applied sciences

MDPI

*Article*

# Deep Learning-Based Segmentation of Various Brain Lesions for Radiosurgery

**Siangruei Wu [1,†], Yihong Wu [2,†], Haoyun Chang [3], Florence T. Su [4], Hengchun Liao [2], Wanju Tseng [5], Chunchih Liao [6], Feipei Lai [7], Fengming Hsu [8] and Furen Xiao [9,*]**

1 Graduate Institute of Communication Engineering, National Taiwan University, Taipei 106, Taiwan; raywu0@gmail.com
2 School of Medicine, National Taiwan University, Taipei 100, Taiwan; patrickwu97@gmail.com (Y.W.); b05401014@ntu.edu.tw (H.L.)
3 Graduate Institute of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan; hyzhang@eda.ee.ntu.edu.tw
4 College of Arts and Sciences, Santa Clara University, Santa Clara, CA 95053, USA; fsu1@scu.edu
5 Data Intelligence and Application Division, QNAP Systems, Inc., New Taipei 221, Taiwan; rowantseng@qnap.com
6 Department of Neurosurgery, Taipei Hospital, New Taipei 242, Taiwan; ns00360@tph.mohw.gov.tw
7 Department of Computer Science & Information Engineering, National Taiwan University, Taipei 106, Taiwan; flai@csie.ntu.edu.tw
8 Department of Oncology, National Taiwan University Hospital, Taipei 100, Taiwan; hsufengming@ntuh.gov.tw
9 Department of Neurosurgery, National Taiwan University Hospital, Taipei 100, Taiwan
* Correspondence: xiao@ntuh.gov.tw; Tel.: +886-2-23123456
† These authors contributed equally to this work.



**Citation:** Wu, S.; Wu, Y.; Chang, H.; Su, F.T.; Liao, H.; Tseng, W.; Liao, C.; Lai, F.; Hsu, F.; Xiao, F. Deep Learning-Based Segmentation of Various Brain Lesions for Radiosurgery. *Appl. Sci.* **2021**, *11*, 9180. https://doi.org/10.3390/app11199180

Academic Editors: Leonardo Rundo, Carmelo Militello, Andrea Tangherloni and Donato Cascio

Received: 27 July 2021
Accepted: 24 September 2021
Published: 2 October 2021



**Featured Application: This study implemented deep learning methods to the task of segmentation of various brain lesions, facilitating the treatment planning process of neurosurgery and radiation oncology.**

**Abstract:** Semantic segmentation of medical images with deep learning models is rapidly being developed. In this study, we benchmarked state-of-the-art deep learning segmentation algorithms on our clinical stereotactic radiosurgery dataset. The dataset consists of 1688 patients with various brain lesions (pituitary tumors, meningioma, schwannoma, brain metastases, arteriovenous malformation, and trigeminal neuralgia), and we divided the dataset into a training set (1557 patients) and test set (131 patients). This study demonstrates the strengths and weaknesses of deep-learning algorithms in a fairly practical scenario. We compared the model performances concerning their sampling method, model architecture, and the choice of loss functions, identifying suitable settings for their applications and shedding light on the possible improvements. Evidence from this study led us to conclude that deep learning could be promising in assisting the segmentation of brain lesions even if the training dataset was of high heterogeneity in lesion types and sizes.

**Keywords:** deep learning; image segmentation; brain tumors; radiosurgery; magnetic resonance imaging

## 1. Introduction

Stereotactic radiosurgery (SRS) is a treatment modality using ionizing radiation, focusing on precisely selected areas of tissue. It is usually delivered in a single session, but the radiation dose can also be fractionated. Targeting accuracy and anatomic precision are critical to successful SRS, but are historically secondary concerns in other types of radiation therapy [1]. Undoubtedly, as technology evolves, standards in this area will have to change. Nevertheless, when root mean square errors can be reduced to approximately 1 mm, a threshold of surgical possibilities is reached both in the brain and throughout the rest of

the body. As the ACR-ASTRO guidelines suggest, a targeting accuracy is approximately 1 mm [2–4]. Although SRS can be performed in many parts of the body, it is best known to treat intracranial lesions. The common indications for intracranial SRS include many different types of brain tumors, vascular malformations (including arteriovenous malformation, AVM), and functional diseases such as trigeminal neuralgia (TN). Brain metastases, vestibular schwannomas, meningiomas, and pituitary adenomas are common tumor types treated by SRS.

Before the delivery of SRS to the target (e.g., a brain tumor), detailed treatment planning with precise contouring of the target is conducted by a neurosurgeon or a radiation oncologist. The contouring is performed on computed tomography (CT) or magnetic resonance images (MRI). Sometimes, both CT and MRI are used, depending on the devices and diseases. Normal organs or tissues sensitive to radiation are also contoured so that radiation dose and risk of injury can be estimated. These normal organs are called critical organs or organs at risk (OARs). In terms of image analysis, "precise" segmentation of targets and OARs is mandatory for SRS treatment planning. In current clinical practice, the segmentation is performed by professional personnel. The manual contouring process is time-consuming and prone to substantial inter-practitioner variability, even amongst experts, and may lead to large variation in care quality. Several pieces of research suggest computer assistance [5–10]. We expect an AI-based assistive tool could improve tumor detection, shorten mean contouring time, and increase inter-clinician agreement [11].

As convolutional neural networks (CNNs), the dominant deep learning models, are leading the breakthrough in computer vision recently, they also dominate MRI segmentation tasks. Havaei et al. (2017) proposed the idea of using a deep learning model to perform brain tumor segmentation tasks on MRI images [12]. They pointed out that both local and global representations are essential to produce better results, and this intuition was later realized in various ways. Kamnitsas et al. (2017) later perfected this idea and achieved state-of-the-art performance with a two-path model [13]. On the other hand, U-Net was first proposed for the cell tracking task [14], but then became widely used in many other segmentation tasks [15,16]. In MICCAI BraTS 2017 competition [17], most participants used U-Net variants, as the winner [18] simply ensembled three kinds of the most common deep learning models, namely FCN (fully convolutional network) [19], V-Net [20], and DeepMedic [13]. Other than deep learning, some studies on brain cancer segmentation took advantage of fuzzy c-means clustering [21–23], cellular automata [24], random walker [8], and so on [5,7,25,26]. However, they are not deep learning by not possessing over two hidden layers and will not be further discussed.

However, few studies apply deep learning methods to the actual SRS datasets. Unlike the BraTS competitions, real applicable models may need to handle much more diversity rather than a single type of disease. Liu et al. (2017) proposed a modification of DeepMedic that outperformed its parent method in segmentation during SRS treatment planning by adding a subpath, with a dice score reaching 0.67 in a cohort of 240 patients [27]. Lu et al. (2019) ensembled two neural networks, namely 3D U-Net and DeepMedic, which were trained with different hyper-parameters so that one neural network focused on small metastases with high sensitivity while the other one addressed overall tumor segmentation with high specificity, yielding a good performance on segmentation within 305 patients, with a median dice score of 0.74 [28]. Fong et al. (2019) trained the convoluted neural network with multiplanar slices, reducing false-positive predictions and yielding a dice score of 0.77 on the 248-patient dataset while maintaining competent 80% isodose coverage [29]. Lu et al. (2021) implemented deep learning in the treatment planning process, decreasing the plethora of time consumed during the planning process as well as enhancing the prediction overlap with ground truth significantly especially in the subgroup of non-experts. However, the cohort size was rather small [11]. Heterogeneity could have been considered a major problem for machine learning decades ago; however, it should be considered a real-world situation. A heterogeneous dataset could help the generalizability and transferability of trained models [11,23,25,30–32]. However, in the

previously-mentioned studies, small sample sizes were important contributors to the lack of confidence to infer the generalization of deep-learning models in clinical practices with heterogeneous lesion types. For the technology to achieve satisfactory performance, we explored the behavior of deep-learning models in a realistic scenario. Therefore, we collected a relatively large dataset with 1688 patients and analyzed the performance of models with various types of settings and architectures. More specifically, we benchmarked the performance of different segmentation models previously proposed for other tasks and also compared the effectiveness of various sampling methods and the choice of loss functions. We used the BRATS dataset to evaluate whether our implementations of deep learning models were correct and comparable to their original implementations.

## 2. Materials and Methods

### *2.1. Dataset*

#### 2.1.1. NTUH (National Taiwan University Hospital) Dataset

The data were extracted from a medical center in northern Taiwan. The SRS device used was CyberKnife (Accuray, Sunnyvale, CA, USA) and commenced operation in January 2008. In the decade until December 2017, there were 2578 treatment courses completed in 2411 patients. Among these, 2036 treatment courses of 1921 patients were intracranial.

We only selected patients undergoing first SRS with contrast-enhanced T1-weighted (T1+C) MRI images available. Finally, there were 1688 patients included in our dataset. Their data were randomly divided into training and test sets (Table 1). However, because treatment targets for patients with trigeminal neuralgia are neither tumor nor vascular malformation, their data were all assigned to the training set.

