advances
in radiation oncology

Scientific Article

Automated Brain Metastases Segmentation With a Deep Dive Into False-positive Detection

Hamidreza Ziyaee, PhD,[a] Carlos E. Cardenas, PhD,[b] D. Nana Yeboa, MD,[c] Jing Li, MD, PhD,[c] Sherise D. Ferguson, MD,[d] Jason Johnson, MD,[e] Zijian Zhou, PhD,[f] Jeremiah Sanders, PhD,[f] Raymond Mumme, BS,[a] Laurence Court, PhD,[a] Tina Briere, PhD,[a] and Jinzhong Yang, PhD[a],*

[a]*Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas;* [b]*Department of Radiation Oncology, The University of Alabama at Birmingham, Birmingham, Alabama;* [c]*Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas;* [d]*Department of Neurosurgery, The University of Texas MD Anderson Cancer Center, Houston, Texas;* [e]*Department of Neuroradiology, The University of Texas MD Anderson Cancer Center, Houston, Texas;* [f]*Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas*

Abstract

Purpose: The clinical management of brain metastases after stereotactic radiosurgery (SRS) is difficult, because a physician must review follow-up magnetic resonance imaging (MRI) scans to determine treatment outcome, which is often labor intensive. The purpose of this study was to develop an automated framework to contour brain metastases in MRI to help treatment planning for SRS and understand its limitations.

Methods and Materials: Two self-adaptive nnU-Net models trained on postcontrast 3-dimensional T1-weighted MRI scans from patients who underwent SRS were analyzed. Performance was evaluated by computing positive predictive value (PPV), sensitivity, and Dice similarity coefficient (DSC). The training and testing sets included 3482 metastases on 845 patient MRI scans and 930 metastases on 206 patient MRI scans, respectively.

Results: In the per-patient analysis, PPV was 90.1% ± 17.7%, sensitivity 88.4% ± 18.0%, DSC 82.2% ± 9.5%, and false positive (FP) 0.4 ± 1.0. For large metastases (≥6 mm), the per-patient PPV was 95.6% ± 17.5%, sensitivity 94.5% ± 18.1%, DSC 86.8% ± 7.5%, and FP 0.1 ± 0.4. The quality of autosegmented true-positive (TP) contours was also assessed by 2 physicians using a 5-point scale for clinical acceptability. Seventy-five percent of contours were assigned scores of 4 or 5, which shows that contours could be used as-is in clinical application, and the remaining 25% were assigned a score of 3, which means they needed minor editing only. Notably, a deep dive into FPs indicated that 9% were TP metastases not identified on the original radiology review, but identified on subsequent follow-up imaging (early detection). Fifty-four percent were real metastases (TP) that were identified but purposefully not contoured for target treatment, mainly because the patient underwent whole-brain radiation therapy before/after SRS treatment.

Conclusions: These findings show that our tool can help radiologists and radiation oncologists detect and contour tumors from MRI, make precise decisions about suspicious lesions, and potentially find lesions at early stages.

Introduction

Approximately 20% to 40% of patients with cancer develop brain metastases.[1] Stereotactic radiosurgery (SRS) is a functional and routinely used treatment for brain metastases,[2,3] using multiple cobalt sources (Gamma Knife) or a linear accelerator (Linac) and delivering a single high dose of radiation to targets.[1,4] Accurately detecting and contouring metastases for treatment planning are important steps to successfully treat brain metastases with SRS.[5,6]

Traditionally, brain metastases are detected manually by a radiologist, and contoured by a radiation oncologist using radiation therapy planning software.[7,8] However, reports exist of missed small and even large brain metastases in clinical practice during the planning of Gamma Knife radiosurgery because of varied human factors.[9,10] Automating metastasis detection could be used as a tool to support clinicians in image evaluation.[11,12] This automation can efficiently reduce human errors, achieve aims with minimal human operation, and augment system performance.[13] Deep learning models have recently shown great potential in medical image analysis, specifically in segmentation, detection, and classification.[11,14]

Several approaches have been introduced for brain metastasis segmentation in magnetic resonance imaging (MRI) using deep learning.[4] These methods use several deep convolutional neural networks (CNNs), including different layers for convolution, pooling, and classification.[15] Previous studies showed the detection of brain metastases with a high sensitivity at >80%.[8] However, those studies usually reported a large number of false-positive (FP) findings and low positive predictive values (PPVs) at the same time, and none reported the early detection of metastases.[10,16-18]

Developing an automated framework to detect metastases from brain MRI and provide accurate contouring could greatly facilitate treatment planning for SRS, as well as treatment outcome prognosis in patient follow up. In addition, having a model to detect metastases in earlier stages before a radiologist is able to see them would have a huge effect in clinical applications. The purpose of this study was to develop and validate an automated framework to contour brain metastases from MRI scans using deep neural networks with improved FP detection and early detection of metastases.

