HHS Public Access

Author manuscript

J Magn Reson Imaging. Author manuscript; available in PMC 2020 May 05.

Published in final edited form as:

J Magn Reson Imaging. 2020 January ; 51(1): 175–182. doi:10.1002/jmri.26766.

Deep Learning Enables Automatic Detection and Segmentation of Brain Metastases on Multisequence MRI

Endre Grøvik, PhD[1,2], **Darvin Yi, MS**[3], **Michael Iv, MD**[1], **Elizabeth Tong, MD**[1], **Daniel Rubin, PhD**[#3], **Greg Zaharchuk, MD, PhD**[#1,*]

[1]Department of Radiology, Stanford University, Stanford, California, USA

[2]Department for Diagnostic Physics, Oslo University Hospital, Oslo, Norway

[3]Department of Biomedical Data Science, Stanford University, Stanford, California, USA

[#] These authors contributed equally to this work.

Abstract

Background: Detecting and segmenting brain metastases is a tedious and time-consuming task for many radiologists, particularly with the growing use of multisequence 3D imaging.

Purpose: To demonstrate automated detection and segmentation of brain metastases on multisequence MRI using a deep-learning approach based on a fully convolution neural network (CNN).

Study Type: Retrospective.

Population: In all, 156 patients with brain metastases from several primary cancers were included.

Field Strength: 1.5T and 3T. [Correction added on May 24, 2019, after first online publication: In the preceding sentence, the first field strength listed was corrected.]

Sequence: Pretherapy MR images included pre- and postgadolinium T_1-weighted 3D fast spin echo (CUBE), postgadolinium T_1-weighted 3D axial IR-prepped FSPGR (BRAVO), and 3D CUBE fluid attenuated inversion recovery (FLAIR).

Assessment: The ground truth was established by manual delineation by two experienced neuroradiologists. CNN training/development was performed using 100 and 5 patients, respectively, with a 2.5D network based on a GoogLeNet architecture. The results were evaluated in 51 patients, equally separated into those with few (1–3), multiple (4–10), and many (>10) lesions.

Statistical Tests: Network performance was evaluated using precision, recall, Dice/F1 score, and receiver operating characteristic (ROC) curve statistics. For an optimal probability threshold, detection and segmentation performance was assessed on a per-metastasis basis. The Wilcoxon rank sum test was used to test the differences between patient subgroups.

[*]Address reprint requests to: G.Z., Department of Radiology, Stanford University, School of Medicine, 1201 Welch Road, Stanford, CA 94305-5488. gregz@stanford.edu.
E.G and D.Y are Co-First authors.

Results: The area under the ROC curve (AUC), averaged across all patients, was 0.98 ± 0.04. The AUC in the subgroups was 0.99 ± 0.01, 0.97 ± 0.05, and 0.97 ± 0.03 for patients having 1–3, 4–10, and >10 metastases, respectively. Using an average optimal probability threshold determined by the development set, precision, recall, and Dice score were 0.79 ± 0.20, 0.53 ± 0.22, and 0.79 ± 0.12, respectively. At the same probability threshold, the network showed an average false-positive rate of 8.3/patient (no lesion-size limit) and 3.4/patient (10 mm^3 lesion size limit).

Data Conclusion: A deep-learning approach using multisequence MRI can automatically detect and segment brain metastases with high accuracy.

Level of Evidence: 3

Technical Efficacy Stage: 2

ATTRIBUTED IN LARGE to advances in effective systemic treatment regimens of primary tumors, there has been an increase in the number of patients with metastatic cancer over the last decade.[1] Brain metastases are one of the most common neurologic complications of cancer, most frequently originating from lung cancer, breast cancer, and malignant melanoma.[2] In a survey including more than 26,000 patients, 12.1% of all patients with metastatic disease had brain metastases at diagnosis.[3] Most patients present with three or fewer metastases to the brain, but 40% of patients have greater than this number. [4,5] Contrast-enhanced magnetic resonance imaging (MRI) is the key imaging technique in the diagnosis of brain metastases and is also used for longitudinal follow-up to assess treatment response.

Delineation of initial tumor volume and volume change in relation to disease progression or therapy are key neuroradiological tasks as part of optimal patient management. Given its importance and high demand for accuracy, manual detection and segmentation of brain tumors is a tedious and time-consuming task, particularly with the growing use of multisequence 3D imaging. Furthermore, the diagnostic methods for assessing treatment response follow the criteria formulated by the Response Assessment in Neuro-Oncology (RANO) working group and are based on measuring the size of the enhancing lesion on gadolinium (Gd)-enhanced T_1-weighted MR images.[6] The traditional metrics used for response evaluation are based on unidimensional measurements, although the value of using volumetric measurements has been increasingly discussed. One concern raised by the RANO group was that volumetric analysis, as performed manually by radiologists, adds cost and complexity and is not available at all centers.

During recent years, advances in machine learning (ML) have suggested the possibility of new paradigms in healthcare. One application of ML in radiology is the detection and segmentation of organs and pathology.[7–10] In particular, there has been a significant effort in developing deep learning (DL) algorithms to learn from the comprehensive voxelwise labeled MRI data for segmenting primary brain tumors.[11–15] However, only a few studies have applied such ML approaches on patients with brain metastases,[16–18] which may require different approaches given their size and multiplicity. To this end, the aim of this work was to develop and assess a fully convolution neural network (CNN) for automatic detection and segmentation of brain metastases using multisequence MRI data as input. Whereas comparable studies in the literature use homogeneous patient cohorts, ie, a single field-

strength/vendor/scanner, or limited to patients receiving stereotactic radiosurgery (SRS), this study work included a heterogeneous cohort of clinical patients not limited to SRS planning, facilitating subgroup analysis based on the total number of brain metastases and their sizes, thus challenging the generalizability of the proposed neural network.

Materials and Methods

Patient Population

This retrospective, single-center study was approved by our Institutional Review Board. Inclusion criteria included the presence of known or possible metastatic disease (ie, presence of a primary tumor), no prior surgical or radiation therapy, and the availability of all required MRI sequences (see below). Only patients with ≥1 metastatic lesion were included. Mild patient motion was not an exclusion criterion. Based on these criteria, a consecutive set of 156 patients was identified, imaged between June 2016 and June 2018, and were included in the study. Details of this cohort are shown in Table 1. Imaging was performed with both 1.5T ($n = 18$; SIGNA Explorer and TwinSpeed, GE Healthcare, Chicago, IL) and 3T ($n = 138$; Discovery 750 and 750w and SIGNA Architect, GE Healthcare; Skyra, Siemens Healthineers, Erlangen, Germany) clinical scanners. Mean patient age was 63 ± 12 years (range: 29–92 years). Primary malignancies included lung ($n = 99$), breast ($n = 33$), melanoma ($n = 7$), genitourinary ($n = 7$), gastrointestinal ($n = 5$), and miscellaneous cancers ($n = 5$). Of the 156 patients included, 64 (41 %) had 1–3 metastases, 47 (30%) had 4–10 metastases, and 45 (29%) had >10 metastases. Lesion sizes varied from 2 mm to over 4 cm and were scattered in every region of the brain parenchyma, ie, the supratentorial and infratentorial region, as well as the cortical and subcortical structures.