**Table 1.** Clinical diagnoses of 1688 patients in the final dataset.

| Types of Brain Lesions | Train | Test |
|---|---|---|
| Metastases | 504 | 53 |
| Meningioma | 314 | 29 |
| Schwannoma | 305 | 20 |
| Pituitary tumor | 147 | 8 |
| Arteriovenous malformation | 80 | 6 |
| Trigeminal neuralgia | 38 | 0 |
| Other tumors | 169 | 15 |
| Total | 1557 | 131 |

For each patient, the target was extracted from the treatment planning system together with axial T1+C MRI. Most of the time, the targets were contoured by a neurosurgeon and then reviewed by a radiation oncologist. Occasionally only one radiation oncologist contoured the targets without review by another physician. An image volume could contain more than one target, particularly in patients with brain metastases. NTUH volumes were registered to CT volumes before contouring, with tumor contours stored in CT coordinates. We had to make an inverse transformation to put the tumor labels in MRI coordinates. In other words, volumes were resampled on a common voxel grid instead of directly cropped on grids of different voxel numbers. The images were retrieved from DICOM format and saved in NIfTI-1 data format, where names, birth dates, and geographic data were removed. After registration and de-identification, these image/label pairs were used for the training and evaluation of deep neural networks. The images were presented in native axial slices, with 1–2 mm slice thicknesses. The number of slices varied from 30 to 233 since the slices did not necessarily cover the full cranial regions, instead, they could only include the region of interest. The in-plane resolution was usually 512 $\times$ 512, and the smallest resolution was 197 $\times$ 197. The field of view in the x–y plane was usually 300 mm, ranging from 200 to 350 mm. The pixel size was mostly 0.5859 $\times$ 0.5859 $mm^2$ and was 1.1719 $\times$ 1.1719 $mm^2$ for some images with lower resolution.

There were a total of 2568 distinct targets in these 1688 image sets. The target volumes ranged from 20 to 72,646 $mm^3$, with a median of 1236 $mm^3$ and a mean of 3696 $\pm$ 6637 $mm^3$. In 1013 image sets, there was only one target. The number of targets may reach up to 34 in a single image set.

#### 2.1.2. BraTS Dataset

The BraTS 2015 dataset is a standard benchmark dataset for MRI segmentation tasks. It includes 220 multi-modal scans of patients with high-grade glioma (HGG) and 54 with low-grade glioma (LGG). T1-weighted, contrast-enhanced T1+C, T2-weighted, and FLAIR images are available. The data had a common dimension of 240 $\times$ 240 $\times$ 155 with 1 $mm^3$ resolution. The annotation contains five classes: 0 for background, 1 for necrotic core (NC), 2 for edema (OE), 3 for non-enhancing core, and 4 for enhancing core. The evaluation follows the rules of the competition by merging the predictions into three sets: whole tumor (classes 1,2,3,4), core (classes 1,3,4), and enhancing core (class 4). The train to test ratio was 10:1 in each of the experiments we conducted.

### *2.2. Preprocessing*

The raw data of the NTUH dataset contains images of different resolutions and fields of view (FOVs). We first used the skull stripping function of Brain-Suite [33] to locate the brain, then utilized the information possessed by the brain masks for centration and cropping of the MRI to make sure the images contain fewer extracranial areas. Brain-Suite was used only to locate the brain center for better cropping. Everything from the scalp to skull remained on the cropped images for reasons that some of our lesions may locate extra-axially. The final input images size was 200 $\times$ 200 $\times$ 200 $mm^3$. Finally, we normalized them by the z-scores.

Images in the BraTS dataset were already registered, cropped, and normalized with bias field corrections. We only normalized the data by the z-scores for every pulse sequence (T1, T2, T1+C, FLAIR).

### *2.3. Data Augmentation*

To perform a fair comparison of the model architectures, we established the following standard data augmentation in the training phase. For 2D models, we performed data augmentation with translation, rotation, shear, zoom, brightness, and elastic distortion [15]. For 3D models, since data augmentation did not yield higher performances of segmentation in the preliminary experiment we performed, we did not perform any type of data augmentation.

### *2.4. Deep Learning Models*

The design of the models we employed could be found online [34]. We will discuss the rationale and the architecture below.

#### 2.4.1. DeconvNet

DeconvNet is an architecture adopted from VGG16, a 16-layered CNN by the Visual Geometry Group, and is rather simple to implement [35]. The objective of this design is to overcome the limitations of FCN, which cannot detect objects that are bigger or smaller than a specific size. In this case, the object may be fragmented or mislabeled. Furthermore, FCN only uses one convolution transpose layer to construct its output, so the output loses much detail. As a consequence, DeconvNet uses several layers of transpose-convolution and up-pooling.

The model can be divided into two parts: the encoder and the decoder, which are formed by convolution and deconvolution operations, respectively. It is worth noting that we replaced the max-pooling and up-sampling operations by setting the stride of Conv and Deconv to 2 in our implementation. This is inspired by the recent proposition of generative adversarial networks.

#### 2.4.2. DeepMedic

DeepMedic is another kind of 3D CNN [13]. It is special for taking two inputs, high resolution, and low resolution. This design seeks to balance fine structures and high-level information. High-resolution inputs for DeepMedic are patches from our preprocessed data. Low-resolution inputs are downsampled using 3D average pooling from each corresponding high-resolution patch. Both inputs go through a series of convolution layers with skip connection, and then it constructs the output by fusing the features of both pathways. This is a state-of-the-art model, and we expect the model to perform well on segmentation of brain lesions based on previous benchmarks [27,36–38].

#### 2.4.3. PSPNet

Pyramid scene parsing network, or PSPNet, is a state-of-the-art model in scene parsing tasks [39]. We included it because it is also suitable for our segmentation task. The PSPNet utilizes the high-level representations extracted by a pretrained network, and a novel design of the pyramid pooling module serves as a backend to predict the segmentations. The pyramid pooling modules pool the extracted feature maps to obtain features of different scales. The pretrained model is typically a ResNet trained on an ImageNet dataset [39]. However, on an MRI dataset, the features are not transferable due to the large consistency and the absence of common pretrained models to process MRI images. In our implementation, we randomly initialized the ResNet backend and also removed the deep supervision loss.

#### 2.4.4. U-Net

U-Net tries to improve the fine structure of segmentations and increase the amount of context used [14]. Traditionally, when a certain amount of pooling is required, if one is intending to train with large patches, it unfortunately degrades the performance such as in FCN and DeconvNet. Hence, the U-Net model utilizes skip connections to forward the unpooled features, thus the model can utilize the information of various scales. In our implementation, we abandoned the max-pooling and up-sampling operations for the same reason as in DeconvNet.

#### 2.4.5. V-Net

V-Net is the adaption of U-Net for 3-dimensional data to capture the relationships in consecutive slices, which were omitted in the 2D models, addressing contiguity problems and yielding better results in the segmentation of various 3D images [20]. It replaces the convolution and pooling operations with 3D versions.

### *2.5. Sampling Method*

Batch samplers were defined in the source code (see Supplementary Material) [40]. We had four batch sampler designs for our models to experiment with the most efficient and accurate settings. We tried each kind of sampling method for each model in our preliminary experiment, but only those sampling methods that did not cause much overfitting, excessive memory consumption, or lower performance than other sampling methods were benchmarked.

#### 2.5.1. Two Dimensional

For two-dimensional models, we split the MRI data slice by slice and performed predictions separately. This may result in noises along the sliced axis due to the loss of spatial contiguity information.

#### 2.5.2. Three Dimensional

For three-dimensional models, the basic strategy is to feed the whole brain image data directly. In the preliminary experiment, three-dimensional patch resulted in high memory consumption when we employed the DeepMedic. Thus, we did not perform a

three-dimensional patch for the DeepMedic. Furthermore, while we experimented with this setting on the BraTS2015 dataset, we found it caused overfitting, and we suspect that this is because many of the voxels are irrelevant and redundant for the prediction. Thus, we added two more three-dimensional sampling methods described below.

#### 2.5.3. Uniform Patch

To reduce the redundant voxels and save memory usage, we sampled small patches within the brain regions. While inferencing, we simply reassembled the patch predictions together. The patch size used was 152 × 128 × 128. Worth noting, for the DeepMedic, uniform patch resulted in generally lower performance in segmentation, sensitivity, and precision compared with the center patch. Hence this method was not employed when training the DeepMedic.

#### 2.5.4. Center Patch

It has been suggested that patches containing foreground regions are crucial to the training [35]. We thus deployed this sampling strategy, which guarantees at least one foreground voxel in the patch. The patch size was default to 64 × 64 × 64. However, we could set the patch size to 96 × 96 × 96 or the same as the size of the uniform patch when we applied the sampling method. When we experimented with the settings in the preliminary test, center patch sizes lower than 190 × 190 would result in tremendously low segmentation performance for V-Net in the NTUH dataset, so this sampling method was not used in the formal benchmark analysis.

### *2.6. Hyperparameters*

We used fixed optimizer settings across all experiments. The optimizer chosen was Adam. The learning rate was initially $1 \times 10^{-4}$, with step decay of factor 0.1 at 50 and 70 epochs. Patch-wise methods were inferenced with patches cropped without overlap and excessive boundaries were filled with zero paddings. Samples were generated on the fly in a patch-wise method.

### *2.7. Loss Functions*

Class imbalance is a major problem in most tumor segmentation problems, and it is even more severe in our task compared to the BraTS glioma dataset because of small target volumes. The imbalance would most likely lead the model to a trivial solution, which predicts all voxels as background. There are several ways to deal with this problem by modifying the loss function.