Methods and Materials

We developed a framework using nnU-Net[19] for brain metastasis detection in postcontrast T1-weighted MRI scans (Fig. 1). In the framework, we trained 2 segmentation models based on a self-adaptive nnU-Net, brain metastases detector (BM-Net) to segment the metastases, and whole-brain network (WB-Net) to segment brain volume as our region of interest for metastases detection.

Patient data

This study was approved by the local institutional review board (protocol number PA16-0379). The data set in this study was comprised of 1 planning MRI scan per patient from 1051 patients who underwent Gamma Knife
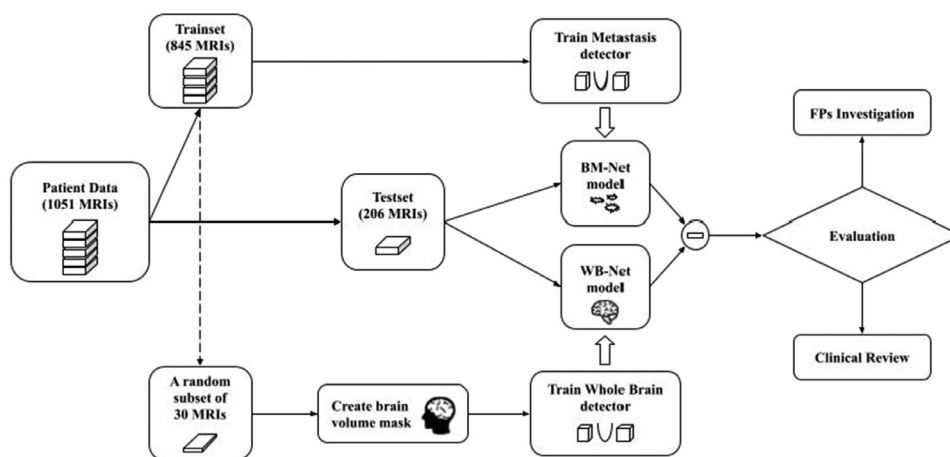

Figure 1   Workflow of building and evaluating our model. Abbreviations: BM-Net = metastases detector; FPs = false positives; MRI = magnetic resonance imaging; WB-Net = whole-brain network.

treatment at MD Anderson Cancer Center between August 2009 and August 2019. We applied no limitation on the size or number of metastases when collecting the data. Patients were scanned with an axial 3-dimensional T1-weighted MRI sequence with and without contrast echo. The acquisition parameters of the MRI system were as follows: GE medical system models Signa PET/MR, Discovery MR750, Optima MR 450, and Siemens model Aera; magnetic field strength 1.5T/3T; modality MRI; repetition time 5.4 to 10 ms; flip angle 12° to 20°; matrix size 256×256, and voxel size $0.94 \times 0.94 \times 1$ mm^3.

The data set included patients ages 16 to 82 years with a total of 4442 brain metastases, the largest of which was 63 mm, and the number of metastases per patient was up to 27. The mean number of metastases was 4 ± 3.5 per patient, and the mean metastasis size was 10 ± 7.4 mm. These metastatic tumors were contoured by experienced oncologists. Only those tumors identified for Gamma Knife treatment were contoured, and these contours were used as ground-truth for segmentation comparison in this study. Of note, some tumors might not be contoured because of varied medical reasons, which will be detailed in the section of FP detection.

The patient data set was randomly split into 845 for the training set and 206 for the testing set (80:20). Table 1 shows the demographics and primary cancer types of patients in the training and testing sets. The average number of metastases per patient and size of metastases (Table 1) show that tumors were distributed with almost the same ratio in the training and testing sets. Moreover, an inclusive subset of 30 patients was randomly selected from the training set, and their brain volumes were manually contoured from the MRI scans to train the WB-Net.