Imaging Protocol

The imaging protocol included pre- and post-Gd T_1-weighted 3D fast spin echo (CUBE), post-Gd T_1-weighted 3D axial IR-prepped FSPGR (BRAVO), and 3D CUBE fluid-attenuated inversion recovery (FLAIR). All sequences with key imaging parameters are summarized in Table 2. For Gd-enhanced imaging, a dose of 0.1 mmol/kg body weight of gadobenate dimeglumine (MultiHance, Bracco Diagnostics, Princeton, NJ) was intravenously administered.

Image Segmentation and Coregistration

Ground truth segmentations were established by two neuroradiologists with 8 (M.I.) and 2 (E.T.) years of experience by manually delineating and cross-checking regions of interest (ROIs) around each enhancing metastatic lesion. The lesions were outlined on each slice on the post-Gd 3D T_1-weighted IR-FSPGR sequence, with additional guidance from the 3D FLAIR and the post-Gd 3D T_1-weighted spin echo data using the OsiriX MD software package (v. 8.0, Geneva, Switzerland).

Pre/postcontrast T_1 CUBE and FLAIR images were coregistered to the IR-FSPGR space by normalized mutual information coregistration using the nordicICE software package (NordicNeuroLab, Bergen, Norway). Prior to network training, the brain was extracted by using the Brain Extraction Tool (BET)[19] and applying the resulting brain masks on the

network's input data. The brain masks were generated from the precontrast T_1-weighted 3D CUBE imaging series and propagated to the other sequences.

CNN Details

Training was performed using a 2.5D fully CNN based on the GoogLeNet architecture[20] (Fig. 1). The network was modified to optimize segmentation by skipping the first and third downsampling max pooling layers and using a stride of one on the first 7 × 7 convolutional layer. As a result, the final downsampling rate throughout the convolutional layers was 4×, rather than 32×. To make the network fully convolutional, GoogLeNet's final fully connected layers were replaced by a single convolutional transpose layer of stride 4 and size 8 × 8. The final prediction was made on a single channel of logit values with a sigmoid cross-entropy loss function. In order to counter learning hurdles introduced by an unbalanced dataset, the loss on positive ground truth voxels were weighted 10× more than the loss on negative ground truth voxels.

To better capture through-plane features without incurring the inefficiencies associated with true 3D CNN's, we implemented a "2.5D" model. The network's input were seven slices from each of the four aforementioned sequences, comprising a single center slice with three slices above and below, resulting in an input channel dimension of 28. Each image was rescaled (if necessary) to a size of 256 × 256. Note that all MR images were originally 256 × 256 or 512 × 512, and that bilinear interpolating the 512 × 512 images down to 256 × 256 gave minimal artifacts. Prior to training, preprocessing and normalization was performed with independent histogram equalization on each slice of the 28-channel input. During training, we randomly flipped and rotated the images in multiples of 90° for data augmentation.

All training was performed on two consumer-grade graphical processing units (GPUs) (NVIDIA GeForce GTX 1080TI). The batch size was 32 with a learning rate of 0.001. Given that there were far more frames (>30×) without lesions compared with frames with lesions, we employed an uneven sampling procedure. For 16 of 32 images in each batch, we sampled the image from the set of frames with at least some lesion. For the other 16 images, we sampled uniform randomly from all frames. This ensured that for each batch at least half of the images was populated with frames including some ground truth lesions. Regularization was performed by an L2 weight decay with a decay constant of $1e^{-5}$. Batch normalization was used following every convolutional layer. We used the ADAM optimization method[21] with default TensorFlow beta values of 0.9 and 0.999. By defining an epoch as the statistical equivalent of seeing every distinct frame of the dataset once, the training continues until convergence, which occurred at about the 10^{th} epoch. The network was trained using TensorFlow, and the resulting output was an image for each slice representing a probability map of whether the voxel represents a metastasis, ranging from 0–1.

The total number of cases were randomly broken into separate train, development, and test sets. None of the cases in the test set were present in the training set. To ensure a representative sample in our test set, we first chose the test cases as follows. First, we determined the number of distinct metastatic lesions in each case in the entire cohort and

then broke the data into groups with (a) 1–3, (b) 4–10, (c) >10 lesions. We then randomly selected 17 cases from each of these groups, leading to a total test set size of 51 cases. The remaining cases were divided into training and development sets in a random 20:1 ratio, giving a final breakdown of 100 training cases, five development cases, and 51 test cases. The test set had a total of856 lesions.

Statistical Analysis

The network's ability to detect metastases on a voxel-by-voxel basis was evaluated using receiver operating characteristic (ROC) curve statistics, measuring the area under the ROC curve (AUC) for each patient in the test set. Only voxels within the brain mask were considered when calculating AUC. Corresponding sensitivity and specificity were determined by using the maximum value of Youden's index as a criterion for selecting the optimal cutoff point. Based on ROC statistics from the development set, the optimal probability threshold for including a voxel as a metastasis was determined, and using this threshold the results were further evaluated in terms of detection accuracy using precision and recall, and segmentation accuracy using the Dice similarity score (also known as the F1 score). In addition to the voxel-by-voxel analysis, the detection performance was also evaluated on a lesion-by-lesion basis by calculating the number of false positives (FPs) per case. These metrics were determined by comparing the ground truth maps and the probability maps, counting the number of overlapping objects using a connecting component approach. The number of FPs were determined both without and with a lesion-size criterion, in which only objects ≥ 10 mm^3 were considered a detected lesion. The detection performance was also investigated as a function of lesion size. This was done by comparing the 3D connective components estimated from the ground truth maps and the probability maps. The latter was done using a probability threshold of 0.1. If there was more than a 10% overlap between the predicted lesion and the ground truth, the networks prediction was labeled "found." If no such overlap existed, the networks prediction was labeled "missed." Furthermore, given that the size threshold for differentiating large and small lesions is not fixed, the detection sensitivity was estimated as a function this size threshold. The lesion size was determined by the longest 3D diameter. Finally, the Wilcoxon rank sum test was used to compare the detection and segmentation metrics between the patient subgroups. A statistical significance level of 5% was used. All statistical analyses were performed using MatLab R2017av. 9.2.0 (MathWorks, Natick, MA).

