Automated brain metastases segmentation with a deep dive into false positive detection

Hamidreza Ziyaee Ph.D., Carlos E. Cardenas Ph.D.,
D. Nana Yeboa M.D., Jing Li M.D.Ph.D.,
Sherise D. Ferguson M.D., Jason Johnson M.D.,
Zijian Zhou Ph.D., Jeremiah Sanders Ph.D.,
Raymond Mumme B.S., Laurence Court Ph.D.,
Tina Briere Ph.D., Jinzhong Yang Ph.D.

PII: S2452-1094(22)00191-9
DOI: https://doi.org/10.1016/j.adro.2022.101085
Reference: ADRO 101085

To appear in: Advances in Radiation Oncology

Received date: 17 July 2022
Accepted date: 15 September 2022

Please cite this article as: Hamidreza Ziyaee Ph.D., Carlos E. Cardenas Ph.D.,
D. Nana Yeboa M.D., Jing Li M.D.Ph.D., Sherise D. Ferguson M.D., Jason Johnson M.D.,
Zijian Zhou Ph.D., Jeremiah Sanders Ph.D., Raymond Mumme B.S., Laurence Court Ph.D.,
Tina Briere Ph.D., Jinzhong Yang Ph.D., Automated brain metastases segmentation with
a deep dive into false positive detection, Advances in Radiation Oncology (2022), doi:
https://doi.org/10.1016/j.adro.2022.101085

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition
of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of
record. This version will undergo additional copyediting, typesetting and review before it is published
in its final form, but we are providing this version to give early visibility of the article. Please note that,
during the production process, errors may be discovered which could affect the content, and all legal
disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc. on behalf of American Society for Radiation Oncology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Automated brain metastases segmentation with a deep dive into false positives detection

[Short Running Title]
Automated brain metastases segmentation

[Author Names]
Hamidreza Ziyaee, Ph.D.¹
Carlos E. Cardenas, Ph.D.²
D. Nana Yeboa, M.D.³
Jing Li, M.D., Ph.D.³
Sherise D. Ferguson, M.D.⁴
Jason Johnson, M.D.⁵
Zijian Zhou, Ph.D.⁶
Jeremiah Sanders, Ph.D.⁶
Raymond Mumme, B.S.¹
Laurence Court, Ph.D.¹
Tina Briere, Ph.D.¹
Jinzhong Yang, Ph.D.¹

[Author Institutions]
¹ Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States
² Department of Radiation Oncology, The University of Alabama at Birmingham, Birmingham, Alabama, United States
³ Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States
⁴ Department of Neurosurgery, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States
⁵ Department of Neuroradiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States
⁶ Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States
[Corresponding Author Name & Email Address]
Jinzhong Yang
JYang4@mdanderson.org

[Author Responsible for Statistical Analysis Name & Email Address]
Hamidreza Ziyaee
hziyaee@mdanderson.org

[Conflict of Interest Statement for All Authors]

Hamidreza Ziyaee
No Disclosures

Carlos E. Cardenas
No Disclosures

D. Nana Yeboa
All support for the present manuscript - Brockman Medical Foundation Grant
Grants or contracts from any entity – MD Anderson Cancer Center Shirly Stein Award (Grant), Robert Wood Johnson Foundation (Grant)

Jing Li
No Disclosures

Sherise D. Ferguson
Grant or contracts from any entity – Research funding from Codiak
Support for attending meeting and/or travel – Travel support from Emerson Collective Cancer

Jason Johnson
Grants or contracts from any entity – Blue Earth Diagnostics
Consulting fees – BioClinica, Kura Oncology, InformAI

Zijian Zhou
No Disclosures

Jeremiah Sanders
Abstract

Purpose: Clinical management of brain metastases after stereotactic radiosurgery is a difficult task because a physician must review follow-up magnetic resonance (MR) images 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 MR imaging to help treatment planning for stereotactic radiosurgery, and to understand its limitations.