#### 2.7.1. Weighted Cross-Entropy

Re-weighting the sparse class is the most common solution to the class imbalance problem. In this study, we set the class weights inversely proportional to the ratio of the class. In particular:

$$C = -\sum_{c=1}^{M} \frac{g_{oc} log(p_{oc})}{r_c} \quad (1)$$

where $M$ is the number of classes and $r_c$ is the ratio of class $c$ in the whole volume/dataset (as an implementation choice); $g_{oc}$ is the ground-truth label of a voxel; and $p_{oc}$ is the predicted label probability of a voxel of class $c$.

#### 2.7.2. Soft-Dice

Milletari et al. (2016) suggest using the differentiable version of the dice score, namely soft-dice, directly as the objective due to its resistibility to class imbalance [20]. It is fairly natural to use this loss function because the dice score is the most common evaluation metric in related tasks. There are two implementations of the soft-dice loss function.

Regarding the cardinality of sets, one can perform summation directly or with squaring. In particular:

$$D1 = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \text{ or } D2 = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i + \sum_i^N g_i} \quad (2)$$

where $p_i$ is the predicted label probability and $g_i$ is the ground-truth label. We found the two versions producing almost identical performances. In this study, we refer to the second version as the soft-dice loss function.

### *2.8. Evaluation Metrics*

#### 2.8.1. Dice Score (Hard Dice)

Dice score is the standard metric for evaluating segmentation results. It is defined as

$$\mathrm{D} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (3)$$

where $X$ and $Y$ are the sets of predicted and labeled lesion voxels.

The previously mentioned soft-dice loss is a modified differentiable version of the dice score. We, therefore, refer to the dice score metric as hard-dice to distinguish the two.

#### 2.8.2. Precision and Sensitivity

Precision and sensitivity (also known as recall) are standard metrics of binary classification, which is a more general scheme for segmentation. Precision quantifies the volume ratio of correctly predicted lesion voxels (TP) to all predicted lesion voxels (TP + FP). It is defined as

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (4)$$

Sensitivity quantifies the volume ratio of correctly predicted lesion voxels (TP) to all labeled lesion voxels (TP + FN). It is defined as

$$\text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (5)$$

### *2.9. Experiments*

In all our experiments, the training and testing were conducted under a 10:1 train-test ratio.

#### 2.9.1. Performances of Models on Segmentation of Brain Lesions in NTUH Dataset

We first experimented with the performances of different models on the segmentation of brain tumors after training with various brain lesions. Different batch samplers and loss functions were employed based on the models used. The hard dice, precision, and sensitivity were the outcomes we were interested in. The higher dice score, precision, and sensitivity were deemed as better segmentation performance.

#### 2.9.2. Performances of Models on Segmentation of Brain Tumors in BraTS Dataset

To compare the variables contributing to the performances of the models trained with the NTUH dataset, we experimented with the segmentation of the brain tumors in the BraTS dataset. We trained our models with either 4-channel or only T1+C inputs to compare the performances of the models trained with the same imaging modality in the NTUH dataset. During training, the labels encompassed either five classes or only tumor cores. The segmentation performances were measured using the hard dice. The evaluation was based on predictions of the tumor cores.

## 3. Results

Three cases from the NTUH dataset showing representative results of different models were shown in Tables 2–4. The overall dice scores of these networks on the NTUH dataset ranged from 0.33 (DeepMedic) to 0.51 (V-Net). Table 5 shows the detailed performance of each network tested with the NTUH dataset.

**Table 2.** Predictions with low dice scores.

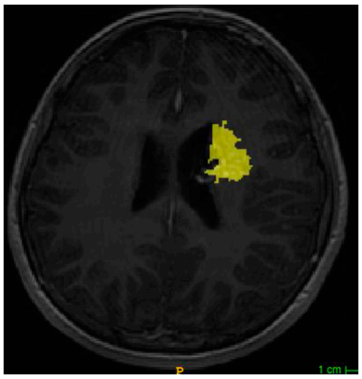
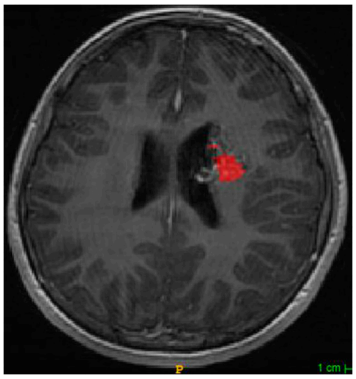
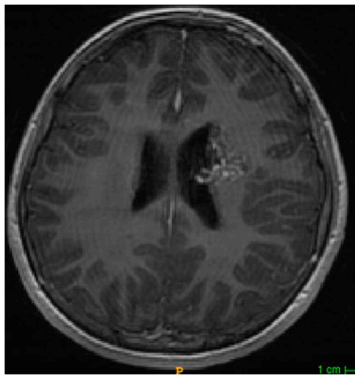
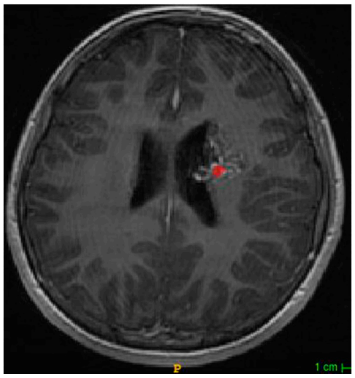
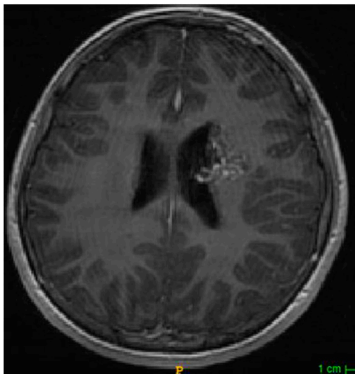
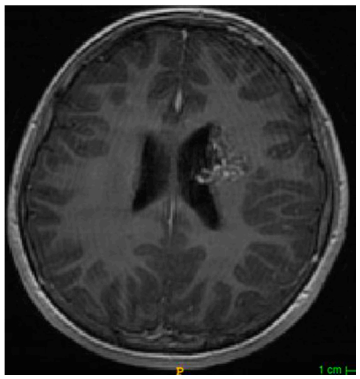
| Ground truth | DeconvNet | DeepMedic |
|---|---|---|
|  |  |  |
| **PSPNet** | **U-Net** | **V-Net** |
|  |  |  |

**Table 3.** Predictions with average dice scores.

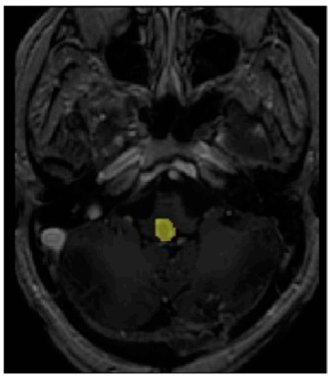
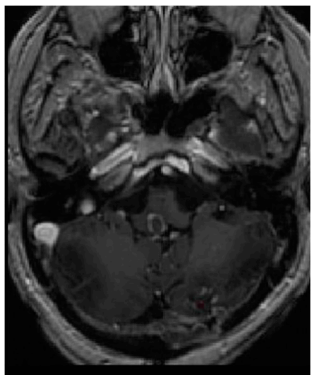
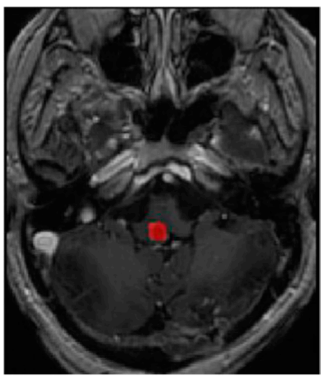
| Ground truth | DeconvNet | DeepMedic |
|---|---|---|
|  |  |  |