Results

The total time for training the neural network was ~15 hours. For processing a test case, the forward pass on a single NVIDIA GTX 1080Ti GPU took less than 200 msec per slice during the test time with a run time of ~1 minute for a full MR volume.

Figure 2 shows an example case demonstrating the resulting probability map as an overlay on the post-Gd FSPGR image series using a lower probability threshold of 0.1. The voxel-by-voxel detection performance showed an area under the ROC-curve, averaged across all patients, of 0.98 ± 0.04, corresponding to a sensitivity and specificity of 94% and 97%, respectively, at the optimal cutoff point. Further, the subgroups showed an area under the

ROC curve of 0.99 ± 0.01 for patients having 1–3 metastases, 0.97 ± 0.05 for 4–10 metastases, and 0.97 ± 0.03 for >10 metastases (Fig. 3). The corresponding sensitivity and specificity are shown in Table 3A. The average optimal probability threshold for including a voxel as a metastasis, measured in the development set, was 0.93. Using this threshold, the precision, recall, and Dice score were 0.79 ± 0.20, 0.53 ± 0.22, and 0.79 ± 0.12, respectively. The distribution of these metrics within the subgroups is shown Table 3B. On a lesion-by-lesion basis, and by using the optimal probability threshold (average sensitivity = 83%), the network showed an average FP rate of 8.3 (no size limit) and 3.4 (10 mm^3 size limit) lesions per case, with the highest sensitivity and lowest numbers of FP in patients with few metastases (Table 3C). The *P*-values, testing the differences in all detection and segmentation metrics between the subgroups, are shown in Table 4A–C. Examples in representative cases with different numbers of metastases are shown in Fig. 4.

The network's ability to detect brain metastases was associated with the lesion size (Fig. 5). For lesions smaller that 7 mm, the neural network showed a sensitivity of 50%. However, for all lesions larger that 22 mm the network achieved a sensitivity of 100%. Figure 6 show the network's sensitivity as a function of size threshold for differentiating large and small lesions. If the size threshold is set to 22 mm, the network's sensitivity was 80% for detecting <22 mm lesions and ~100% for detecting >22 mm lesions.

Discussion

This study demonstrated that a modified 2.5D GoogLeNet CNN can detect and segment brain metastases on multisequence MRI with high accuracy. By testing on a large number of patients, thus facilitating subgroup analysis, this work demonstrates the network's clinical performance and potential, in addition to better understanding of its generalizability. To our knowledge, no previous study has reported on subgroup analysis using deep learning in brain metastases segmentation.

In recent years, many DL approaches have been developed and tested for automatic segmentation of gliomas,[22] thanks in part to the publicly available BRAin Tumor Segmentation (BraTS) dataset. In contrast, only a few studies have used this approach for brain metastases. Liu et al investigated the use of a CNN-based segmentation for SRS planning and reported a Dice score of 0.67 and an AUC of 0.98.[16] The performance of the current method is superior to this prior method based on the average Dice score (0.79 vs. 0.67), but showed similar AUC performance. Charron et al used a 3D CNN (DeepMedic) for automatic detection and segmentation of brain metastases.[17] By using segmented metastases to be irradiated as the ground truth, their network was trained using three MRI sequences from 146 patients as input and further tested on 18 patients. Similar to our study, their network was trained using three MRI sequences that proved to outperform networks trained on a single MRI contrast. Their network showed a sensitivity of 98% and 7.2 FP per patient. Sunwoo et al developed a computer-aided diagnostic (CAD) system for detecting brain metastases and a neural network for FP reduction.[23] Their CAD system significantly improved the diagnostic performance of the reviewers and showed an overall sensitivity of 87%. One feature that separates our work from these previous studies is the strength of having diverse data, which may make it more challenging to demonstrate an overall high

performance. We included cases from both 1.5T and 3T using multivendor scanners, and our data were not limited to patients receiving SRS, thus including more patients with extensive metastases disease (>10). This is supported by our results, which indicate that the neural networks ability to detect and segment brain metastases were reduced in patients with a higher number of metastases. However, our patient cohort is more representative of real-world data. Furthermore, there are also differences in network architecture. Whereas Charron et al[17] used a full 3D network, we used a 2.5D network. The results indicate that our 2.5D network achieves the same segmentation performance as a 3D network, which reduces the computational and memory requirements for training. However, further studies systematically comparing 2D, 2.5D, and 3D neural architectures must be performed to adequately answer this question. Also note that, using comparable graphic cards, our 2.5D network required ~1 minute to perform a forward-pass (inference), while a 3D network would require ~20 minutes for the same task.

The split between training/validation/testing in this study is somewhat unusual compared with similar studies in the literature. However, we chose to test on a large number of cases to understand how generalizable the network was and to facilitate subgroup analysis, enabling a better understanding of the network's clinical performance and potential. In earlier stages of the study, we found that the network had high performance training on approximately half of the current training set, and that increasing the number of training cases did not provide significant improvement, thus justifying our use of a larger test set. Our results indicate that the networks ability to detect metastatic voxels, as measured by the AUC, is best in patients with few (1–3) metastases and further that the segmentation performance, as measured by the Dice score, is slightly better for patients with 4–10 metastases. On a lesion-by-lesion basis, our results suggest that the network performs best on patients with few metastases, both in terms of sensitivity and the number of FPs. One hypothesis is that these results may be associated with an optimal tradeoff between total number and individual size of the metastases. This is supported by our result suggesting that the network has a higher sensitivity for detecting large lesions. Larger lesions are often less subtle and may exhibit more textural features that may make them easier to detect. However, through the multiple layers of downsampling with higher order strides and pooling layers, the network may have a harder time distinguishing lesions without enough pixel information. It should also be noted that in training segmentation networks, a large lesion comprises many more voxels than a smaller lesion, which may cause a data imbalance problem.

Multiple network architectures were considered for this project, including residual networks,[24] dense networks,[25] U-Nets,[26] Pyramid Scene Parsing (PSP) Nets,[27] and Feature Pyramid Networks (FPNs).[28] However, after running preliminary experiments with the 2014 GoogLe (or Inception v1) network, we found that this was already capable of overfitting the training data. Thus, given that network complexity and capacity were not driving issues in the project, and that the GoogLeNet enables high computational efficiency, both in memory and speed, this became our choice of architecture. The compact size of the network allows it to be run on even the smallest mobile GPUs, such as the NVIDIA Tegra chip. However, note that more ample gains from advanced network architectures could be attained with larger datasets.