Methods and Materials: Two self-adaptive nnU-Net models trained on post-contrast 3D T1-weighted MR images from stereotactic radiosurgery patients. The performance was evaluated by computing positive predictive value (PPV), sensitivity, and Dice similarity coefficient (DSC).

The training and testing sets, respectively, included 3482 metastases on 845 patient MR images and 930 metastases on 206 patient MR images.

Results: On the per-patient, PPV was 90.1±17.7%, sensitivity was 88.4±18.0%, DSC was 82.2±9.5%, and false positive (FP) per patient was 0.4±1.0; for large metastases (≥ 6mm), the per-patient PPV was 95.6±17.5%, sensitivity was 94.5±18.1%, DSC was 86.8±7.5%, and FP was 0.1±0.4. The quality of auto-segmented true positive contours was also assessed by two physicians using a 5-point scale for clinical acceptability. 75% of contours were assigned scores of 4 or 5, which shows 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 true-positive (TP) metastases that were not identified on the original radiology review but were identified on subsequent follow-up imaging (early detection).

54% were real metastases (TP) that were identified but purposefully not contoured for target treatment mainly because patient underwent whole brain radiotherapy pre/post SRS treatment.

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

1. Introduction
Around 20 - 40% of cancer patients develop brain metastases.\(^1\) Stereotactic radiosurgery (SRS) is a functional and routinely used treatment for brain metastases.\(^2,3\) This method uses multiple cobalt sources (Gamma Knife) or a linear accelerator (Linac) and delivers 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 manually detected by a radiologist, and contoured by a radiation oncologist using radiotherapy planning software.\(^7,8\) However, there were reports of missing small and even large brain metastases in clinical practice during the planning of Gamma Knife radiosurgery due to 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 potentials 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 of above 80%.\(^8\) However, those studies usually reported a large number of false positive (FP) findings and low positive predictive value (PPV) at the same time, and none have reported the early detection of metastases.\(^10,16-18\)

Developing an automated framework to detect metastases from brain MRIs 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 impact in clinical applications. The purpose of this study is to develop and validate an automated framework to contour brain metastases from MR images using deep neural networks with improved false positive detection and early detection of metastases.

## 2. Methods and materials

We developed a framework using nnU-Net\textsuperscript{19} for brain metastasis detection in postcontrast T1-weighted MRIs (Figure 1). In the framework, we trained two segmentation models based on self-adaptive nnU-Net, a brain metastases detector (BM-Net) for segmenting the metastases, and a whole-brain network (WB-Net) for segmenting brain volume as our region of interest (ROI) for metastases detection.

### 2.1. Patient data

This study was approved by the local institutional review board (protocol number removed for review). The dataset in this study was comprised of one planning MRI per patient from 1051 patients who underwent Gamma Knife treatment at [xxx institution (anonymized for review)] 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 3D T1-weighted MRI sequence with and without contrast echo. Acquisition parameters of 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 MR, repetition time 5.4-10 ms, flip angle 12°-20°, matrix size 256 × 256, voxel size 0.94 × 0.94 × 1 mm\textsuperscript{3}. The dataset included patients 16 to 82 years old with a total of 4442 brain metastases, the largest of which was 63 mm, and metastases per
patient was up to 27. The mean number of metastases was $4 \pm 3.5$ per patient, and the mean metastasis size was $10 \pm 7.4$ mm. These metastatic tumors were contoured by experienced oncologists. Only those tumors that were identified for Gamma Knife treatment were contoured and these contours were used as ground-truth for segmentation comparison in this study. It is worth noting that some tumors might not be contoured due to varied medical reasons, which will be detailed in the section of false positive detection.

The patient dataset was randomly split to 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 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 was manually contoured from the MRI images to train the WB-Net.