**Table 3.** *Cont.*

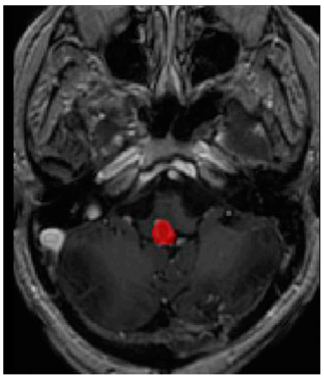
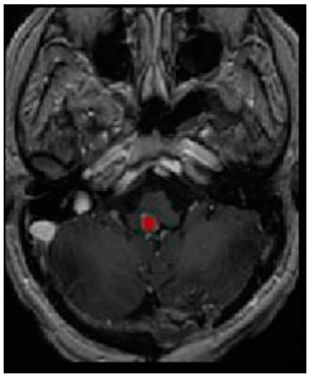
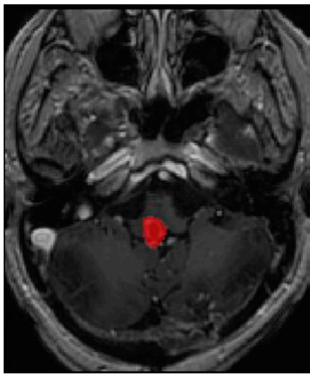
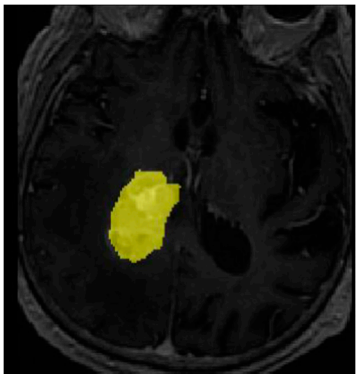
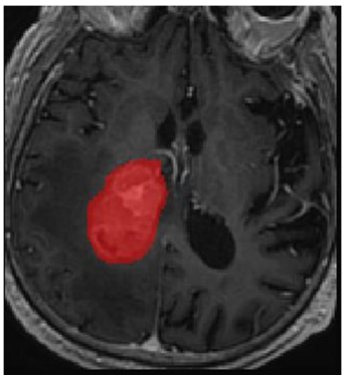
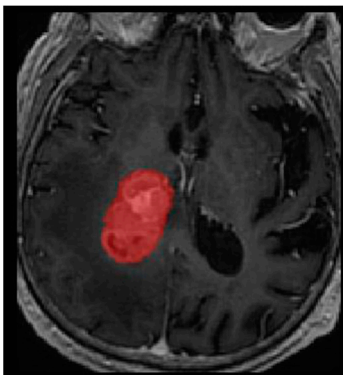
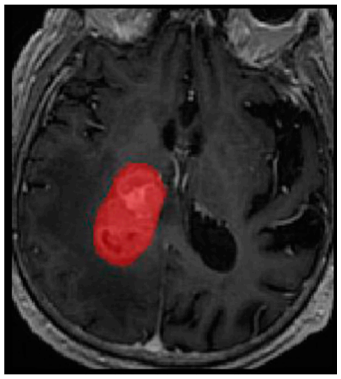
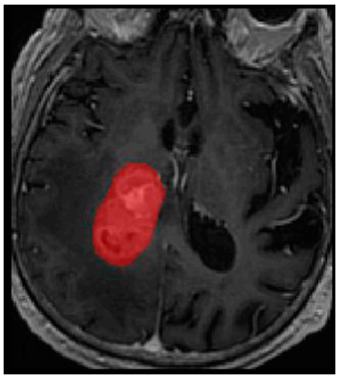
| PSPNet | U-Net | V-Net |
| --- | --- | --- |
|  |  |  |

**Table 4.** Predictions with high dice scores.

| Ground truth | DeconvNet | DeepMedic |
| --- | --- | --- |
|  |  |  |
| **PSPNet** | **U-Net** | **V-Net** |
|  |  |  |

On the NTUH datasets, the performance was also affected by the types of lesions. As shown in Figure 1, we obtained better results for brain metastases, meningiomas, and schwannomas, while all models performed poorly on pituitary tumors, AVMs, and other tumor types. Detailed tables are attached as Appendix A, Appendix B, Appendix C.

**Table 5.** Performance of different models on the NTUH dataset.

| Model | Numbers of Parameters | Batch Samplers | Loss Function | Precision | Sensitivity | Hard Dice |
|---|---|---|---|---|---|---|
| DeconvNet | 12,544,324 | two_dim | Cross-entropy minus log(soft-dice) | 0.46 | 0.48 | 0.43 |
| U-Net | 34,524,034 | two_dim | Cross-entropy minus log(soft-dice) | **0.48** | 0.48 | 0.43 |
| PSPNet | 28,280,773 | two_dim | Cross-entropy minus log(soft-dice) | 0.47 | 0.48 | 0.43 |
| V-Net | 8,232,274 | uniform_patch3d | Cross-entropy minus log(soft-dice) | 0.39 | 0.54 | 0.41 |
| V-Net | 8,232,274 | three_dim | Cross-entropy | 0.2 | 0.56 | 0.25 |
| V-Net | 8,232,274 | three_dim | Cross-entropy minus log(soft-dice) | **0.48** | 0.51 | 0.46 |
| V-Net dropout 0.1 | 8,232,274 | three_dim | Cross-entropy minus log(soft-dice) | 0.47 | **0.66** | **0.51** |
| DeepMedic | 1,301,478 | center_patch3d | Cross-entropy minus log(soft-dice) | 0.36 | 0.43 | 0.35 |
| DeepMedic | 1,301,478 | center_patch3d | Cross-entropy | 0.37 | 0.43 | 0.33 |

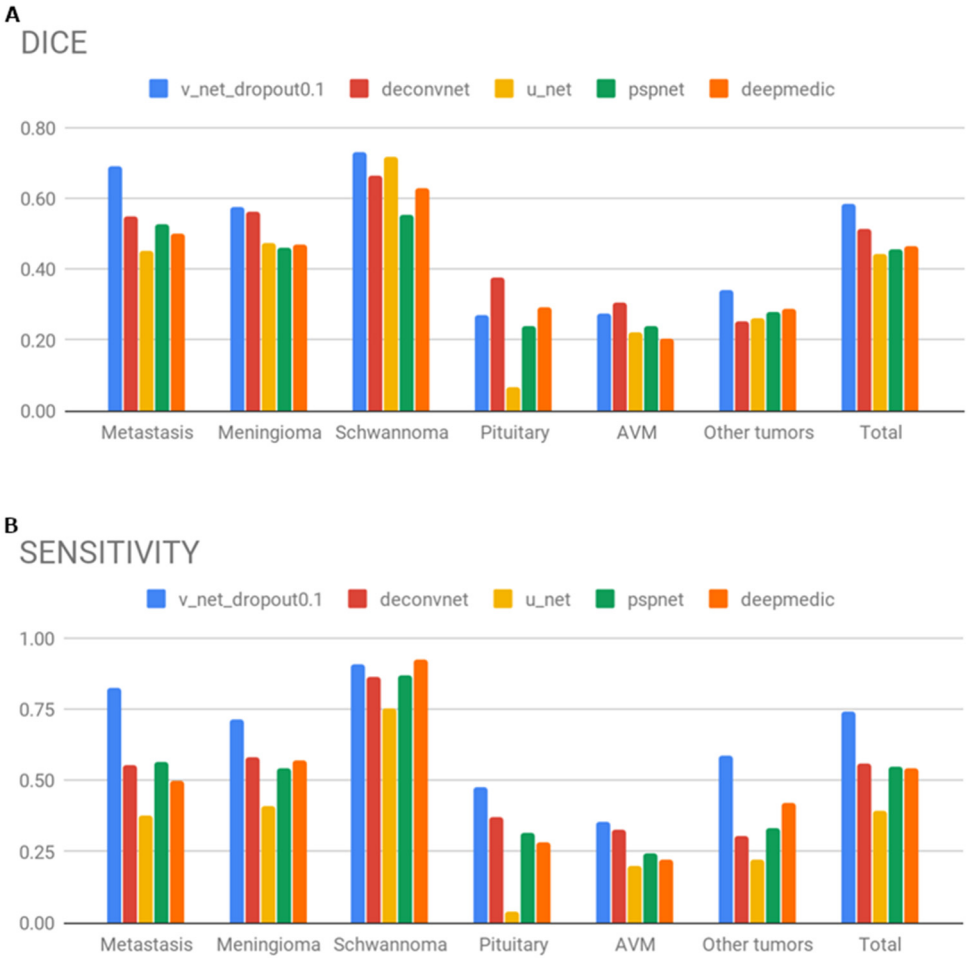



**Figure 1.** *Cont.*

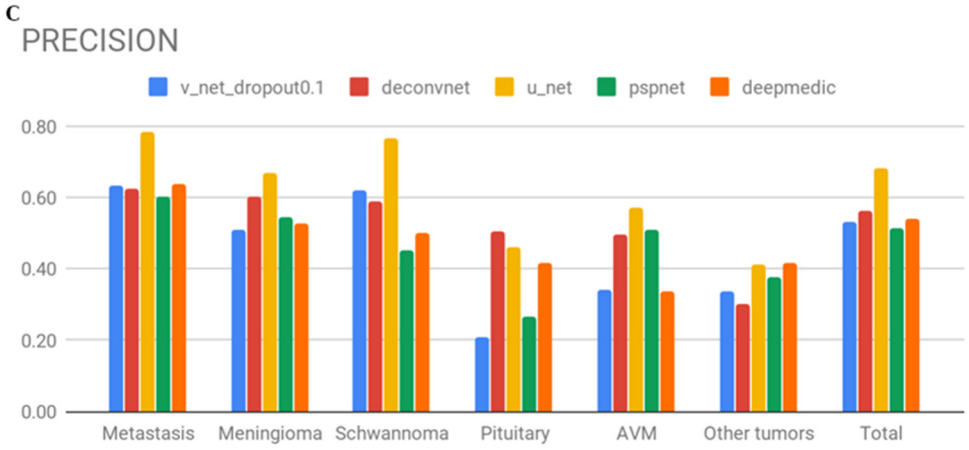



**Figure 1.** Bar chart results of (**A**) dice score; (**B**) sensitivity; and (**C**) precision of deep learning-based segmentation versus ground truth on different lesion types in the NTUH dataset.

As shown in Figure 2, lesions with smaller target volumes introduce lower average dice performance for each deep-learning model. V-Net, the best-performing model in the current study, obtained a fairly satisfactory dice score when lesion size exceeded the median size of all targets.