Typical workflow in radiotherapy planning requires accurate detection by a radiologist, followed by segmentation by a radiation oncologist. Both steps are time-consuming and subject to interobserver variation. Detection requires manual visualization and annotation. Fatigue and image quality are a few factors that may affect the accuracy.[29,30] Special imaging techniques have been proposed to improve this process. For instance, double-dose Gd-based contrast-enhanced thin-slice MRI produced more precise delineation of lesions compared with using a single dose.[31] The addition of overlapping CUBE maximum intensity projection (MIP) images, which have a better contrast-to-noise ratio of metastatic lesions than post-Gd 3D isometric high-resolution sequences, are often used to enhance the sensitivity of detection. However, even with the addition of CUBE MIP images, the interrater agreement for identification of metastases between two experienced radiologists was reported as only fair-to-moderate in one study.[32] Segmentation requires tracing the contours of the lesions on the 2D images slice-by-slice. Even though there is semiautomatic software available for segmentation, extensive manual editing is often required, thus generating nonreproducible operator-dependent results.[33,34] Accurate segmentation of the metastases is imperative in radiation therapy planning to minimize damage to adjacent normal tissue. Our neural network essentially combines visualization, quantification, and segmentation into one fluid step, producing results that can be directly applied to radiotherapy planning, with minimal user interaction.

While this study shows high accuracy and performance using DL for segmenting brain metastasis, several potential study limitations exist. First, the results must be interpreted in light of the limited sample size in this single-center, retrospective study. This is partly related to the time required for manual segmentation. Future studies will investigate the use of "coarse" segmentation, which is by far less time-consuming compared with fine segmentation, and how this may affect the network's ability to detect and segment brain metastases. Also, testing of the network performance on multisite data remains a key step towards understanding its clinical value. Second, the network sometimes fails in terms of reporting FP. This is particularly true in and near vascular structures at the skull base such as venous sinuses, or over the cortex. Finally, as our neural network is trained on four distinct MRI contrasts, the use of this method is limited to sites acquiring all sequences. However, future studies will address the issue of having other or even lacking model inputs with the aim of making the neural network more robust and versatile towards different input channels.

In conclusion, our study shows that a deep learning network can automatically detect and segment brain metastases on multisequence MRI with high accuracy and thus illustrates the potential use of this technique in a clinically relevant setting.

Acknowledgment

Contract grant sponsor: South-Eastern Norway Regional Health Authority; Contract grant numbers: 2016102 and 2013069; Contract grant sponsor: Research Council of Norway; Contract grant number: 261984; Contract grant sponsor: Norwegian Cancer Society; Contract grant numbers: 6817564 and 3434180.

References

1. Arvold ND, Lee EQ, Mehta MP, et al. Updates in the management of brain metastases. Neuro Oncol 2016;18:1043–1065. [PubMed: 27382120]
2. Nayak L, Lee EQ, Wen PY. Epidemiology of brain metastases. Curr Oncol Rep 2012;14:48–54. [PubMed: 22012633]
3. Cagney DN, Martin AM, Catalano PJ, et al. Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: A population-based study. Neuro Oncol 2017;19:1511–1521. [PubMed: 28444227]
4. Fabi A, Felici A, Metro G, et al. Brain metastases from solid tumors: Disease outcome according to type of treatment and therapeutic resources of the treating center. J Exp Clin Cancer Res 2011;30:10. [PubMed: 21244695]
5. Nussbaum ES, Djalilian HR, Cho KH, Hall WA. Brain metastases: Histology, multiplicity, surgery, and survival. Cancer 1996;78:1781–1788. [PubMed: 8859192]
6. Lin NU, Lee EQ, Aoyama H, et al. Response assessment criteria for brain metastases: Proposal from the RANO group. Lancet Oncol 2015; 16:e270–e278. [PubMed: 26065612]
7. Goceri E, Songul C. Computer-based segmentation, change detection and quantification for lesions in multiple sclerosis. In: IEEE Int Conf Comput Sci Eng; 2017:177–182.
8. Moghbel M, Mashohor S, Mahmud R, Bin Saripan MI. Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. Artif Intell Rev 2017;1–41.
9. Grossiord E, Talbot H, Passat N, Meignan M, Najman L. Automated 3D lymphoma lesion segmentation from PET/CT characteristics. In: IEEE 14th Int Symp Biomed Imaging; 2017:174–178.
10. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: State of the art and future directions. J Digit Imaging 2017;30:449–459. [PubMed: 28577131]
11. Havaei M, Guizard N, Larochelle H, Jodoin P-M. Deep learning trends for focal brain pathology segmentation in MRI. Mach Learn Heal Inform 2016;125–148.
12. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 2016;35:1240–1251. [PubMed: 26960222]
13. Zhao L, Jia K. Deep feature learning with discrimination mechanism for brain tumor segmentation and diagnosis. In: IEEE Int Conf Intell Inf Hiding Multimed Signal Process; 2015:306–309.
14. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78. [PubMed: 27865153]
15. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med Image Anal 2018;43:98–111. [PubMed: 29040911]
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. [PubMed: 28985229]
17. 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. [PubMed: 29455079]
18. Liu Y, Stojadinovic S, Hrycushko B, et al. Automatic metastatic brain tumor segmentation for stereotactic radiosurgery applications. Phys Med Biol 2016;61:8440–8461. [PubMed: 27845915]
19. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002;17:143–155. [PubMed: 12391568]
20. Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions. In: IEEE Conf Comput Vis Pattern Recognit; 2015:1–9.
21. Kingma DP, Ba JA. A method for stochastic optimization. arXiv Prepr 2014:arXiv; 1412.6980.
22. Isýn A, Direkoglu C, Sah M. Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Comput Sci 2016;102:317–324.

23. Sunwoo L, Kim YJ, Choi SH, et al. Computer-aided detection of brain metastasis on 3D MR imaging: Observer performance study. PLoS One 2017;12:e0178265. [PubMed: 28594923]
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Conf Comput Vis Pattern Recognit; 2016:770–778.
25. Huang G, Liu Z, Maaten L van der, Weinberger KQ. Densely connected convolutional networks. In: IEEE Conf Comput Vis Pattern Recognit; 2017:2261–2269.
26. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Int Conf Med Image Comput Comput Interv; 2015:234–241.
27. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: IEEE Conf Comput Vis Pattern Recognit; 2017:6230–6239.
28. Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: IEEE Conf Comput Vis Pattern Recognit; 2017:936–944.
29. Ramkumar A, Dolz J, Kirisli HA, et al. User interaction in semi-automatic segmentation of organs at risk: A case study in radiotherapy. J Digit Imaging 2016;29:264–277. [PubMed: 26553109]
30. Park SH, Gao Y, Shi Y, Shen D. Interactive prostate segmentation using atlas-guided semi-supervised learning and adaptive feature selection. Med Phys 2014;41:111715. [PubMed: 25370629]
31. Subedi KS, Takahashi T, Yamano T, et al. Usefulness of double dose contrast-enhanced magnetic resonance imaging for clear delineation of gross tumor volume in stereotactic radiotherapy treatment planning of metastatic brain tumors: A dose comparison study. J Radiat Res 2013; 54:135–139. [PubMed: 22843378]
32. Yoon BC, Saad AF, Rezaii P, Wintermark M, Zaharchuk G, Iv M. Evaluation of thick-slab overlapping MIP images of contrast-enhanced 3D T1-weighted CUBE for detection of intracranial metastases: A pilot study for comparison of lesion detection, interpretation time, and sensitivity with nonoverlapping CUBE MIP, CUBE, and inversion-recovery-prepared fast-spoiled gradient recalled brain volume. Am J Neuroradiol 2018;39: 1635–1642. [PubMed: 30093483]
33. Whitfield GA, Price P, Price GJ, Moore CJ. Automated delineation of radiotherapy volumes: Are we going in the right direction? Br J Radiol 2013;86:20110718. [PubMed: 23239689]
34. Heckel F, Moltz JH, Tietjen C, Hahn HK. Sketch-based editing tools for tumour segmentation in 3D medical images. Comput Graph Forum 2013;32:144–157.