Segmentation

As shown in Figure 1, the training set was used to train the BM-Net to detect metastases, and the subset of 30 MRI scans was used to train the WB-Net to contour brain volume. The self-adaptive nnU-Net framework was used to train both networks. The nnU-Net uses the U-Net architecture with an automated pipeline, including data augmentation, preprocessing, and postprocessing, which require no changes in the architecture of the network and take care of hyper-parameter tuning.[19] Specifically, the data augmentation and preprocessing steps in the nnU-Net includes Gaussian blur, Gaussian noise, scaling, rotation, simulation of low resolution, brightness, gamma correction, contrast, and mirroring. No additional preprocessing was performed other than that.

The U-Net[20] network consists of a series of convolutional layers that reduce the dimension of input image (256×256), and stack together several 3×3 convolutional

maps with padding, followed by a rectified linear unit. The U-Net uses 2×2 max-pooling layers with stride 2, and contracts until the bottommost is reached. This is the encoder part where the model grows the predetermined number of channels. From the bottommost, the model starts on the upconvolution path, which is called decoder or expanding part. At every stage of the upconvolution, the model concatenates the results of the corresponding step from the downsampling path, which adds robustness to the network.

The training configuration of the nnU-Net was as follows: An initial learning rate of 0.001 and stochastic gradient descent with Nesterov momentum ($\mu = 0.99$) was used for the optimizer. Networks were trained for 1000 epochs, with each epoch defined as iteration >250 mini-batches. The loss function is the sum of Dice loss and cross-entropy.[19] The nnU-Net trained 2 configurations (2- and 3-dimensional U-Net) in a 5-folder cross-validation, and used the optimal ensemble of these models to choose the best model, which could also be a combination of 2 models according to validation performance.[19,20]

Evaluation

Objective evaluation

We calculated PPV, sensitivity, and Dice similarity coefficient (DSC) of the entire segmentation mask for each patient based on a per-patient level and metastasis level in 3 categories of metastasis size (<3 mm, ≥3 mm and <6 mm, and ≥6 mm), as well as for all tumor sizes, to enable a comparison with previous studies of brain metastasis contouring. PPV tells us the probability that a predicted metastasis is truly positive, and sensitivity shows the likelihood that a negative result (no metastasis) is truly negative. In other words, a model with high PPV and sensitivity has a low FP and false negative (FN) rate, respectively. The definitions of these metrics are as follows:

$$PPV = \frac{\text{TP}}{TP + FP}$$

$$Sensitivity = \frac{\text{TP}}{TP + FN}$$

where true positive (TP) reflects the number of metastases correctly identified, FP the number of metastases incorrectly identified, and FN the number of metastases not identified. DSC computes the overlap of automatic segmentation U$_s$ and ground truth segmentation U$_g$ as follows:

$$DSC = \frac{2 * \| \, |\text{U}_s \cap U_g \, \|}{\left\| U_s + U_g \right\|}$$

Table 1   Patient demographics and primary cancer types in training and testing sets

	Training set	Testing set	Total
Number of patients	845	206	1051
Mean age, y	59.4	57.5	59.3
Sex, male:female	406:439	101:105	507:544
Magnetic field, 1.5T:3T)	806:39	192:14	998:53
Mean size of metastases, mm	10.1 ± 7.4	12.3 ± 9.2	10.5 ± 7.8
Mean metastases per patient, n	4.1 ± 3.6	4.1 ± 3.1	4.1 ± 3.5
Number of metastases <3 mm	78	60	138
Number of metastases ≥3 mm and <6 mm	1267	324	1591
Number of metastases ≥6 mm	2137	546	2683
Primary cancer type, n (%)			
Non-small cell lung	349 (41.3)	61 (29.6)	410 (39)
Melanoma	166 (19.6)	53 (25.7)	219 (20.8)
Breast	117 (13.8)	38 (18.4)	155 (14.7)
Renal	87 (10.3)	21 (10.4)	108 (10.3)
Gastrointestinal	30 (3.6)	9 (4.4)	39 (3.7)
Head and neck	20 (2.4)	4 (1.9)	24 (2.3)
Genitourinary	19 (2.2)	7 (3.4)	26 (2.5)
Thyroid	18 (2.1)	3 (1.5)	21 (2)
Sarcoma	14 (1.7)	4 (1.9)	18 (1.7)
Small cell lung	13 (1.5)	2 (1)	15 (1.4)
Neuroendocrine carcinoma	8 (0.9)	3 (1.5)	11 (1)
Thymic	3 (0.4)	0 (0.0)	3 (0.3)
Nonmelanoma skin	1 (0.1)	1 (0.5)	2 (0.2)

On a per-patient level, PPV and sensitivity values were first evaluated for each patient using FPs, FNs, and TPs, and then the mean was calculated among all patients in each category. When calculating PPV and sensitivity per patient in each category, for special cases in which TP, FP, and TN were all 0, we defined the PPV and sensitivity measures as 1. If TP was 0 and FN and FP were >0, we defined the PPV and sensitivity measures as 0. On the per-metastasis level, the PPV and sensitivity values were calculated based on the total number of TPs, FPs, and FNs in the testing set.

