have explored the RANO-BM criteria as a tool for objective response assessment. Douri et al. [48] evaluated the RANO-BM criteria's current threshold in a cohort of 50 patients with brain metastasis treated by SRS. Their findings show that the current RANO-BM thresholds are useful in assessing diameter increases caused by tumor progression and pseudo progression, but may need adjustments to identify clinically relevant tumour progression reliably. Fischedick et al. [49] compared the 2D linear and 3D volumetric measurement methods for post-SRT monitoring of brain metastasis. The 2D and 3D measurements were categorized according to the RANO-BM criteria and Matthew J. et al. [50], respectively. They concluded that results obtained from the 2D and 3D measurements are highly comparable. While the criteria proposed for volumetric analysis in the RANO-BM guidelines are incomplete due to lack of research to support specific recommendations, Oft et al. [41] adopted the basic concept from the RANO-BM guideline to derive volumetric criteria and investigated the predictors for volumetric regression after SRT. Their result show that volumetric regression post-SRT does not occur at a constant rate, and a cut-off of $\geq$20% regression for the volumetric definition of response at 3 months post-SRT was predictive for subsequent control. Further research is required to validate specific threshold recommendations for volumetric monitoring of brain metastasis after radiotherapy. The automatic system proposed in this paper can facilitate such investigations in future and is a step forward towards a volumetric radiotherapy response assessment paradigm.

There is a huge interest in finding reliable clinical and/or imaging features that would assist in distinguishing ARE from tumour progression to limit the number of cases triaged for diagnostic biopsy or surgical resection [51]. Various methods such as those based on the qualitative [17] and quantitative [52] assessment of T1/T2 matching, and perfusion [53], [54], and CEST [38] MRI have shown relatively effective with different degrees of accuracy to differentiate ARE from tumour progression. Accurate segmentation of tumour on MRI is a prerequisite for all these methods. Wiggenraad et al. have investigated the use of cine-loops for monitoring tumour size changes in brain metastasis after SRT to identify pseudo-progression (ARE) [55]. They created the cine-loops for ten patients using the axial slice with largest tumour diameter on pre-treatment contrast-enhanced T1-weighted MRI and the corresponding slices in the co-registered follow-up images. The cine-loops were evaluated by a group of radiation oncologists and neuroradiologists for interpretation of events after SRT, where it was concluded that the use of cine-loops was superior to assessment of separate MRI scans. To our knowledge, no previous study has investigated the application of automatic brain tumour segmentation on serial MRI for monitoring the pattern of tumour size changes to detect ARE.

One potential limitation associated with this study is its relatively small cohort size. Here, several MRI datasets acquired at different imaging sessions for each patient were applied to evaluate the proposed framework. While the results presented are encouraging and pave the way for future studies, more investigations are required for further evaluation of the proposed methodologies on larger patient populations and possibly multi-centre imaging data. The patients in this study had relatively large brain metastases treated with hypo-fractionated SRT. Although tumours with size of 5 mm and above were included in this study, future studies focusing on tumours with size of less than 1 cm are required for further assessment of the performance of the framework on smaller brain metastases typically treated with SRS. ARE in this study was diagnosed clinicoradiologically based on serial imaging, and/or histological confirmation. Diagnosing ARE clinicoradiologically without histological confirmation, however, may be prone to errors due to misinterpretation of images in complicated cases. As such, future studies on imaging datasets with ground truth histology for all ARE cases are necessary for further validation of the results of this study.

The proposed segmentation framework demonstrated good generalizability in longitudinal segmentation of brain tumours on serial MRI, while it was only trained on the baseline images of the training set. The generalizability of the proposed framework makes it an appropriate fit for the task of automatic therapy outcome assessment. Implementation of the proposed system in clinical settings can potentially accelerate longitudinal tumour size analyses, streamline image-guided therapy outcome evaluation workflows, e.g., for local response assessment and ARE detection, and facilitate precision oncology through regular and high-throughput response assessment. This is particularly important in case of patients with multiple brain metastases where manually segmenting tumours on several follow-up scans puts a substantial burden on clinical workflow. The system can possibly be coupled with PACS-based databases to perform online and/or offline tumour size analyses on serial imaging and act as an invaluable decision support tool in clinic. Although a more comprehensive study is a prerequisite to further validate the results of this study and the clinical utility of the proposed system, the promising results obtained here and the prospect of its real-world applications highlight the importance of the findings in this paper.