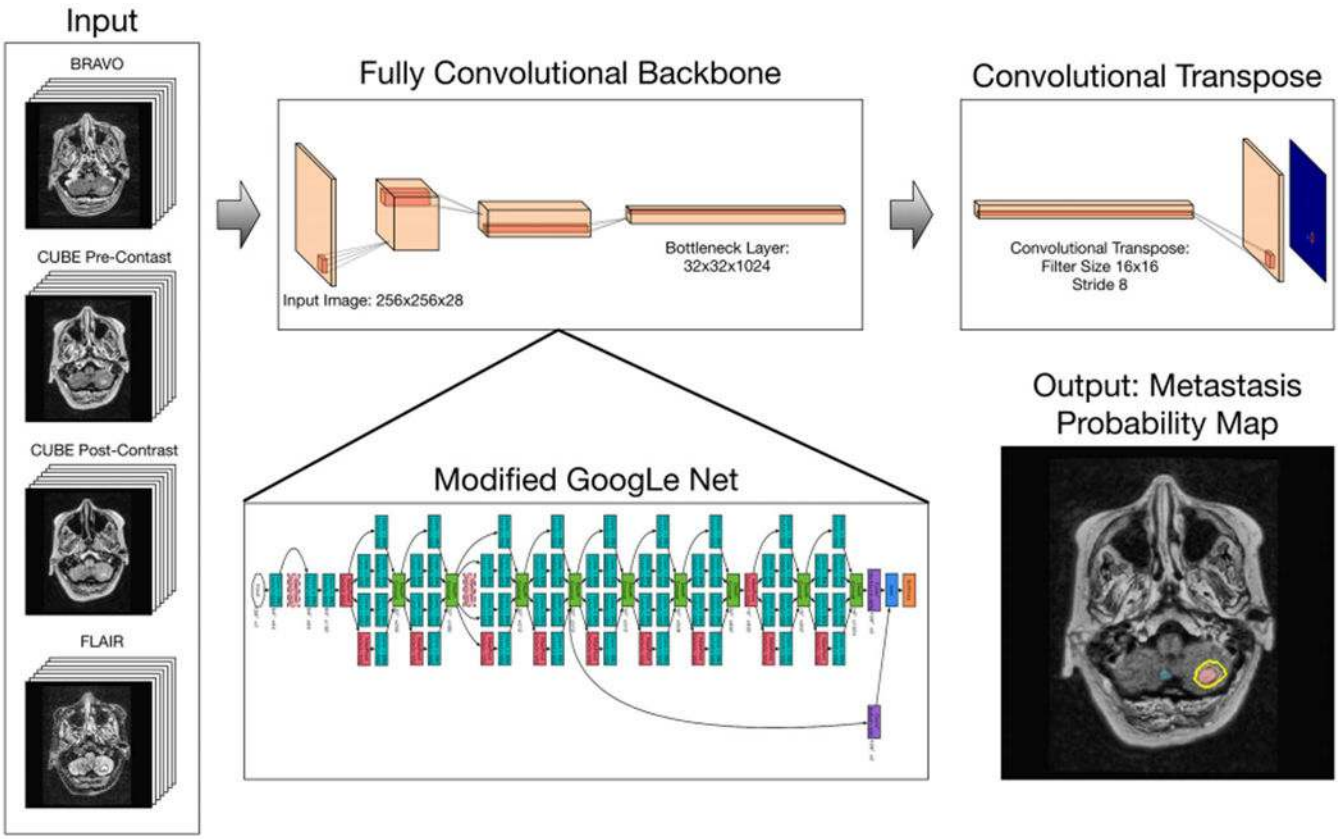

FIGURE 1:
Flowchart showing the four image inputs used to train the neural network, the modified GoogLeNet architecture, and the resulting output color map (overlaid on a postcontrast BRAVO image) representing a probability map of whether the voxel represents a metastasis, ranging from 0–1.