Subjective evaluation

The quality of our contours was assessed by 2 central nervous system radiation oncologists, who assigned scores from 1 to 5 as follows: 5 means strongly agree and the contour can be used as is; 4 means agree, minor edits are there but not necessary; 3 means neither agree nor disagree, minor edits are necessary; 2 means disagree, major edits are necessary; and 1 means strongly disagree, unusable. A total of 24 patients' MRI scans were randomly sampled from the testing set, so each reviewer assessed 12 MRI scans. To avoid bias in their scoring, we only shared the predicted contours with the reviewers without showing the ground truth.

False positive investigation

The BM-Net detected metastasis, and WB-Net segmented the brain volume as the region of interest in the MRI scans. All metastases that were detected outside of the brain in the WB-Net were removed using the BM-Net in a postprocessing step. The initial FP detection of brain metastases based on the segmentation model was further investigated by using a combination of radiology reports and follow-up MRI as reference to recategorize the FP detection. If a current FP metastasis was marked as a tumor in the radiology report, the FP metastasis was recategorized as a TP metastasis. If the radiology report did not consider the FP metastasis a tumor but follow-up MRI confirmed the FP metastasis as a tumor, this FP metastasis was recategorized as TP detection as well. These recategorized metastases were added to the testing data set, and updated the number testing metastases in Table 1. Table 2 lists 5 possible categories of the initial FP metastases and their recategorization after this further examination.

Results

False positive detection

Our segmentation model identified a total of 164 FP metastases based on the contours drawn for Gamma-knife treatment planning. After further FP investigation by including radiology reports and follow-up MRI scans, these 164 FP metastases were grouped into 5 different categories as shown in Table 2. This recategorization reduced the FP metastases to 94. The other 70 metastases were real metastases not contoured for Gamma-knife treatment because of a medical reason or being indiscernible at the time of treatment. These metastases were added to the testing metastases to calculate model performance. Figure 2 presents 1 example from each group of this false FP detection (groups 2-5 in Table 2).

Model performance

By moving these actual noncontoured metastases (false FP detections) to the TP detection list, our tool detected only 94 cases that were actual FP detection. The performance of the model was evaluated by calculating TP, FP, FN, PPV, sensitivity, and DSC in 3 size categories in addition to all sizes on the metastases and per-patient levels. Table 3 shows the evaluated metrics of our model. On a per-patient level for all metastasis sizes, PPV, sensitivity, DSC, and FPs per patient were $90.1\% \pm 17.7\%$, $88.4\% \pm 18.0\%$, $82.2\% \pm 9.5\%$, and 0.4 ± 1.0, respectively. For metastases ≥ 6 mm, these metrics were $95.6\% \pm 17.5\%$, $94.5\% \pm 18.1\%$, $86.8\% \pm 7.5\%$, and $0.1\% \pm 0.4$, respectively. The mean number of metastases in the testing set was 4.1 ± 3.1 per patient.

Figure 3 shows the box-whisker plot of PPV and sensitivity in the 4 categories of sizes on a per-patient level. PPV was higher than sensitivity in all categories, and the model had the best performance in detecting metastases ≥ 6 mm for both metrics. Table 3 also shows the evaluated metrics on the per-metastasis level in the 4 size categories. With all metastasis sizes included, the model PPV was 790/884 (89.4 %), sensitivity was 790/930 (84.9%), and DSC was $80.4\% \pm 15.9\%$. For metastases ≥ 6 mm, the PPV, sensitivity, and DSC values were 523/540 (96.9%), 523/546 (95.8 %), and $86.5\% \pm 9.8\%$, respectively.

Subjective evaluation

The review assessments from the 2 radiation oncologists were very similar. The first physician evaluated the quality of 65 metastases, and 77% of the autosegmented contours had a score of ≥ 4 and 23% had a score of 3. The second physician evaluated the quality of 52 metastasis, and 74% had a score of ≥ 4 and 26% had a score of 3. None of the metastases that the reviewers assessed had a score <3. Overall, the quality of the autosegmented contours can be assessed as 4.0, which shows that contours predicted by the model can be used as is or with only minor editing.