2.2. Segmentation

As shown in Figure 1, the training set was used to train the BM-Net to detect metastases, and the subset of 30 MRIs was used to train 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 requires no changes in the architecture of the network and takes care of hyper-parameter tuning. Specifically, the data augmentation and preprocessing step in 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\textsuperscript{20} network consists of a series of convolutional layers which reduce the dimension of input image ($256 \times 256$) and stack together several $3 \times 3$ convolutional maps with padding followed by a rectified linear unit (ReLU). It uses $2 \times 2$ max-pooling layers with stride 2 and contracts until it reaches at the bottommost. This is the encoder part where the model grows the predetermined number of channels. From the bottommost, the model starts on the up-convolution path, which is called decoder or expanding part. At every stage of the up-convolution, the model concatenates the results of the corresponding step from the down-sampling path, which adds robustness to the network.

The training configuration of 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 optimizer. Networks were trained for 1000 epochs with each epoch being defined as iteration over 250 mini-batches. The loss function is the sum of Dice loss and cross-entropy.\textsuperscript{19} The nnU-Net trained two configurations (2D U-Net, 3D U-Net) in a five-folder cross validation and used the optimal ensemble of these models to choose the best model, which could also be a combination of two models according to validation performance.\textsuperscript{19,20}

2.3. Evaluation

2.3.1. 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 three categories of metastasis size (less than 3 mm, greater than or equal to 3 mm and less than 6 mm, and 6 mm or larger) as well as for all tumor sizes to enable comparison with previous studies of brain
metastasis contouring. PPV tells us the probability that a predicted metastasis is truly positive, and sensitivity says the likelihood that a negative result, no metastasis, is truly negative. In other words, a model with high PPV and sensitivity has a low false positive rate and false negative rate, respectively. The definitions of these metrics are given below:

\[
\text{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

Where true positive (TP) reflects the number of metastases correctly identified. The false positive (FP) reflects the number of metastases incorrectly identified. The false negative (FN) reflects the number of metastases not identified.

DSC compute the overlap of automatic segmentation \( U_s \) and ground truth segmentation \( U_g \):

\[
\text{DSC} = \frac{2 * ||U_s \cap U_g||}{||U_s + U_g||}
\]

On a per-patient level, the 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 the FN and FP were larger than 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.

2.3.2. Subjective evaluation

The quality of our contours was assessed by two CNS radiation oncologists and assigned scores from 1 to 5 where:
- 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
- 1 means strongly disagree, unusable

Twenty-four patients’ MRIs were randomly sampled from the testing set, so each reviewer assessed 12 MRIs. To avoid bias in their scoring, we only shared the predicted contours with the reviewers without showing the ground truth.

2.4. False positive investigation

The BM-Net detected metastasis, and WB-Net segmented the brain volume as the ROI in the MR images. All metastases that were detected outside the brain in WB-Net were removed using 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 MRIs as reference to re-categorize the FP detection. If a current FP metastasis was marked as a tumor in the radiology report, the FP metastasis was re-categorized as a TP metastasis. If the radiology report did not consider the FP metastasis as a tumor, but the follow-up MRIs confirmed the FP metastasis as a tumor, this FP metastasis was re-categorized as TP detection as well. These re-categorized metastases were added to testing dataset and updated the number testing metastases in Table 1. Table 2 lists 5 possible categories of the initial FP metastases and their re-categorization after this further examination.

3. Results
3.1. 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 MRIs, these 164 FP metastases were grouped into 5 different categories as shown in Table 2. This re-categorization reduced the FP metastases to 94. The other 70 metastases were real metastases that were not contoured for Gamma-knife treatment due to some medical reason or indiscernible at the time of treatment. They were added to testing metastases to calculate the model performance. Figure 2 presents one example from each group of this false FP detection (groups 2 – 5 in Table 2).