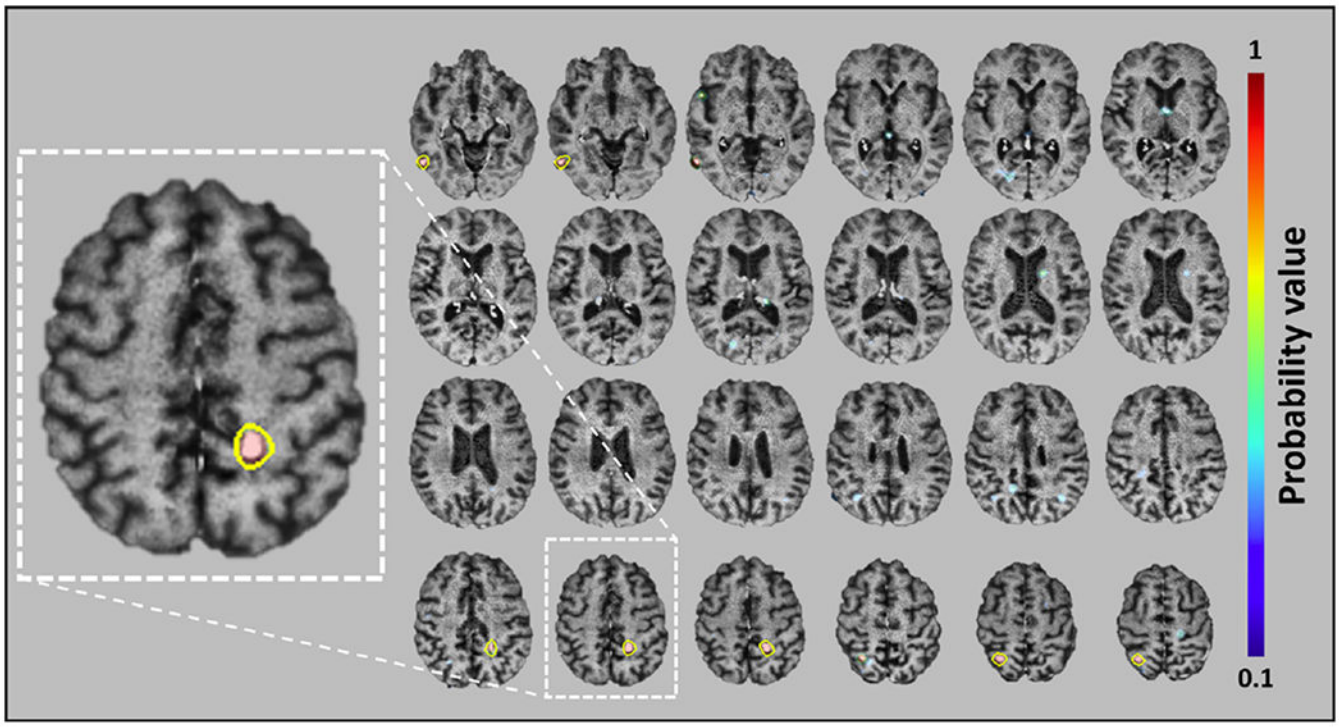

FIGURE 2:
Example case of a 47-year-old female patient presenting with three brain metastases from lung cancer. The image mosaic shows the predictions (probability maps as indicated by the color bar), generated by the neural network, and manually delineated metastases (yellow lines) overlaid on the postcontrast image.

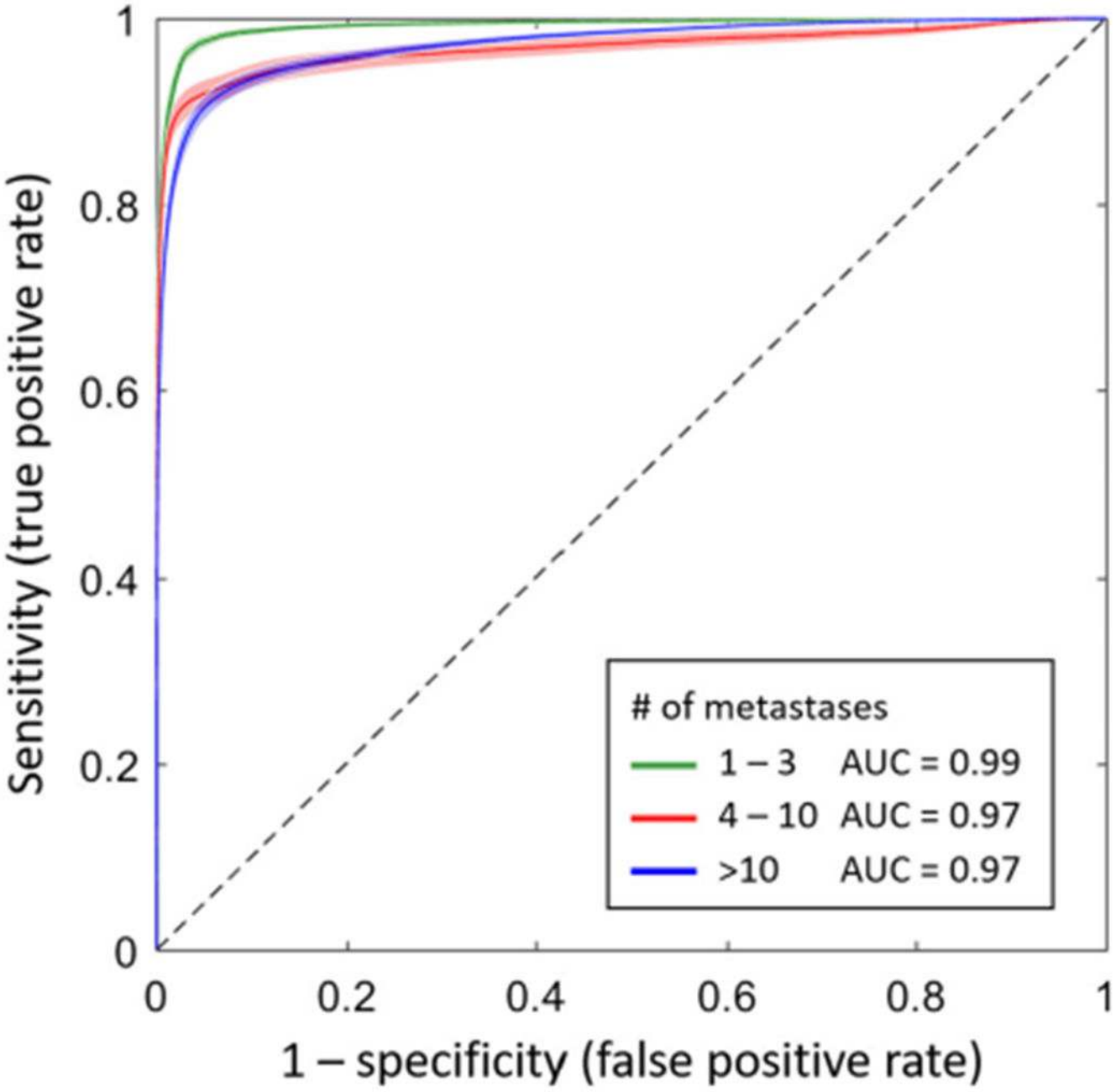

FIGURE 3:
ROC curves with 95% confidence intervals (shaded areas) for the three subgroups having 1–3 metastases (green), 4–10 metastases (red), and >10 metastases (blue). The average area under the ROC curve was 0.98, ranging from 0.79–1.00 for all cases.