Discussion

This work presents an evaluation of a deep learning approach for brain metastasis segmentation. The model detected $88.4\% \pm 18.0\%$ of tumors on average for each patient with a PPV of $90.1\% \pm 17.7\%$. Our results show that the performance of the model depends on the metastases size, and the model has better performance in detecting and contouring larger metastases (≥ 6 mm) than smaller ones.

The model achieved an overall DSC of $82.2\% \pm 9.5\%$, which shows good segmentation performance. However, interobserver variability in target volume delineations of brain metastases for SRS has been always a challenge for quality assurance in clinical trials.[9,21] Therefore, assessing the quality of contours by experienced physicians seems a better way of evaluating segmentation. The subjective

Table 2   Investigation of 164 false-positive cases using radiology report and follow-up magnetic resonance imaging scans

| Group	Number of metastases	Radiology report	Ground truth (treatment)	Follow up	Classification
1	94	Not tumor		Not tumor	False positive
2	38	Tumor		Unknown	True positive
3	15	Not tumor	Not contoured	Tumor	True positive/early detection
4	6	Suspect tumor		Tumor	True positive
5	11	Tumor (identified for whole brain treatment)		Unknown	True positive
Unknown indicates that follow-up magnetic resonance imaging had not been checked because of confirmation of a tumor on the radiology report.					
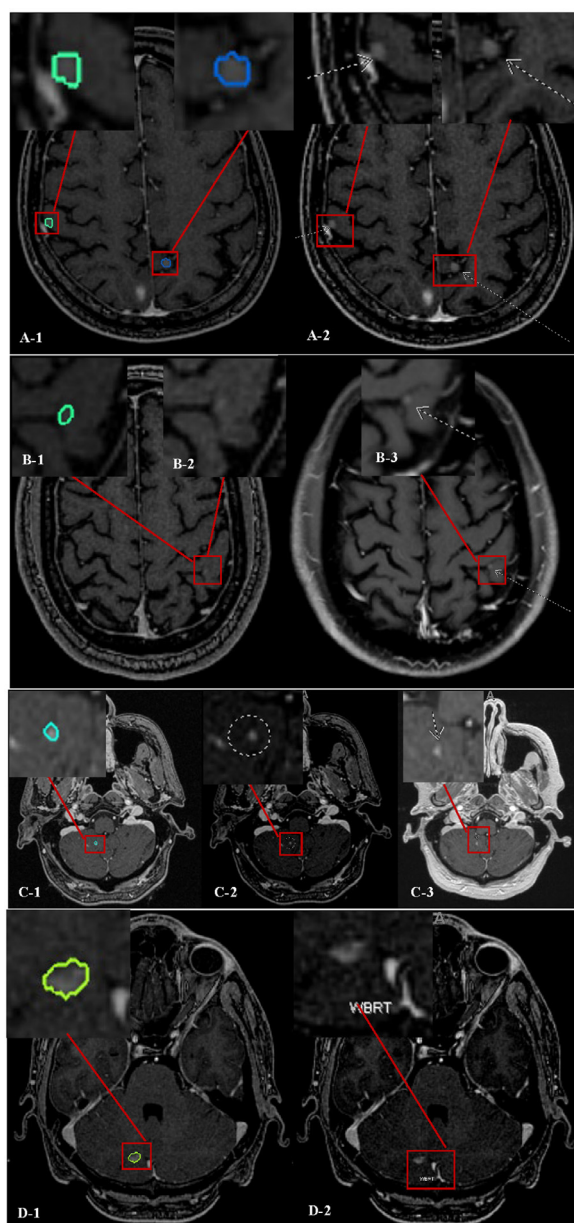
Figure 2 A, False false-positive detection, with the left image showing our model-predicted tumors. The radiologist also marked the metastases as tumors (right image), but the physician did not contour these tumors for medical reasons; B, early detection, with the left image showing our contour, but the radiologist did not report a tumor at this location. In the follow-up magnetic resonance imaging scan 4 months later (right image), the radiologist marked this location as a tumor; C, false false-positive detection, with the left image showing the predicted contours, and the middle image is the radiologist's image showing a suspected tumor. However, in a follow-up image 4 months later (right image), the radiologist indicated confidence of a tumor at this location; and D, false false-positive detection. Although our model correctly detected the metastasis in the left image, since the patient had undergone whole brain radiation therapy, treatment of the lesion was not required, and the lesion was not contoured in the ground truth image.

evaluation showed that 75% of the autocontours were clinically acceptable and could be used as is, and the remaining 25% needed only minor editing. This finding shows consistent and accurate contouring for clinical application.