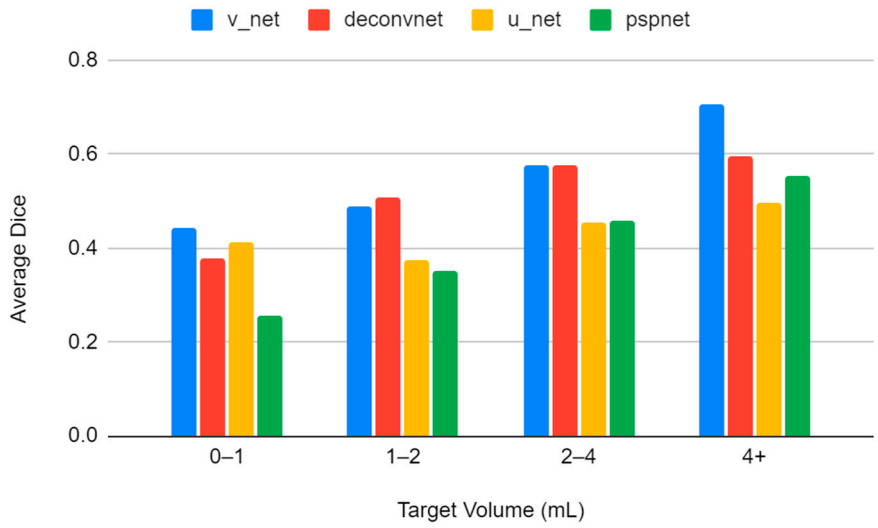



**Figure 2.** The performance of each model on lesions within different volume ranges.

To compare the performances of different models trained with one-channel input on the segmentation of brain lesions, we performed another experiment in which models were trained with just T1+C input with tumor core labels of the BraTS dataset. The evaluation was based on prediction of the tumor cores. As shown in Figure 3, V-Net had the highest Dice score when trained with 4-channel input with 5-class labels. Interestingly, all models performed better in this circumstance than when trained with only T1+C images. Of note, V-Net and PSPNet could not yield comparable results when trained with only T1+C images, implying that they are more sensitive to the change from multimodal to single modality inputs. While the models trained with one-channel inputs yielded lower performances in segmentation, they still performed better than their counterparts trained with the NTUH dataset.


**Figure 3.** The performance of each model trained with four-channel or one-channel inputs.

Because of the nature of PSPNet and DeepMedic, they took a significantly longer time for inference, as shown in Table 6. V-Net had the least number of parameters and the shortest inference time. We also found that adding dropouts in V-Net further improved its performance, which we have noted in the table with 0.1 being the dropout rate.

**Table 6.** Inference time on our hardware and parameters of different networks.

| | V-Net Dropout 0.1 | DeconvNet | U-Net | PSPNet | DeepMedic |
|---|---|---|---|---|---|
| Inference time (minutes:seconds) | 02:51 | 04:00 | 04:01 | 14:43 | 17:73 |
| Number of parameters | 8.23M | 12.5M | 34.5M | 28.3M | 13M |

## 4. Discussion

### *4.1. Segmentation Performance: NTUH vs. BraTS Dataset*

The performance on our radiosurgery dataset was inferior to that on BraTS. Many factors might lead to such a result. First of all, the tumor volumes in the NTUH dataset are typically smaller than those in BraTS 2015. On average, the tumor occupied 1.23% of the whole image volume in the BraTS dataset, but only 0.145% in ours. It should also be noted that a significant portion of our dataset contained multiple targets, which is much less likely for glioma patients (BraTS). The lesions in NTUH dataset are thus more difficult to detect.

Moreover, there is significant heterogeneity in our dataset. To evaluate whether our model could achieve similar segmentation performance under a more realistic scenario, we used the dataset containing cranial lesions of various pathology, which is different from the BraTS dataset with only glioma cases. In a strict sense, we also have some images of non-neoplastic diseases such as AVM. Additionally, some of the tumors are extra-axial (outside the brain parenchyma) and may even extend extracranially, so we cannot perform skull stripping like BraTS. Due to the heterogeneity of tumor types and sites, we may need a much larger dataset to reach similar performance.

Our results indicated that better performance was correlated with more training samples (as in metastases, meningioma, and schwannoma, Figure 2) and larger lesion dimensions (Figure 3). We also report the effect of input channels (of BraTS) in this revision.

Another reason is that we only used one image set (T1+C) to predict instead of four sequences used in the BraTS dataset. Less information might lead to deteriorated performance.

It is also worth mentioning that our dataset is quite imbalanced disease-wise. From the performance of the models we trained, we could observe that this imbalance resulted in serious bias issues for minority patients. We found it quite difficult to train a model by the traditional soft-dice loss or cross-entropy loss. Using the weighted cross-entropy loss gave us a 0.25 dice score, while our modification of subtracting a log-soft-dice term improved the dice score to 0.40. Such difference may result from tumor size since tumors in our dataset were of fewer voxels on average. In addition to the data variety, the weighted cross-entropy function could be very unstable and thus harmful to the optimization. Empirically, we found that the model will most likely fail in 10 epochs and predict nothing but the background for all inputs. By adding another term with the dice score, the new loss function provides better guidance to the model, and we could empirically observe the significant improvements.

We added images of trigeminal neuralgia in the training set as negative samples, in which there was no real space-occupying lesion. We did not expect the machine to learn how to identify trigeminal neuralgia. Instead, it can be considered that images of trigeminal neuralgia are examples of the heterogeneity of real clinical datasets. This artificial impurity was meant to mimic the systematic bias that could occur in a larger and unpurified dataset to infer the availability of deep learning models.

Although the targets in our dataset were defined and contoured by experienced clinicians, it should be noted that they were the targets we wanted to treat. Therefore, in very few cases, not every lesion detected by human experts was labeled. For example, a patient with brain metastases may also have a small meningioma, which may be stable and will not be labeled and treated by radiosurgery. If an algorithm detects that meningioma is this rare, decreased precision and dice score can be expected. However, from the clinical experience of our expert neurosurgeons and radiation oncologists, the rate of intentionally ignored meningiomas and pituitary adenomas was estimated around 1%. This estimation was in parallel with the reported prevalence of intracranial incidentaloma. On the other hand, the estimated rate for ignored brain metastases was much higher (5%), because our clinical experts might decide not to treat small lesions (less than 5–10 mm or visible only on one axial slice) in patients with multiple brain metastases [41,42]. As a result, this should not impede the training due to its rarity, and most meningiomas were labeled.

### *4.2. Performance on Different Types of Tumor*

We can see that these models performed better for brain metastases, meningiomas, and schwannomas, where there were more than 300 cases each. They performed best for schwannomas, probably because most of these are vestibular ones, whose locations are always around internal auditory meatus.

On the other hand, these models performed poorly for pituitary tumors, AVMs, and other tumor types. Besides the relatively small number of cases for training, pituitary tumors and AVMs are not always readily visible for humans using only the T1+C series. For example, dynamic contrast-enhanced MRI may be required to visualize pituitary tumors. AVMs are sometimes not visible even using time-of-flight (TOF) MRI, so computed tomography angiography and/or digital subtraction angiography may be required for target contouring.

### *4.3. Comparison between Deep Learning Models*

With respect to the input format, there are two classes of model architectures. The 2D model predicts tumors in just one slice and completely discards the information along the *z*-axis, while the 3D model utilizes the full information on the MRI volume. This results in a trade-off between features and overfitting. When receiving more features, it is more likely to overfit the unrelated noise, especially with such a small dataset. Patch sizes in previous works range from $16 \times 16 \times 16$ to $64 \times 64 \times 64$ mm$^3$ [18,43–55], of which Kamnitsas et al. outperformed the others. Thus, in the current proposed work, we restrained the receptive field and predicted on inputs patches with the size of $64 \times 64 \times 64$ mm$^3$. We examined

this trade-off in our benchmark experiment on the BraTS dataset. Surprisingly, when experimenting with V-Net on our dataset, small patch-wise prediction became detrimental, but receiving the full brain volume guaranteed the best performance.

Overall, the 3D models seem to be more appealing. The 3D models present the full potential of convolution networks, reducing the number of parameters and becoming far more efficient due to their convolution nature. Specifically, V-Net has approximately 1/30 of the parameters compared to U-Net, shortest inference time, and the best performance on dice metric. The only shortcoming of 3D models is the requirement of GPU RAM due to the large input. In our experiments, we solve this by using a smaller batch-size. Furthermore, replacing batch normalization with dropout is quite effective in preventing overfitting because of the small batch size.

We compared the performances of the models trained with one-channel inputs of the NTUH and BraTS datasets. When the models were trained with one-channel inputs, the segmentation performances were slightly better than when they were trained with four-channel inputs. It could be inferred that the models perform better on a dataset with less heterogeneity in lesion types as well as lesion sizes.

### *4.4. Comparison to Previous Studies Addressing Deep Learning-Based Segmentation in SRS Treatment Planning*

Efforts to identify the targets and the OARs prior to SRS treatment are crucial for dosimetry planning to protect the organs other than the lesions themselves. Several studies have benefited from deep learning methods on the classification and nomenclature standardization of the OARs [56,57]. The above-mentioned studies could advance computer-assisted radiation therapy.

To evaluate the benchmark performed in this study on the segmentation of brain lesions, previous studies addressing the segmentation of brain tumors in the treatment planning process during SRS will be reviewed. Of all types of brain lesions, asymptomatic or unresectable metastases warrant SRS without maximal surgical resection. As SRS serves as the first-line treatment for oligometastatic lesions, which denotes metastases of lesser than five lesions, contouring the lesions is of important clinical significance. The models previously used included modified DeepMedic [11,27], an ensemble of DeepMedic and 3D U-Net [28], and CNN [29].