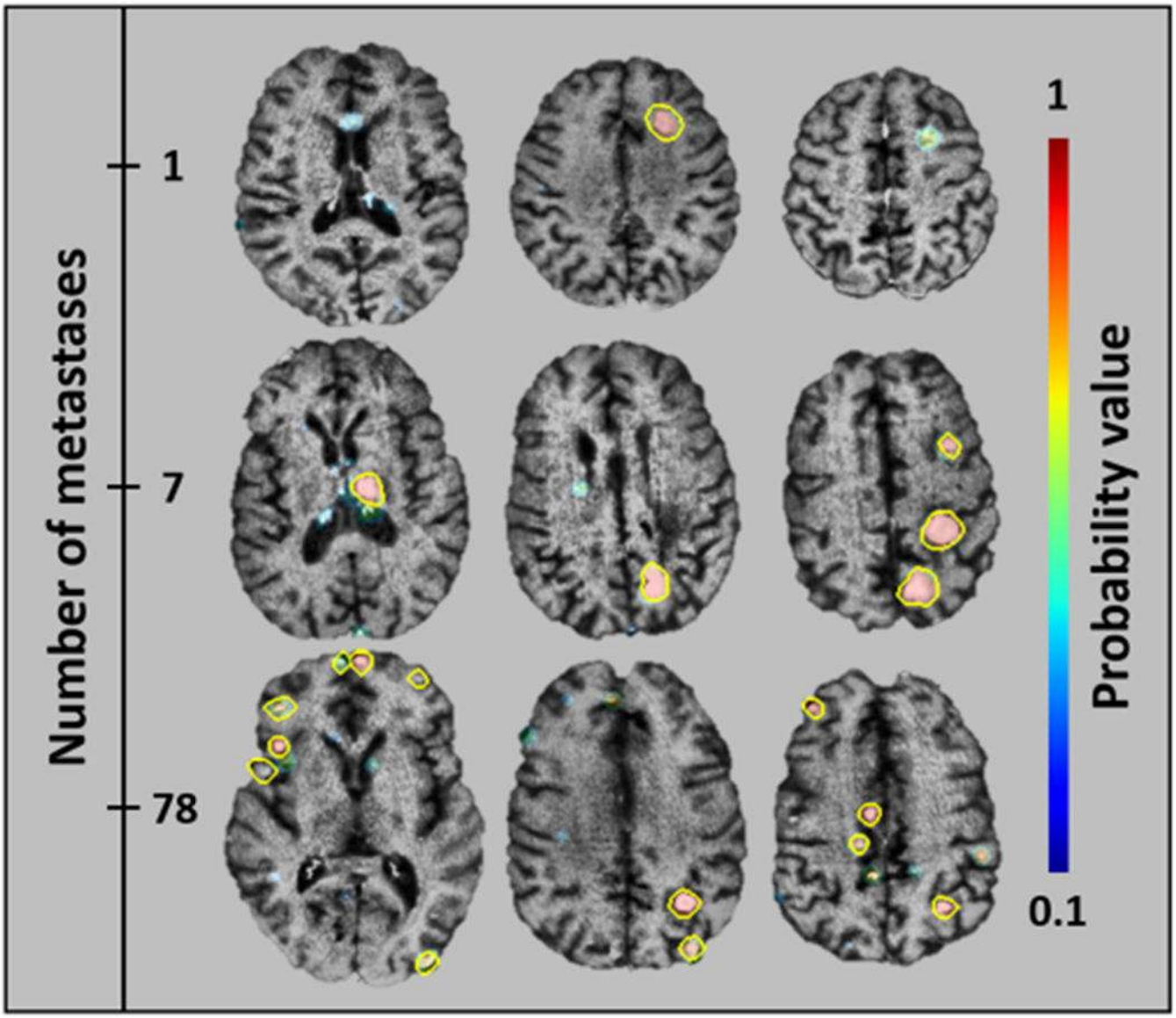

FIGURE 4:

Examples of representative cases with few (top row), moderate (middle row), and extensive (bottom row) metastatic disease. The top row shows a 70-year-old woman presenting with one brain metastasis from colon carcinoma. The middle row shows a 68-year-old man presenting with seven brain metastases from lung cancer. The bottom row shows a 37-year-old man presenting with 78 brain metastases from lung cancer. The network's predictions are shown as probability maps and the yellow lines show the manually delineated lesions.

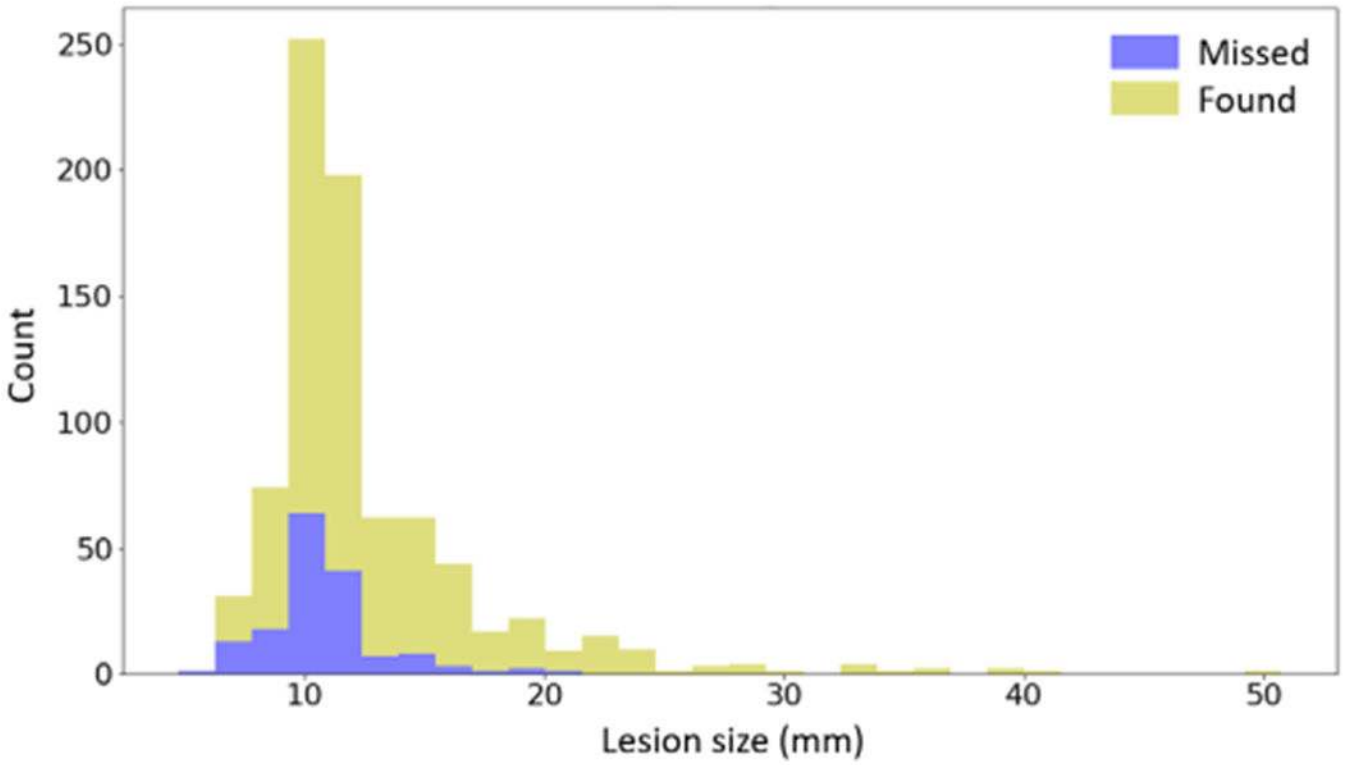

FIGURE 5:
Stacked histogram showing the number of missed and found lesions as a function of lesion size (greatest diameter). As seen in the histogram, the proposed network does not miss any lesion of size larger than 22 mm.