Our investigation of FPs showed the potential of our model for clinical application in assisting radiologists and other physicians in detecting indiscernible metastases in early stages. The model can also help radiologists make more precise decisions when a lesion is suspected of being a tumor or abscess. At the metastasis level, the model achieved near-perfect detection performance in large metastases ($\geq$6 mm), with PPV and sensitivity of approximately 97% and 96%, respectively.

In reviewing the chart for cases with a false FP detection in Table 2, we found that the main medical reason for not contouring those metastases for Gamma Knife treatment was whole brain radiation therapy (WBRT) or after Gamma Knife treatment. As a general practice, for patients who had received WBRT before Gamma Knife treatment, physicians would typically treat only lesions that showed progression of disease since the time of completion of WBRT. Separately, in some cases, physicians treated patients with Gamma Knife before planned WBRT, essentially offering the boost dose before WBRT, which could include treatment of the largest lesions, symptomatic lesions, or lesions potentially in eloquent areas to increase the likelihood of treatment response to these sites.

For patients with numerous metastases, there would be too many lesions to treat with Gamma Knife, so a physician might plan on treating only the lesions most concerning first and resort to WBRT for other lesions. Lesions noted by the radiologist as indeterminate/suspected but not definitive were also not typically treated to avoid mistreatment or unnecessary toxicity. In routine clinics, these suspected lesions are further evaluated during follow up. Importantly, this study focuses on the ability to detect all lesions and not necessarily whether those lesions are clinically indicated for treatment targeting. Thus, there are understandable differences in the detection or contouring of identified lesions compared with the lesions marked and contoured for treatment, including history of WBRT or deferring treatment of indeterminate lesions.

Several previous works have investigated the fully automatic segmentation of brain metastases. Xue et al.[10] built a cascaded 3-dimensional fully convolution network to detect and segment brain metastases using 1201 patient T1 contrast MRI scans, and split them into training and testing sets with a 75:25 ratio. The researchers reported sensitivity and DSC rates of 96% and 85% $\pm$ 8%, respectively, but these results were limited to large metastases ($\geq$6 mm) only. Zhang et al.[22] trained and tested a Faster region-based convolutional neural network (Faster R-CNN) on 270 and 45 patients T1 contrast MRI scans, respectively, and reported a sensitivity rate and FP per patient of 96% $\pm$ 12% and 20 $\pm$ 13, respectively.

Table 3   Model metrics on per-patient level in 4 size categories

Metric	<3 mm	≥3 mm and <6 mm	≥6 mm	All sizes
Per-patient level				
Positive predictive value, %	85.0 ± 34.9	90.7 ± 26.2	95.6 ± 17.5	90.1 ± 17.7
Sensitivity, %	94.1 ± 23.0	83.4 ± 30.8	94.5 ± 18.1	88.4 ± 18.0
Dice similarity coefficient, %	31.2 ± 14.5	72.0 ± 10.1	86.8 ± 7.5	82.2 ± 9.5
True positive	0.2 ± 0.6	1.1 ± 1.6	2.5 ± 1.8	4.0 ± 3.1
False positive	0.2 ± 0.5	0.2 ± 0.6	0.1 ± 0.4	0.4 ± 1.0
False negative	0.1 ± 0.4	0.5 ± 0.9	0.1 ± 0.4	0.7 ± 1.2
Per-metastasis level				
Positive predictive value, %	50.6	85.9	96.9	89.4
Sensitivity, %	68.3	69.8	95.8	84.9
Dice similarity coefficient, %	30.7 ± 16.7	71.0 ± 12.2	86.5 ± 9.8	80.4 ± 15.9
True positive	41	226	523	790
False positive	40	37	17	94
False negative	19	98	23	140