3.2 Model Performance

By moving these actual non-contoured metastases “false-FP detections” to the TP detection list, our tool detected only 94 cases that are actual FP detection. The performance of the model was evaluated by calculating TP, FP, FN, PPV, sensitivity, and DSC in three size categories in addition to all sizes on the metastases level and per-patient level. 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 ± 17.7%, 88.4 ± 18.0%, 82.2 ± 9.5%, and 0.4 ± 1.0, respectively. For metastases 6 mm or larger, these metrics were 95.6 ±17.5%, 94.5 ± 18.1%, 86.8 ± 7.5%, and 0.1 ± 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 four categories of sizes on a per-patient level. PPV was higher than sensitivity in all categories, and for both metrics the model had the best performance in detecting metastases 6 mm or larger.
Table 3 also shows the evaluated metrics on the per-metastasis level in the four 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 ± 15.9 %. For metastases 6 mm or larger, the PPV, sensitivity, and DSC values were 523/540 (96.9 %), 523/546 (95.8 %), and 86.5 % ± 9.8 %, respectively.

3.3. Subjective evaluation

The review assessments from two radiation oncologists were very similar. The first physician evaluated the quality of 65 metastases, and 77 % of the auto-segmented contours had a score of 4 or higher and 23 % had a score of 3. The second physician evaluated the quality of 52 metastasis, and 74 % had a score of 4 or higher and 26 % had a score of 3. None of the metastases that reviewers assessed had a score below 3. Overall, the quality of the auto-segmented 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.

4. Discussion

This work presents an evaluation of a deep learning approach for brain metastasis segmentation. The model detected 88.4 ± 18.0 % of tumors on average for each patient with a PPV of 90.1 % ± 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 ± 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.\textsuperscript{9,21}

Therefore, assessing the quality of contours by experienced physicians seems a better way of evaluating segmentation. The subjective evaluation showed that 75\% of the auto-contours 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 decision when they suspect a lesion of being a tumor or abscess. At the metastasis level, the model achieved near-perfect detection performance in large metastases (\geq 6 \text{ mm}), with PPV and sensitivity of about 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 the whole-brain radiation therapy (WBRT) treatment before or after Gamma Knife. As a general practice, for patients who had received WBRT prior to Gamma Knife, physicians would typically treat only lesions that showed progression of disease since the time of completion of WBRT.

Separately, in some cases, physicians sometimes treated patients with Gamma Knife prior to planned WBRT, essentially offering the boost dose prior to WBRT. This could include treatment of the largest lesions, symptomatic lesions, or lesions potentially in eloquent areas in order 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 that were 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 in order to avoid mistreatment or unnecessary toxicity. In routine clinic,
these suspected lesions will further be evaluated during follow-up. Importantly, this study focuses on the ability to detect all lesions and not necessarily whether those lesions were clinically indicated for treatment targeting. Thus, there are understandable differences in the detection or contouring of identified lesions in comparison to the lesions marked and contoured for treatment. This could include history of WBRT or deferring treatment of indeterminate lesions.