Tumor volume tremendously affects the performance of segmentation; higher variety in tumor sizes and smaller lesions usually imply adversity in segmentation. Smaller lesions, while not affecting dice scores much, are not easily detected in methods with lower sensitivity. Liu et al. (2017) [27] proposed a modification of DeepMedic and managed to reach a dice score of 0.67. In their study, the number of brain metastases per case varied from 1 to 93 (5.679 $\pm$ 8.917), and the mean tumor size was 672 $\pm$ 1994 $mm^3$. Lu et al. (2019) [28] ensembled two neural networks, namely 3D U-Net and DeepMedic, yielding a good performance in segmentation with a median dice score of 0.74. The median size of the tumors in their dataset was 980 $mm^3$, while the smallest tumor was 3 $mm^3$. Fong et al. (2019) [29] trained the convoluted neural network with multiplanar slices, yielding a dice score of 0.77. Lu et al. (2021) [11] implemented an ensemble of 3D U-Net and DeepMedic and enhanced the prediction significantly, especially for non-experts. In their dataset, the median volume of the lesion was 890 $mm^3$. In our dataset, the lesions possessed a median size of 656 $mm^3$ and a mean of 2833 $\pm$ 6389 $mm^3$, while the smallest lesion was 13.05 $mm^3$. Generally speaking, with the highest dice score of 0.51, sensitivity of 0.66, and precision of 0.48, the lesions in our dataset had a higher size variety and smaller median size. The inconsistency in the lesion characteristics could cause difficulties for the deep learning models to extract features and hinder the prediction.

Ensemble models introduced higher segmentation performance than a single model in the previous studies [11,27–29]. Although in our study, V-Net with a dropout rate of 0.1 outperformed other methods in segmentation of brain lesions in the NTUH dataset, we did not perform a benchmark on the ensemble methods. It remains undetermined whether ensemble models yield better performance as well as which models ensembled could enhance segmentation the most.

As the difference in the imaging sequences used in the training process is a determinant of segmentation performance, the sequences used in previous works are discussed. Liu et al. (2017) used contrast-enhanced T1-weighted images [27] while Lu et al. (2019) used CT and T1-weighted MRI scans with contrast as the input [28]. Multiplanar slices of MPRAGE (magnetization-prepared rapid acquisition with gradient echo) images were taken as input in Fong et al. (2019) [29]. Lu et al. (2021) used contrast-enhanced CT and T1-weighted MR scans [11]. Out of the three studies, methods with MPRAGE as the input sequence yielded the highest dice score compared to the ground truth. Brain tumors on FLAIR, which is often used to contour the clinical target volume (CTV), mostly appeared as confluent hyperintense signals, introducing higher sensitivity and lower precision. On the other hand, brain tumors were mostly discrete on MPRAGE, an MRI modality taking advantage of gradient echo [58]. Despite the fact that higher precision could be achieved with MPRAGE, it is currently of lower significance in contouring before SRS. Of note, studies have shown that simultaneous use of different imaging modalities promised a better performance in segmentation compared to single modality use [38]. In our study, only contrast-enhanced T1-weighted MR images were used, and this could be a determinant of lower segmentation performance.

The required dataset size to yield high performances could not be confirmed, as we collected the data available to train the models and only draw conclusions from the current dataset. It is probably true that a larger dataset may generate better or different results, but such a dataset was not available to us.

### *4.5. Limitation of This Study*

Compared to previous works investigating samples that underwent SRS, a relatively large dataset was implemented in the current study. However, the results suggest that the numbers of pituitary tumors, AVMs, and other tumors are probably insufficient for good results. Since the numbers of above lesions in a single institute may be insufficient, federated learning can be a potentially practical approach for better results.

Contrast-enhanced T1-weighted MR imaging was the only modality used as input in our study. Some tumors such as low-grade glioma or pituitary tumors are non-enhancing, introducing great difficulty in the detection and segmentation of these types of lesions. Simultaneous use of multiple imaging modalities could be the solution to this. Reviewing previous works, the sensitivity for detection of smaller brain lesions (<3 mm) with 3D U-Net, whether trained with black-blood or gradient echo modalities, decreased significantly compared to larger brain lesions ($\geq$10 mm, 0.981, 3–10 mm 0.829, <3 mm 0.235) [59]. The same trend could be observed in studies performed with 2-stage MetNet ($\geq$6 mm 0.99, 3–6 mm 0.87, $\leq$3 mm 0.25) [60] or GoogLeNet [61]. The 2-stage MetNet [60] and BMDS net [62] could achieve satisfactory segmentation prediction on tumors larger than 6 mm, with dice scores of 0.87 and 0.83, respectively. In our dataset counterpart, the diameters of 10.5% lesions were smaller than 6 mm, 45% lesions smaller than 10 mm, and 95.7% smaller than 3 cm. The small lesion sizes in our NTUH dataset contributed to the dice score lower than 0.6 predicted with V-Net.

The way dice score is derived could mask the effect of contouring small lesions. In our work, dice score was calculated per voxel, which favored larger lesions compared to dice score derived per lesion. Clinically, SRS is indicated and is of significant importance for patients with smaller brain lesions, whereas for patients who are surgical candidates with larger lesions, standard care remains surgery with adjuvant stereotactic radiation therapy or whole-brain radiation therapy. As for patients with diffuse lesions, whole-

brain radiation therapy is the standard treatment due to the lack of level 1 evidence to support the use of SRS in the patient population [63] (p. 865). Contouring deflection on the gross tumor volume (GTV) of such small lesions could introduce a huge impact on later target contouring, compromising organs at risk (OAR). Take brain metastases, for example, current guidelines for contouring for SRS generally indicate a 1.5 cm expansion from GTV to generate CTV. In our dataset, the smallest volume of brain lesion being 20 $mm^3$ implies a 3.4 mm diameter, and the volume difference of CTV with GTV is about 3000 $mm^3$. This expansion in target volume significantly differs if a small lesion was not correctly contoured. As a consequence, a dice score per lesion provides benefit in some circumstances.

Evidence derived from trials concerning treatment response to SRS based on either deep-learning segmentation or manual segmentation is still an unmet need. Several studies implemented multiple modalities (PET/MRI) in order to train machine learning models for tumor segmentation, which suggested that biological target volume (BTV) could be promising in helping CTV definition during SRS treatment and their ability to indicate dose escalation on biologically active targets [64,65]. Despite the effort in assisting CTV definition by taking advantage of the training set of multi-modalities, whether the addition in modalities to either of the learning methods improves clinical treatment response is yet undetermined.

## 5. Conclusions

We benchmarked five commonly used deep learning segmentation models on our SRS dataset. We confirmed that these approaches also work on a heterogeneous dataset, but with decreased performance. We discovered that the V-Net architecture worked best for this specific task. With the top dice scores, the smallest size of the model, and the shortest inference time, V-Net may be a good choice to improve upon. We also found that when training on the dataset with such heterogeneity and class imbalance, using weighted cross-entropy loss with log-soft-dice term significantly improved the performance.

**Supplementary Materials:** GitHub repository, https://github.com/raywu0123/Brain-Tumor-Segmentation, (accessed on 25 September 2021).

**Author Contributions:** Conceptualization, S.W., Y.W., H.C., and F.X.; Methodology, S.W., H.C., and Y.W.; Software, S.W. and H.C.; Validation, S.W., Y.W., and F.X.; Formal analysis, Y.W., C.L., and F.X.; Investigation, S.W., H.C., Y.W., and F.X.; Resources, F.X.; Data curation, S.W., H.C., Y.W., and F.X.; Writing—original draft preparation, Y.W., C.L., and F.X.; Writing—review and editing, Y.W., F.T.S., H.L., W.T., C.L., F.L., F.H., and F.X.; Visualization, C.L., Y.W., and F.X.; Supervision, F.X.; Project administration, F.X.; Funding acquisition, F.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Ministry of Science and Technology, Taiwan, ROC, grant numbers 107-2634-F-002-015, 110-2634-F-002-032, 110-2314-B-002-161.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by Research Ethics Committee of National Taiwan University Hospital (201708071RINC, 6 October 2017).