Dikici et al.[23] developed a 2-stage model to first detect image points with high probability of representing brain metastasis and a custom-built CNN to classify these points. The researchers reported sensitivity as 90% and FP as 9 ± 3 per patient. Grøvik et al.[18] trained a CNN model based on GoogLeNet[24] architecture using 105 and 51 patient multisequence MRI scans for training and testing, respectively. The researchers reported PPV, sensitivity, DSC, and FP per patient as 79% ± 20%, 53% ± 22%, 79% ± 12%, and 8.3 ± 13, respectively. Zhou et al.[8] developed a 2-stage deep learning algorithm using 748 and 186 postcontrast T1-weighted MRI scans for training and testing, respectively. Their model included a single-shot detector to first detect regions containing metastases, followed by a fully convolutional network to segment the metastases from these regions. The researchers reported PPV as 58% ± 25%, sensitivity as 88% ± 19%, DSC as 85% ± 13%, and FPs as 3 ± 3 per patient.
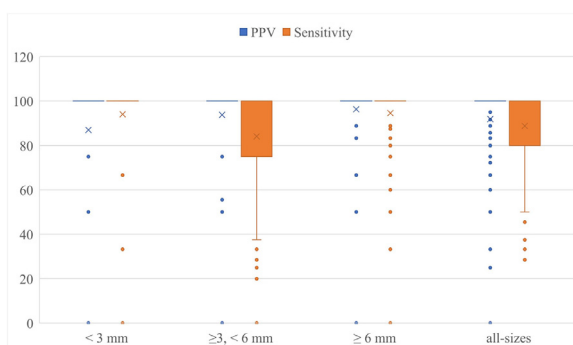
Figure 3 Comparison of box-whisker plot of positive predictive value and sensitivity of the model in 4 size categories on a per-patient level.

Charron et al[17] adapted an existing 3-dimensional convolutional neural network (DeepMedic[25]) to detect and segment brain metastases using 164 and 18 multisequence MRI scans for training and testing, respectively. The researchers obtained a sensitivity of 93% ± 20%, DSC of 79% ± 21%, and FPs of 8 ± 7 per patient. Lui et al.[16] developed a CNN-based algorithm using 225 and 15 postcontrast T1 MRI scans for training and testing, respectively, and reported DSC of 67% ± 3%. Yoo et al.[26] developed a deep learning model, using 58 patient MRI scans for training and 12 for testing with small brain metastases (volume <67 cc). The researchers applied training techniques to the well-known 2-dimensional U-Net, and obtained a sensitivity of 97%, average FP rate of 1.25 per patient, and DSC of 75%.

Grøvik et al.[27] also developed a deep learning model in which a neural network was trained on 4 distinct MRI sequences using an input-level dropout layer. The researchers used a training set of multisequence MRI for 100 patients, and validated/tested on 10/55 patients, and the test set was missing 1 of 4 MRI sequences used for training. Grøvik et al. reported PPV as 79%, sensitivity as 67%, and DSC as 75.5%. By comparison, our model, with a PPV of 90% ± 17%, achieved the best detection performance and, with 0.4 ± 1.0 FPs per patient, achieved the lowest FP value. Moreover, our data set is one of the largest data sets in existing similar studies, and we examined all different metastasis sizes. The higher PPV and sensitivity rates using our method implies that our approach may save more time in contouring these tumors for treatment.

One major limitation of our approach is the small lesion detection. The PPV for tumor size <3 mm is 50.6%, showing that around half of detected lesions are

FP. Sensitivity is 68%, showing that around 30% of lesions cannot be correctly detected. The low accuracy in small lesion detection might be caused by the imaging noise/artifacts and small vessels in the brain, because their appearance on MRI scans is similar to that of small lesions. Increasing training data to include more small lesions could potentially reduce the effect from imaging noise/artifacts or small brain vessels.

In Table 2, we show that a total of 21 lesions initially not marked as tumors or marked as suspect tumors were confirmed as tumors during a follow-up visit, which shows the potential of this tool in helping identify lesions in an early stage. However, because of the different time spans from treatment to follow-up imaging and given the many FP detections in small lesions, we are unable to draw a definitive conclusion that this tool can identify lesions in an early stage. Radiologists need to use their best judgment to determine whether a tumor or not by using this tool as an auxiliary to make their decision.

Conclusions

We developed an automated framework to detect brain metastases on MRI scans. Our model achieved a high PPV and sensitivity value with a high level of clinical acceptability that can provide consistent and accurate contouring for clinical application. This model could potentially help radiologists and oncologists detect and contour tumors, make precise decisions on suspected metastases, and potentially find metastases at an early stage. A deep dive into FPs indicated that approximately 40% of metastases were TP metastases not contoured in ground truth because of other medical reasons or being indiscernible at the time of treatment, but identified later by a radiologist on subsequent follow-up imaging.