Several previous works have investigated the fully automatic segmentation of brain metastases. Xue et al\textsuperscript{10} built a cascaded 3D fully convolution network to detect and segment brain metastases using 1201 patient T1 contrast MRIs and split them into training and testing set with 75:25 ratio. They reported the sensitivity and DSC of 96\% and 85 $\pm$ 8\% respectively but these results were limited to large metastases ($\geq$ 6 mm) only. Zhang et al\textsuperscript{22} trained and tested a Faster region-based convolutional neural network (Faster R-CNN) on 270 and 45 patients T1 contrast MRIs, respectively. They reported sensitivity rate and false-positive per patient as 96$\pm$12\% and 20$\pm$13, respectively. Dikici et al\textsuperscript{23} developed a two stages model to first detect image points with high probability of representing brain metastasis and next a custom-built CNN to classify these points. They reported sensitivity as 90\% and false-positive as 9$\pm$3 per patient. Grøvik et al\textsuperscript{18} trained a CNN model based on GoogLeNet\textsuperscript{24} architecture using 105 and 51 patient multi sequence MRIs for training and testing, respectively. They reported PPV, sensitivity, DSC, and false-positive per patient as 79$\pm$20\%, 53$\pm$22\%, 79$\pm$12\%, and 8.3$\pm$13, respectively. Zhou et al\textsuperscript{8} developed a 2-stage deep learning algorithm using 748 and 186 post-contrast T1-weighted MRI for training and testing respectively. Their model included a single-shot detector to first detect regions containing metastases followed by a fully convolutional network to segment the metastases from these regions. They reported PPV as 58$\pm$25\%, sensitivity as 88$\pm$19\%, DSC as 85$\pm$13\%, and FPs as 3$\pm$3.
Charron et al. adapted an existing 3D convolutional neural network (DeepMedic) to detect and segment brain metastases using 164 and 18 multi sequence MRIs for training and testing respectively. They obtained a sensitivity of $93 \pm 20\%$, DSC of $79 \pm 21\%$, and FPs of $8 \pm 7$ per patient. Lui et al. developed a CNN-based algorithm using 225 and 15 postcontrast T1 MRIs for training and testing respectively. They reported DSC of $67 \pm 3\%$. Yoo et al. developed a DL using 58 patients MRIs for training and 12 patients MRIS for testing with small brain metastases (volume less than 67cc). They applied training techniques to the well-known 2D U-Net and obtained sensitivity of $97\%$, average false positive rate of $1.25$ per patient, and dice coefficient of $75\%$. Grøvik et al. also developed a deep learning model (ILD model) in which a neural network was trained on four distinct MRI sequences using an input-level dropout layer. They used a training set of multisequence MRI for 100 patients and validated/tested on 10/55 patients. The test set was missing one of the four MRI sequences used for training. They reported PPV as $79\%$, sensitivity as $67\%$, and Dice as $75.5\%$. By comparison, our model, with a PPV of $90\% \pm 17\%$, achieved the best detection performance and, with $0.4 \pm 1.0$ FPs per patient, achieved the lowest FP value. Moreover, our dataset is one of the largest datasets in existing similar studies and we examined all different metastasis sizes. The higher PPV and sensitivity using our method implies that our approach may be able to 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 < 3mm is $50.6\%$, showing that around half the detected lesions are false positive. 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, which appearance on MR images is similar to small lesions. By increasing training data to
include more small lesions could potentially reduce the impact from imaging noise/artifacts or small brain vessels. In table 2, we have showed that a total of 21 lesions that initially was not marked as tumors or marked as suspect tumors were confirmed as tumors during follow-up visit. This shows the potentials of this tool in helping identify lesions in an early stage. However, due to different time span from treatment to follow-up imaging and giving the many false positive detection in small lesions, we are not able to draw a definitive conclusion that this tool can identify lesions in an early stage. Yet still, radiologists need to use their best judgment to determine whether this is a tumor or not by using this tool as an auxiliary tool to make their decision.

5. Conclusions

We have developed an automated framework to detect brain metastases in MR images. 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 early stages. A deep dive into FPs indicated that around 40% of them were true-positive metastases that were not contoured in ground-truth due to other medical reasons or indiscernible at the time of treatment but identified later by radiologist on subsequent follow-up imaging.

Declaration of interests
The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

D. Nana Yeboa reports financial support was provided by Brockman Foundation.
References


5. Patricia S. Steeg KAC and QRS. Brain metastases as preventive and therapeutic targets. Published online 2011.


Figure Legends

**Figure 1.** The workflow of building and evaluating our model. Abbreviations: FPs, false positives; BM-Net, metastases detector; MRI, magnetic resonance images; WB-Net whole-brain network.

![Workflow Diagram](attachment:workflow_diagram.png)

**Figure 2.**

A. (false-FP detection) The left image [A-1] shows our model-predicted tumors. The radiologist also marked them as tumors (right image [A-2]), but the physician did not contour these tumors for a medical reason.