**Informed Consent Statement:** Informed consent was waived by the committee.

**Data Availability Statement:** The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Conflicts of Interest:** The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Appendix A

| DICE | v_net_dropout0.1 | Deconvnet | u_net | Pspnet | Deepmedic |
|---|---|---|---|---|---|
| Metastasis | 0.69 | 0.55 | 0.45 | 0.52 | 0.5 |
| Meningioma | 0.57 | 0.56 | 0.48 | 0.46 | 0.47 |
| Schwannoma | 0.73 | 0.66 | 0.72 | 0.55 | 0.63 |
| Pituitary | 0.27 | 0.38 | 0.07 | 0.24 | 0.29 |
| AVM | 0.27 | 0.31 | 0.22 | 0.24 | 0.2 |
| Other tumors | 0.34 | 0.25 | 0.26 | 0.28 | 0.29 |
| Total | 0.59 | 0.52 | 0.44 | 0.46 | 0.46 |

## Appendix B

| SENSITIVITY | v_net_dropout0.1 | Deconvnet | u_net | Pspnet | Deepmedic |
|---|---|---|---|---|---|
| Metastasis | 0.82 | 0.55 | 0.37 | 0.56 | 0.5 |
| Meningioma | 0.71 | 0.58 | 0.41 | 0.54 | 0.57 |
| Schwannoma | 0.91 | 0.86 | 0.75 | 0.87 | 0.92 |
| Pituitary | 0.48 | 0.37 | 0.04 | 0.31 | 0.29 |
| AVM | 0.36 | 0.33 | 0.2 | 0.24 | 0.22 |
| Other tumors | 0.58 | 0.3 | 0.22 | 0.33 | 0.42 |
| Total | 0.74 | 0.56 | 0.39 | 0.55 | 0.54 |

## Appendix C

| PRECISION | v_net_dropout0.1 | Deconvnet | u_net | Pspnet | Deepmedic |
|---|---|---|---|---|---|
| Metastasis | 0.63 | 0.62 | 0.78 | 0.6 | 0.64 |
| Meningioma | 0.51 | 0.6 | 0.67 | 0.54 | 0.53 |
| Schwannoma | 0.62 | 0.59 | 0.76 | 0.45 | 0.5 |
| Pituitary | 0.21 | 0.5 | 0.46 | 0.27 | 0.42 |
| AVM | 0.34 | 0.49 | 0.57 | 0.51 | 0.34 |
| Other tumors | 0.34 | 0.3 | 0.41 | 0.38 | 0.42 |
| Total | 0.53 | 0.56 | 0.68 | 0.52 | 0.54 |

## References


1. Adler, J.R., Jr.; Colombo, F.; Heilbrun, M.P.; Winston, K. Toward an expanded view of radiosurgery. *Neurosurgery* **2004**, *55*, 1374–1376. [CrossRef] [PubMed]
2. Chao, S.T.; Dad, L.K.; Dawson, L.A.; Desai, N.B.; Pacella, M.; Rengan, R.; Xiao, Y.; Yenice, K.M.; Rosenthal, S.A.; Hartford, A. ACR–ASTRO Practice Parameter for the Performance of Stereotactic Body Radiation Therapy. *Am. J. Clin. Oncol.* **2020**, *43*, 545–552. [CrossRef] [PubMed]
3. Schell, M.C.; Bova, F.J.; Larson, D.A.; Leavitt, D.D.; Latz, W.R.; Podgorsak, E.B.; Wu, A. *Stereotactic Radiosurgery*; AAPM Report NO. 54; American Association of Physicists in Medicine: Alexandria, VA, USA, 1995.
4. Seung, S.K.; Larson, D.A.; Galvin, J.M.; Mehta, M.P.; Potters, L.; Schultz, C.J.; Yajnik, S.V.; Hartford, A.C.; Rosenthal, S.A. American College of Radiology (ACR) and American Society for Radiation Oncology (ASTRO) Practice Guideline for the Performance of Stereotactic Radiosurgery (SRS). *Am. J. Clin. Oncol.* **2013**, *36*, 310–315. [CrossRef] [PubMed]
5. Shin, H.-C. Hybrid clustering and logistic regression for multi-modal brain tumor segmentation. In Proceedings of the MICCAI-BRATS 2012, Nice, France, 1 October 2012.
6. Bauer, S.; Fejes, T.; Slotboom, J.; Wiest, R.; Nolte, L.-P.; Reyes, M. Segmentation of brain tumor images based on integrated hierarchical classification and regularization. In Proceedings of the MICCAI-BRATS 2012, Nice, France, 1 October 2012.
7. Zhao, L.; Wu, W.; Corso, J.J. Brain tumor segmentation based on GMM and active contour method with a model-aware edge map. In Proceedings of the MICCAI-BRATS 2012, Nice, France, 1 October 2012; pp. 19–23.
8. Xiao, Y.; Hu, J. Hierarchical random walker for multimodal brain tumor segmentation. In Proceedings of the MICCAI-BRATS 2012, Nice, France, 1 October 2012.
9. Subbanna, N.; Arbel, T. Probabilistic gabor and markov random fields segmentation of brain tumours in mri volumes. In Proceedings of the MICCAI-BRATS 2012, Nice, France, 1 October 2012; pp. 28–31.
10. Zikic, D.; Glocker, B.; Konukoglu, E.; Shotton, J.; Criminisi, A.; Ye, D.; Demiralp, C.; Thomas, O.M.; Das, T.; Jena, R. Context-sensitive classification forests for segmentation of brain tumor tissues. In Proceedings of the MICCAI-BRATS 2012, Nice, France, 1 October 2012; pp. 22–30.

11. Lu, S.-L.; Xiao, F.-R.; Cheng, J.C.-H.; Yang, W.-C.; Cheng, Y.-H.; Chang, Y.-C.; Lin, J.-Y.; Liang, C.-H.; Lu, J.-T.; Chen, Y.-F.; et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. *Neuro-Oncology* **2021**, *23*, 1560–1568. [CrossRef]
12. Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Pal, C.; Jodoin, P.-M.; Larochelle, H. Brain tumor segmentation with Deep Neural Networks. *Med. Image Anal.* **2017**, *35*, 18–31. [CrossRef]
13. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.J.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Rueckert, D.; Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. [CrossRef] [PubMed]
14. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
15. Dong, H.; Yang, G.; Liu, F.; Mo, Y.; Guo, Y. Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2017; pp. 506–517.
16. Livne, M.; Rieger, J.; Aydin, O.U.; Taha, A.A.; Akay, E.M.; Kossen, T.; Sobesky, J.; Kelleher, J.D.; Hildebrand, K.; Frey, D.; et al. A U-Net Deep Learning Framework for High Performance Vessel Segmentation in Patients With Cerebrovascular Disease. *Front. Neurosci.* **2019**, *13*, 97. [CrossRef]
17. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117. [CrossRef]
18. Kamnitsas, K.; Bai, W.; Ferrante, E.; McDonagh, S.; Sinclair, M.; Pawlowski, N.; Rajchl, M.; Lee, M.; Kainz, B.; Rueckert, D.; et al. Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation. In *Machine Learning and Knowledge Discovery in Databases*; Springer: Cham, Switzerland, 2017; pp. 450–462.
19. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional models for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 640–651. [CrossRef]
20. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]
21. Militello, C.; Rundo, L.; Vitabile, S.; Russo, G.; Pisciotta, P.; Marletta, F.; Ippolito, M.; D'Arrigo, C.; Midiri, M.; Gilardi, M.C. Gamma Knife treatment planning: MR brain tumor segmentation and volume measurement based on unsupervised Fuzzy C-Means clustering. *Int. J. Imaging Syst. Technol.* **2015**, *25*, 213–225. [CrossRef]
22. Hamamci, A.; Kucuk, N.; Karaman, K.; Engin, K.; Unal, G. Tumor-Cut: Segmentation of Brain Tumors on Contrast Enhanced MR Images for Radiosurgery Applications. *IEEE Trans. Med. Imaging* **2011**, *31*, 790–804. [CrossRef]
23. Hu, M.; Zhong, Y.; Xie, S.; Lv, H.; Lv, Z. Fuzzy System Based Medical Image Processing for Brain Disease Prediction. *Front. Neurosci.* **2021**, *15*, 965. [CrossRef]
24. Rundo, L.; Militello, C.; Russo, G.; Vitabile, S.; Gilardi, M.C.; Mauri, G. GTV cut for neuro-radiosurgery treatment planning: An MRI brain cancer seeded image segmentation method based on a cellular automata model. *Nat. Comput.* **2018**, *17*, 521–536. [CrossRef]
25. Wu, X.; Bi, L.; Fulham, M.; Feng, D.D.; Zhou, L.; Kim, J. Unsupervised brain tumor segmentation using a symmetric-driven adversarial network. *Neurocomputing* **2021**, *455*, 242–254. [CrossRef]
26. Rundo, L.; Militello, C.; Tangherloni, A.; Russo, G.; Vitabile, S.; Gilardi, M.C.; Mauri, G. NeXt for neuro-radiosurgery: A fully automatic approach for necrosis extraction in brain tumor MRI using an unsupervised machine learning technique. *Int. J. Imaging Syst. Technol.* **2018**, *28*, 21–37. [CrossRef]
27. Liu, Y.; Stojadinovic, S.; Hrycushko, B.; Wardak, Z.; Lau, S.; Lu, W.; Yan, Y.; Jiang, S.B.; Zhen, X.; Timmerman, R.; et al. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. *PLoS ONE* **2017**, *12*, e0185844. [CrossRef]
28. Lu, S.; Hu, S.; Weng, W.; Chen, Y.; Lu, J.; Xiao, F.; Hsu, F. Automated Detection and Segmentation of Brain Metastases in Stereotactic Radiosurgery Using Three-Dimensional Deep Neural Networks. *Int. J. Radiat. Oncol.* **2019**, *105*, S69–S70. [CrossRef]
29. Fong, A.; Swift, C.; Wong, J.; McVicar, N.; Giambattista, J.; Kolbeck, C.; Nichol, A. Automatic Deep Learning-based Segmentation of Brain Metastasis on MPRAGE MR Images for Stereotactic Radiotherapy Planning. *Int. J. Radiat. Oncol.* **2019**, *105*, E134. [CrossRef]
30. Sachdeva, J.; Kumar, V.; Gupta, I.; Khandelwal, N.; Ahuja, C.K. Segmentation, Feature Extraction, and Multiclass Brain Tumor Classification. *J. Digit. Imaging* **2013**, *26*, 1141–1150. [CrossRef]
31. Gros, C.; Lemay, A.; Cohen-Adad, J. SoftSeg: Advantages of soft versus binary training for image segmentation. *Med. Image Anal.* **2021**, *71*, 102038. [CrossRef]
32. Wong, J.; Huang, V.; Wells, D.; Giambattista, J.; Giambattista, J.; Kolbeck, C.; Otto, K.; Saibishkumar, E.P.; Alexander, A. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: A workflow study at two cancer centers. *Radiat. Oncol.* **2021**, *16*, 101. [CrossRef] [PubMed]
33. Shattuck, D.W.; Leahy, R.M. BrainSuite: An automated cortical surface identification tool. *Med. Image Anal.* **2002**, *6*, 129–142. [CrossRef]