References

[1] R. M. Auchter et al., "A multiinstitutional outcome and prognostic factor analysis of radiosurgery for resectable single brain metastasis," *Int. J. Radiat. Oncol.*, vol. 35, no. 1, pp. 27–35, Apr. 1996, doi: 10.1016/S0360-3016(96)85008-5.

[2] B. D. Fox, V. J. Cheung, A. J. Patel, D. Suki, and G. Rao, "Epidemiology of metastatic brain tumors," *Neurosurgery Clin. North Amer.*, vol. 22, no. 1, pp. 1–6, Jan. 2011, doi: 10.1016/j.nec.2010.08.007.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA. Cancer J. Clin.*, vol. 69, no. 1, pp. 7–34, Jan. 2019, doi: 10.3322/caac.21551.

[4] P. W. Sperduto et al., "Estimating survival in patients with lung cancer and brain metastases," *JAMA Oncol.,* vol. 3, no. 6, pp. 827–831, Jun. 2017, doi: 10.1001/jamaoncol.2016.3834.

[5] P. W. Sperduto et al., "Estimating survival in patients with gastrointestinal cancers and brain metastases: An update of the graded prognostic assessment for gastrointestinal cancers (GI-GPA)," *Clin. Transl. Radiat. Oncol.*, vol. 18, pp. 39–45, Sep. 2019, doi: 10.1016/j.ctro.2019.06.007.

[6] D. N. Cagney et al., "Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: A population-based study," *Neuro-Oncol.*, vol. 19, no. 11, pp. 1511–1521, Oct. 2017, doi: 10.1093/neuonc/nox077.

[7] P. W. Sperduto et al., "Estimating survival in melanoma patients with brain metastases: An update of the graded prognostic assessment for melanoma using molecular markers (Melanoma-molGPA)," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 99, no. 4, pp. 812–816, Nov. 2017, doi: 10.1016/j.ijrobp.2017.06.2454.

[8] J. Wong, A. Hird, A. Kirou-Mauro, J. Napolskikh, and E. Chow, "Quality of life in brain metastases radiation trials: A literature review," *Curr. Oncol.*, vol. 15, no. 5, pp. 25–45, Oct. 2008.

[9] M. C. Chamberlain, "Brain metastases: A medical neuro-oncology perspective," *Expert Rev. Neurotherapeutics*, vol. 10, no. 4, pp. 563–573, Apr. 2010, doi: 10.1586/ern.10.30.

[10] P. W. Sperduto et al., "Diagnosis-specific prognostic factors, indexes, and treatment outcomes for patients with newly diagnosed brain metastases: A multi-institutional analysis of 4,259 patients," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 77, no. 3, pp. 655–661, Jul. 2010, doi: 10.1016/j.ijrobp.2009.08.025.

[11] C. Nieder, O. Spanne, M. P. Mehta, A. L. Grosu, and H. Geinitz, "Presentation, patterns of care, and survival in patients with brain metastases," *Cancer*, vol. 117, no. 11, pp. 2505–2512, Jun. 2011, doi: 10.1002/cncr.25707.

[12] T. K. Owonikoko et al., "Current approaches to the treatment of metastatic brain tumours," *Nature Rev. Clin. Oncol.*, vol. 11, no. 4, pp. 203–222, 2014, doi: 10.1038/nrclinonc.2014.25.Current.

[13] P. D. Brown, M. S. Ahluwalia, O. H. Khan, A. L. Asher, J. S. Wefel, and V. Gondi, "Whole-brain radiotherapy for brain metastases: Evolution or revolution?," *J. Clin. Oncol.*, vol. 36, no. 5, pp. 483–491, 2018, doi: 10.1200/JCO.2017.75.9589.

[14] N. U. Lin et al., "Response assessment criteria for brain metastases: Proposal from the RANO group," *Lancet Oncol.*, vol. 16, no. 6, pp. e270–e278, Jun. 2015, doi: 10.1016/S1470-2045(15)70057-4.

[15] P. K. Sneed et al., "Adverse radiation effect after stereotactic radiosurgery for brain metastases: Incidence, time course, and risk factors," *J. Neurosurgery*, vol. 123, no. 2, pp. 373–386, Aug. 2015, doi: 10.3171/2014.10.JNS141610.

[16] M. T. Truong et al., "Results of surgical resection for progression of brain metastases previously treated by gamma knife radiosurgery," *Neurosurgery*, vol. 59, no. 1, pp. 86–97, Jul. 2006, doi: 10.1227/01.NEU.0000219858.80351.38.