B. (Early detection) The left image shows our contour [B-1], but the radiologist did not report a tumor at this location [B-2]. In the follow-up MRI 4 months later (right image [B-3]), the radiologist marked this location as a tumor.

C. (false-FP detection) The left image [C-1] shows the predicted contours, and the middle image [C-2] is the radiologist’s image showing a suspected tumor. However, in a follow-up image 4 months later (right image [C-3]), the radiologist indicated confidence of a tumor at this location.

D. (false-FP detection) Although our model correctly detected the metastasis in the left image [D-1], since the patient
had undergone whole-brain radiation therapy (WBRT), the treatment of the lesion was not required, and it was not contoured in the ground truth image [D-2].
Figure 3. Comparing box-whisker plot of positive predictive value (PPV) and sensitivity of the model in four size categories on a per-patient level
**Table 1.** Patient demographics and primary cancer types in the training and testing sets

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Testing set</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>845</td>
<td>206</td>
<td>1051</td>
</tr>
<tr>
<td>Mean age (years)</td>
<td>59.4</td>
<td>57.5</td>
<td>59.3</td>
</tr>
<tr>
<td>Gender (M, F)</td>
<td>406:439</td>
<td>101:105</td>
<td>507:544</td>
</tr>
<tr>
<td>Magnetic field (1.5 T, 3 T)</td>
<td>806:39</td>
<td>192:14</td>
<td>998:53</td>
</tr>
<tr>
<td>Mean size of metastases</td>
<td>10.1±7.4 mm</td>
<td>12.3±9.2 mm</td>
<td>10.5±7.8 mm</td>
</tr>
<tr>
<td>Mean metastases per patient</td>
<td>4.1±3.6</td>
<td>4.1±3.1</td>
<td>4.1±3.5</td>
</tr>
<tr>
<td>Number of metastases &lt; 3 mm</td>
<td>78</td>
<td>60</td>
<td>138</td>
</tr>
<tr>
<td>Number of metastases ≥ 3 mm, &lt; 6 mm</td>
<td>1267</td>
<td>324</td>
<td>1591</td>
</tr>
<tr>
<td>Number of metastases ≥ 6 mm</td>
<td>2137</td>
<td>546</td>
<td>2683</td>
</tr>
<tr>
<td>Primary cancer type (No., %)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-small cell lung</td>
<td>349 (41.3%)</td>
<td>61 (29.6%)</td>
<td>410 (39%)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>166 (19.6%)</td>
<td>53 (25.7%)</td>
<td>219 (20.8%)</td>
</tr>
<tr>
<td>Breast</td>
<td>117 (13.8%)</td>
<td>38 (18.4%)</td>
<td>155 (14.7%)</td>
</tr>
<tr>
<td>Renal</td>
<td>87 (10.3%)</td>
<td>21 (10.4%)</td>
<td>108 (10.3%)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>30 (3.6%)</td>
<td>9 (4.4%)</td>
<td>39 (3.7%)</td>
</tr>
<tr>
<td>Head and neck</td>
<td>20 (2.4%)</td>
<td>4 (1.9%)</td>
<td>24 (2.3%)</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>19 (2.2%)</td>
<td>7 (3.4%)</td>
<td>26 (2.5%)</td>
</tr>
<tr>
<td>Thyroid</td>
<td>18 (2.1%)</td>
<td>3 (1.5%)</td>
<td>21 (2%)</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>14 (1.7%)</td>
<td>4 (1.9%)</td>
<td>18 (1.7%)</td>
</tr>
<tr>
<td>Small cell lung</td>
<td>13 (1.5%)</td>
<td>2 (1%)</td>
<td>15 (1.4%)</td>
</tr>
<tr>
<td>Neuroendocrine carcinoma</td>
<td>8 (0.9%)</td>
<td>3 (1.5%)</td>
<td>11 (1%)</td>
</tr>
<tr>
<td>Thymic</td>
<td>3 (0.4%)</td>
<td>0 (0.0%)</td>
<td>3 (0.3%)</td>
</tr>
<tr>
<td>Non-melanoma skin</td>
<td>1 (0.1%)</td>
<td>1 (0.5%)</td>
<td>2 (0.2%)</td>
</tr>
</tbody>
</table>
Table 2. Investigation of 164 false positive cases using radiology report and follow-up MRIs.