34. Wu, S.-R.; Wu, P.Y.; Chang, H.Y. Brain-Tumor-Segmentation/Models at Master • raywu0123/Brain-Tumor-Segmentation. Available online: https://github.com/raywu0123/Brain-Tumor-Segmentation/tree/master/models (accessed on 25 September 2021).
35. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
36. Pennig, L.; Shahzad, R.; Caldeira, L.; Lennartz, S.; Thiele, F.; Goertz, L.; Zopfs, D.; Meißner, A.-K.; Fürtjes, G.; Perkuhn, M.; et al. Automated Detection and Segmentation of Brain Metastases in Malignant Melanoma: Evaluation of a Dedicated Deep Learning Model. *Am. J. Neuroradiol.* **2021**, *42*, 655–662. [CrossRef] [PubMed]
37. Jünger, S.T.; Hoyer, U.C.I.; Schaufler, D.; Laukamp, K.R.; Goertz, L.; Thiele, F.; Grunz, J.; Schlamann, M.; Perkuhn, M.; Kabbasch, C.; et al. Fully Automated MR Detection and Segmentation of Brain Metastases in Non-small Cell Lung Cancer Using Deep Learning. *J. Magn. Reson. Imaging* **2021**. [CrossRef]
38. Charron, O.; Lallement, A.; Jarnet, D.; Noblet, V.; Clavier, J.-B.; Meyer, P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Comput. Biol. Med.* **2018**, *95*, 43–54. [CrossRef] [PubMed]
39. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
40. Wu, S.R.; Wu, P.Y.; Chang, H.Y. Brain-Tumor-Segmentation/Models/Batch_Samplers at Master • raywu0123/Brain-Tumor-Segmentation. Available online: https://github.com/raywu0123/Brain-Tumor-Segmentation/tree/master/models/batch_samplers (accessed on 25 September 2021).
41. Neugut, A.I.; Sackstein, P.; Hillyer, G.C.; Jacobson, J.S.; Bruce, J.; Lassman, A.B.; Stieg, P.A. Magnetic Resonance Imaging-Based Screening for Asymptomatic Brain Tumors: A Review. *Oncologist* **2019**, *24*, 375–384. [CrossRef]
42. Nakasu, S.; Notsu, A.; Nakasu, Y. Prevalence of incidental meningiomas and gliomas on MRI: A meta-analysis and meta-regression analysis. *Acta Neurochir.* **2021**, 1–15. [CrossRef]
43. Andermatt, S.; Pezold, S.; Cattin, P. Multi-dimensional gated recurrent units for brain tumor segmentation. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017; pp. 15–19.
44. Amorim, P.H.A.; Chagas, V.S.; Escudero, G.G.; Oliveira, D.D.C.; Pereira, S.M.; Santos, H.M.; Scussel, A.A. 3D u-nets for brain tumor segmentation in miccai 2017 brats challenge. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017.
45. Castillo, L.S.; Daza, L.A.; Rivera, L.C.; Arbeláez, P. Volumetric multimodality neural network for brain tumor segmentation. In Proceedings of the 13th International Conference on Medical Information Processing and Analysis, San Andres Island, Colombia, 5–7 October 2017; p. 105720E.
46. Feng, X.; Meyer, C. Patch-based 3d u-net for brain tumor segmentation. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017.
47. Zhou, C.; Ding, C.; Lu, Z.; Zhang, T. Brain tumor segmentation with cascaded convolutional neural networks. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017; pp. 328–333.
48. Isensee, F.; Kickingereder, P.; Wick, W.; Bendszus, M.; Maier-Hein, K.H. Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. In *International MICCAI Brainlesion Workshop*; Springer: Cham, Switzerland, 2018; pp. 287–297. [CrossRef]
49. Li, Y.; Shen, L. MvNet: Multi-view deep learning framework for multimodal brain tumor segmentation. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017.
50. Pourreza, R.; Zhuge, Y.; Ning, H.; Miller, R. Brain Tumor Segmentation in MRI Scans Using Deeply-Supervised Neural Networks. In *International MICCAI Brainlesion Workshop*; Springer: Cham, Switzerland, 2018; pp. 320–331.
51. Zhou, F.; Li, T.; Li, H.; Yu, K.; Wang, Y.; Zhu, H. TP-CNN: A two-phase convolution neural network based model to do automatic brain tumor segmentation by using BRATS 2017 data. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017; pp. 334–341.
52. Zhu, J.; Wang, D.; Teng, Z.; Lio, P. A multi-pathway 3d dilated convolutional neural network for brain tumor segmentation. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017; pp. 342–347.
53. Hu, Y.; Xia, Y. Automated brain tumor segmentation using a 3D deep detection-classification model. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017.
54. Chen, S.; Ding, C.; Zhou, C. Brain tumor segmentation with label distribution learning and multi-level feature representation. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017.
55. Beers, A.; Chang, K.; Brown, J.; Sartor, E.; Mammen, C.; Gerstner, E.; Rosen, B.; Kalpathy-Cramer, J. Sequential 3d u-nets for brain tumor segmentation. In Proceedings of the International MICCAI BraTS Challenge 2017, Quebec City, QC, Canada, 14 September 2017; pp. 20–23.
56. Yang, Q.; Chao, H.; Nguyen, D.; Jiang, S. A Novel Deep Learning Framework for Standardizing the Label of OARs in CT. In *Workshop on Artificial Intelligence in Radiation Therapy*; Springer: Cham, Switzerland, 2019; pp. 52–60.
57. Yang, Q.; Chao, H.; Nguyen, D.; Jiang, S. Mining Domain Knowledge: Improved Framework Towards Automatically Standardizing Anatomical Structure Nomenclature in Radiotherapy. *IEEE Access* **2020**, *8*, 105286–105300. [CrossRef]
58. Brant-Zawadzki, M.; Gillan, G.D.; Nitz, W.R. MP RAGE: A three-dimensional, T1-weighted, gradient-echo sequence—Initial experience in the brain. *Radiology* **1992**, *182*, 769–775. [CrossRef]

59. Park, Y.W.; Jun, Y.; Lee, Y.; Han, K.; An, C.; Ahn, S.S.; Hwang, D.; Lee, S.-K. Robust performance of deep learning for automatic detection and segmentation of brain metastases using three-dimensional black-blood and three-dimensional gradient echo imaging. *Eur. Radiol.* **2021**, *31*, 6686–6695. [CrossRef]
60. Zhou, Z.; Sanders, J.W.; Johnson, J.M.; Gule-Monroe, M.; Chen, M.; Briere, T.M.; Wang, Y.; Son, J.B.; Pagel, M.D.; Ma, J.; et al. MetNet: Computer-aided segmentation of brain metastases in post-contrast T1-weighted magnetic resonance imaging. *Radiother. Oncol.* **2020**, *153*, 189–196. [CrossRef] [PubMed]
61. Grøvik, E.; Yi, D.; Iv, M.; Tong, E.; Rubin, D.; Zaharchuk, G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J. Magn. Reson. Imaging* **2020**, *51*, 175–182. [CrossRef] [PubMed]
62. Xue, J.; Wang, B.; Ming, Y.; Liu, X.; Jiang, Z.; Wang, C.; Liu, X.; Chen, L.; Qu, J.; Xu, S.; et al. Deep learning–based detection and segmentation-assisted management of brain metastases. *Neuro-Oncology* **2020**, *22*, 505–514. [CrossRef]
63. Hansen, E.K.; Roach, M., III. *Handbook of Evidence-Based Radiation Oncology*; Springer: Berlin/Heidelberg, Germany, 2018.
64. Wang, X.; Cui, H.; Gong, G.; Fu, Z.; Zhou, J.; Gu, J.; Yin, Y.; Feng, D. Computational delineation and quantitative heterogeneity analysis of lung tumor on 18F-FDG PET for radiation dose-escalation. *Sci. Rep.* **2018**, *8*, 10649. [CrossRef]
65. Rundo, L.; Stefano, A.; Militello, C.; Russo, G.; Sabini, M.G.; D'Arrigo, C.; Marletta, F.; Ippolito, M.; Mauri, G.; Vitabile, S.; et al. A fully automatic approach for multimodal PET and MR image segmentation in gamma knife treatment planning. *Comput. Methods Programs Biomed.* **2017**, *144*, 77–96. [CrossRef] [PubMed]