[17] H. Kano, D. Kondziolka, J. Lobato-Polo, O. Zorro, J. C. Flickinger, and L. D. Lunsford, "T1/T2 matching to differentiate tumor growth from radiation effects after stereotactic radiosurgery," *Neurosurgery*, vol. 66, no. 3, pp. 486–492, Mar. 2010, doi: 10.1227/01.NEU.0000360391.35749.A5.

[18] E. Karami et al., "Quantitative MRI biomarkers of stereotactic radiotherapy outcome in brain metastasis," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, Art. no. 19830, doi: 10.1038/s41598-019-56185-5.

[19] A. Tiwari, S. Srivastava, and M. Pant, "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019," *Pattern Recognit. Lett.*, vol. 131, pp. 244–260, Mar. 2020, doi: 10.1016/j.patrec.2019.11.020.

[20] Z. Liu et al., "Deep learning based brain tumor segmentation: A survey," *Complex Intell. Syst.*, Jul. 2020, doi: 10.1007/s40747-022-00815-5.

[21] N. Gordillo, E. Montseny, and P. Sobrevilla, "State of the art survey on MRI brain tumor segmentation," *Magn. Reson. Imag.*, vol. 31, no. 8, pp. 1426–1438, 2013, doi: 10.1016/j.mri.2013.05.002.

[22] A. Jalalifar, H. Soliman, M. Ruschin, A. Sahgal, and A. Sadeghi-Naini, "A brain tumor segmentation framework based on outlier detection using one-class support vector machine," in *Proc. IEEE 42nd Annu. Int. Conf. Eng. Med. Biol. Soc.*, Jul. 2020, pp. 1067–1070, doi: 10.1109/EMBC44109.2020.9176263.

[23] E. Karami et al., "An automatic framework for segmentation of brain tumours at Follow-up scans after radiation therapy," in *Proc. IEEE 41st Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2019, pp. 463–466, doi: 10.1109/EMBC.2019.8856858.

[24] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation," *Med. Image Anal.*, vol. 43, pp. 98–111, Jan. 2018, doi: 10.1016/j.media.2017.10.002.

[25] M. Havaei et al., "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[26] A. Işın, C. Direkoğlu, and M. Şah, "Review of MRI-based brain tumor image segmentation using deep learning methods," *Procedia Comput. Sci.*, vol. 102, pp. 317–324, 2016, doi: 10.1016/j.procs.2016.09.407.

[27] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 11045, Cham, Switzerland: Springer, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[28] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9901, S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, Eds. Cham, Switzerland: Springer, 2016, pp. 424–432, doi: 10.1007/978-3-319-46723-8_49.

[29] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Medical Image Understanding and Analysis*, vol. 723, M. Valdés Hernández and V. González-Castro, Eds. Cham, Switzerland: Springer, 2017, pp. 506–517, doi: 10.1007/978-3-319-60964-5_44.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[31] R. Ranjbarzadeh et al., "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, Art. no. 10930, doi: 10.1038/s41598-021-90428-8.

[32] M. Islam, V. S. Vibashan, V. J. M. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain tumor segmentation and survival prediction using 3D attention UNet," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes*, vol. 11992, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2019, pp. 262–272, doi: 10.1007/978-3-030-46640-4_25.

[33] T. Magadza and S. Viriri, "Deep learning for brain tumor segmentation: A survey of State-of-the-Art," *J. Imag.*, vol. 7, no. 2, Jan. 2021, Art. no. 19, doi: 10.3390/jimaging7020019.

[34] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[35] E. Gates, J. G. Pauloski, D. Schellingerhout, and D. Fuentes, "Glioma segmentation and a simple accurate model for overall survival prediction," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. New York, NY, USA: Springer, 2019, pp. 476–484.

[36] L. Pei, L. Vidyaratne, M. M. Rahman, and K. M. Iftekharuddin, "Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 19726, doi: 10.1038/s41598-020-74419-9.

[37] Y. Zhu et al., "Semi-Automatic segmentation software for quantitative clinical brain glioblastoma evaluation," *Academic Radiol.*, vol. 19, no. 8, pp. 977–985, Aug. 2012, doi: 10.1016/j.acra.2012.03.026.

[38] H. Mehrabian, K. L. Desmond, H. Soliman, A. Sahgal, and G. J. Stanisz, "Differentiation between radiation necrosis and tumor progression using chemical exchange saturation transfer," *Clin. Cancer Res.*, vol. 23, no. 14, pp. 3667–3675, Jul. 2017, doi: 10.1158/1078-0432.CCR-16-2265.

[39] A. Sinha and J. Dolz, "Multi-scale guided attention for medical image segmentation," *IEEE J. Biomed. Heal. Inform.*, vol. 25, no. 1, pp. 121–130, Jan. 2021, doi: 10.1109/JBHI.2020.2986926.