The word “Unknown” indicated that follow-up MRI had not been checked due to the confirmation of a tumor on the radiology report.

<table>
<thead>
<tr>
<th>Group</th>
<th>No. Mets</th>
<th>Radiology report</th>
<th>Ground truth (Treatment)</th>
<th>Follow-up Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
<td>Not tumor</td>
<td>Not contoured</td>
<td>Not tumor</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>Tumor</td>
<td>Unknown</td>
<td>Tumor</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>Not tumor</td>
<td>True positive/early detection</td>
<td>True positive</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>Suspect tumor</td>
<td>Tumor</td>
<td>True positive</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>Tumor (identified for whole-brain treatment)</td>
<td>Unknown</td>
<td>True positive</td>
</tr>
</tbody>
</table>
Table 3. The model metrics on a per-patient level in the four size categories. PPV: positive predictive value; DSC: Dice similarity coefficient; TP: true positive; FP: false positive; FN: false negative.

<table>
<thead>
<tr>
<th>Metric</th>
<th>&lt; 3 mm</th>
<th>≥ 3 mm, &lt; 6 mm</th>
<th>≥ 6 mm</th>
<th>All sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>per-patient level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPV (%)</td>
<td>85.0 ± 34.9</td>
<td>90.7 ± 26.2</td>
<td>95.6 ± 17.5</td>
<td>90.1 ± 17.7</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>94.1 ± 23.0</td>
<td>83.4 ± 30.8</td>
<td>94.5 ± 18.1</td>
<td>88.4 ± 18.0</td>
</tr>
<tr>
<td>DSC (%)</td>
<td>31.2 ± 14.5</td>
<td>72.0 ± 10.1</td>
<td>86.8 ± 7.5</td>
<td>82.2 ± 9.5</td>
</tr>
<tr>
<td>TP</td>
<td>0.2 ± 0.6</td>
<td>1.1 ± 1.6</td>
<td>2.5 ± 1.8</td>
<td>4.0 ± 3.1</td>
</tr>
<tr>
<td>FP</td>
<td>0.2 ± 0.5</td>
<td>0.2 ± 0.6</td>
<td>0.1 ± 0.4</td>
<td>0.4 ± 1.0</td>
</tr>
<tr>
<td>FN</td>
<td>0.1 ± 0.4</td>
<td>0.5 ± 0.9</td>
<td>0.1 ± 0.4</td>
<td>0.7 ± 1.2</td>
</tr>
<tr>
<td><strong>per-metastasis level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPV (%)</td>
<td>50.6</td>
<td>85.9</td>
<td>96.9</td>
<td>89.4</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>68.3</td>
<td>69.8</td>
<td>95.8</td>
<td>84.9</td>
</tr>
<tr>
<td>DSC (%)</td>
<td>30.7 ± 16.7</td>
<td>71.0 ± 12.2</td>
<td>86.5 ± 9.8</td>
<td>80.4 ± 15.9</td>
</tr>
<tr>
<td>TP</td>
<td>41</td>
<td>226</td>
<td>523</td>
<td>790</td>
</tr>
<tr>
<td>FP</td>
<td>40</td>
<td>37</td>
<td>17</td>
<td>94</td>
</tr>
<tr>
<td>FN</td>
<td>19</td>
<td>98</td>
<td>23</td>
<td>140</td>
</tr>
</tbody>
</table>