Acknowledgments

References

1. Soffietti R, Ruda R, Mutani R. Management of brain metastases. *J Neurol*. 2002;249:1357-1369.
2. Arvold ND, Lee EQ, Mehta MP, et al. Updates in the management of brain metastases Enhanced Reader. *Neuro Oncol*. 2016;18:1043-1065.
3. Müller-Riemenschneider F, Bockelbrink A, Ernst I, et al. Stereotactic radiosurgery for the treatment of brain metastases. *Radiother Oncol*. 2009;91:67-74.
4. Nayak L, Lee EQ, Wen PY. Epidemiology of brain metastases. *Curr Oncol Rep*. 2012;14:48-54.
5. Steeg PS, Camphausen KA, Smith QR. Brain metastases as preventive and therapeutic targets. *Nat Rev Cancer*. 2011;11:352-363.
6. Suh JH. Stereotactic radiosurgery for the management of brain metastases. *N Engl J Med*. 2010;362:1119-1127.
7. Canon C, Beecham Chick JF, DeQuesada I, Gunderman RB, Hoven N, Prosper AE. Physician burnout in radiology: Perspectives from the field. *AJR Am J Roentgenol*. 2022;218:370-374.
8. Zhou Z, Sanders JW, Johnson JM, et al. MetNet: Computer-aided segmentation of brain metastases in post-contrast T1-weighted magnetic resonance imaging. *Radiother Oncol*. 2020;153:189-196.
9. Higuchi Y, Yamamoto M, Serizawa T, Aiyama H, Sato Y, Barfod BE. Modern management for brain metastasis patients using stereotactic radiosurgery: Literature review and the authors' Gamma Knife treatment experiences. *Cancer Manag Res*. 2018;10:1889-1899.
10. Xue J, Wang B, Ming Y, et al. Deep learning-based detection and segmentation-assisted management of brain metastases. *Neuro Oncol*. 2020;22:505-514.
11. Pennig L, Shahzad R, Caldeira L, et al. Automated detection and segmentation of brain metastases in malignant melanoma: Evaluation of a dedicated deep learning model. *AJNR Am J Neuroradiol*. 2021;42:655-662.
12. Conson M, Cella L, Pacelli R, et al. Automated delineation of brain structures in patients undergoing radiotherapy for primary brain tumors: From atlas to dose-volume histograms. *Radiother Oncol*. 2014;112:326-331.
13. Qeshmy DE, Makdisi J, Ribeiro da Silva E, Angelis JJ. Managing human errors: Augmented reality systems as a tool in the quality journey. *Proc Manufact*. 2019;28:24-30.
14. Bousabarah K, Ruge M, Brand JS, et al. Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. *Radiat Oncol*. 2020;15:13014.
15. Leibe B, Matas J, Sebe N, Welling M. *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. New York City, NY; . Springer International Publishing; 2016.
16. Liu Y, Stojadinovic S, Hrycushko B, et al. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. *PLoS One*. 2017;12: e0185844.
17. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Comput Biol Med*. 2018;95:43-54.
18. 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.
19. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203-211.
20. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. *IEEE Access*. 2021;9:16591-16603.
21. Chang ATY, Tan LT, Duke S, Ng WT. Challenges for quality assurance of target volume delineation in clinical trials. *Front Oncol*. 2017;7:221.
22. Zhang M, Young GS, Chen H, et al. Deep-learning detection of cancer metastases to the brain on MRI. *J Magn Reson Imaging*. 2020;52:1227-1236.
23. Dikici E, Ryu JL, Demirer M, et al. Automated brain metastases detection framework for T1-weighted contrast-enhanced 3D MRI. *IEEE J Biomed Health Inform*. 2020;24:2883-2893.
24. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA; . IEEE; 2015.
25. Withey DJ, Koles ZJ. A review of medical image segmentation: Methods and available software. *Methods*. 2008;10:125-148.
26. Yoo SK, Kim TH, Chun J, et al. Deep-learning-based automatic detection and segmentation of brain metastases with small volume for stereotactic ablative radiotherapy. *Cancers (Basel)*. 2022;14:2555.
27. Grøvik E, Yi D, Iv M, et al. Handling missing MRI sequences in deep learning segmentation of brain metastases: a multicenter study. *NPJ Digit Med*. 2021;4:33.