[40] F. Isensee, P. F. Jaeger, S. A. 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, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[41] D. Oft et al., "Volumetric regression in brain metastases after stereotactic radiotherapy: Time course, predictors, and significance," *Front. Oncol.*, vol. 10, Jan. 2020, Art. no. 590980, doi: 10.3389/fonc.2020.590980.

[42] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 8, no. 4, Jul. 2018, Art. no. e1249, doi: 10.1002/widm.1249.

[43] B. Vellayappan et al., "Diagnosis and management of radiation necrosis in patients with brain metastases," *Front. Oncol.*, vol. 8, 2018, Art. no. 395, doi: 10.3389/fonc.2018.00395.

[44] J. Xue et al., "Deep learning-based detection and segmentation-assisted management of brain metastases," *Neuro-Oncol.*, vol. 22, no. 4, pp. 505–514, 2020, doi: 10.1093/neuonc/noz234.

[45] J. Cho et al., "Deep learning-based computer-aided detection system for automated treatment response assessment of brain metastases on 3D MRI," *Front. Oncol.*, vol. 11, 2021, Art. no. 739639, doi: 10.3389/fonc.2021.739639.

[46] M. Nishino et al., "Personalized tumor response assessment in the era of molecular medicine: Cancer-specific and therapy-specific response criteria to complement pitfalls of RECIST," *Amer. J. Roentgenol.*, vol. 198, no. 4, pp. 737–745, Apr. 2012, doi: 10.2214/AJR.11.7483.

[47] N. U. Lin et al., "Challenges relating to solid tumour brain metastases in clinical trials, part 1: Patient population, response, and progression. A report from the RANO group," *Lancet Oncol.*, vol. 14, no. 10, pp. e396–e406, Sep. 2013, doi: 10.1016/S1470-2045(13)70311-5.

[48] K. Douri et al., "RANO-BM response criteria verification study in a SRS-treated cohort," *Can. J. Neurol. Sci.*, vol. 49, no. s1, pp. S45–S45, Jun. 2022, doi: 10.1017/cjn.2022.229.

[49] G. Fischedick, U. Haverkamp, and A. Fischedick, "Are three-dimensional volumetric measurements of brain metastasis the future for disease control? A comparative study," *Int. J. Radiat. Oncol., Biol., Phys.*, vol. 99, no. 2, pp. E659–E660, Oct. 2017, doi: 10.1016/j.ijrobp.2017.06.2193.

[50] M. J. Follwell et al., "Volume specific response criteria for brain metastases following salvage stereotactic radiosurgery and associated predictors of response," *Acta Oncol. (Madr).*, vol. 51, no. 5, pp. 629–635, May 2012, doi: 10.3109/0284186X.2012.681066.

[51] A. L. Stockham et al., "Conventional MRI does not reliably distinguish radiation necrosis from tumor recurrence after stereotactic radiosurgery," *J. Neuro-Oncol.*, vol. 109, no. 1, pp. 149–158, Aug. 2012, doi: 10.1007/s11060-012-0881-9.

[52] I. M. Dequesada, R. G. Quisling, A. Yachnis, and W. A. Friedman, "Can standard magnetic resonance imaging reliably distinguish recurrent tumor from radiation necrosis after radiosurgery for brain metastases? A Radiographic-pathological study," *Neurosurgery*, vol. 63, no. 5, pp. 898–904, Nov. 2008, doi: 10.1227/01.NEU.0000333263.31870.31.

[53] J. S. Detsky et al., "Differentiating radiation necrosis from tumor progression in brain metastases treated with stereotactic radiotherapy: Utility of intravoxel incoherent motion perfusion MRI and correlation with histopathology," *J. Neuro-Oncol.*, vol. 134, no. 2, pp. 433–441, Sep. 2017, doi: 10.1007/s11060-017-2545-2.

[54] Y. Yunqi et al., "Quantitative MR perfusion for the differentiation of recurrence and radionecrosis in hypoperfusion and hyperperfusion brain metastases after gamma knife radiosurgery," *Front. Neurol.*, vol. 13, 2022, Art. no. 823731, doi: 10.3389/fneur.2022.823731.

[55] R. Wiggenraad et al., "Pseudo-progression after stereotactic radiotherapy of brain metastases: Lesion analysis using MRI cine-loops," *J. Neuro-Oncol.*, vol. 119, no. 2, pp. 437–443, Sep. 2014, doi: 10.1007/s11060-014-1519-x.