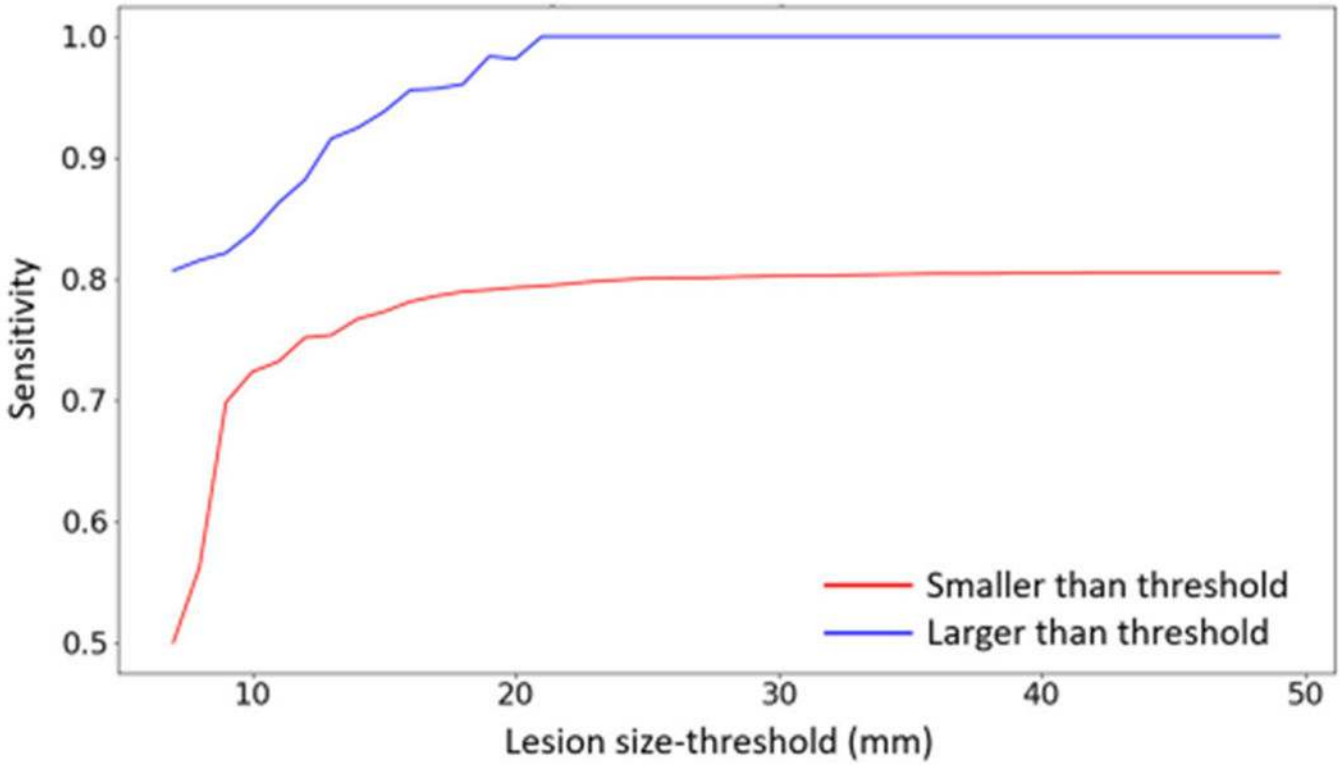

FIGURE 6:
Detection sensitivity as a function of the size used to differentiate large and small lesions. The curves show the different sensitivities of lesions both larger (blue) and smaller (red) than any given threshold value.

TABLE 1.

Demographics

Total number of patients	**156**
Gender	105 Female / 51 Male
Primary cancer:	
Lung	99 (63%)
Breast	33 (21%)
Skin/melanoma	7 (5%)
Genitourinary	7 (5%)
Gastrointestinal	5 (3%)
Miscellaneous	5 (3%)
Number of metastases:	Number of patients
≤3	64 (41%)
4–10	47 (30%)
>10	45 (29%)

Table 2.

Overview of MRI Pulse Sequences and Key Imaging Parameters

Technique	3D T1 BRAVO	Pre/Post 3D T1 CUBE	3D CUBE FLAIR
TR (msec) *	12.02 / 8.24	550 / 602	6000
TE (msec) *	5.05 / 3.24	9.54 / 12.72	119 / 136
Flip angle *	20 / 13	90	90
FOV (mm^2)	240 × 240	250×250	240 × 240
Inversion time (msec) *	300 / 400	—	1880 / 1700
Acquisition matrix	256 × 256	256 × 256	256 × 256
Slice thickness (mm)	1	1	1–1.6
# of slices	160	270–320	270–320
Slice acquisition plane	Axial	Sagittal	Sagittal

TR = repetition time; TE = echo time; FOV = field-of-view; BRAVO–T1-weighted inversion recovery prepped fast spoiled gradient-echo; CUBE–T1-weighted fast spin-echo; FLAIR–fluid attenuated inversion recovery.

* In case of varying parametric values between field strength,'/' notation is given (1.5T / 3T)

TABLE 3.

Summary of Detection and Segmentation Metrics (Mean Value ± Standard Deviation)

A: Voxel-by-voxel detection accuracy using ROC statistics*			
# of metastases	**AUC**	**Sensitivity**	**Specificity**
1 to 3	0.99 ± 0.01	98 ± 3%	98 ± 2%
4 to 10	0.97 ± 0.05	92 ± 10%	97 ± 3%
>10	0.97 ± 0.03	92 ± 7%	95 ± 3%
All cases	0.98 ± 0.04	94 ± 8%	97 ± 3%
B: Detection and segmentation accuracy at an optimal probability threshold**			
# of metastases	**Dice**	**Recall**	**Precision**
1 to 3	0.76 ± 0.20	0.54 ± .026	0.79 ± 0.27
4 to 10	0.83 ± 0.04	0.59 ± 0.21	0.76 ± 0.22
>10	0.78 ± 0.05	0.44 ± 0.18	0.81 ± 0.11
All cases	0.79 ± 0.12	0.53 ± 0.22	0.79 ± 0.20
C: Lesion-by-lesion detection accuracy at an optimal probability threshold**			
# of metastases	**Sensitivity**	**FP (no size limit)**	**FP (10 mm^3 size limit)**
1 to 3	92 ± 25%	3.2 ± 4.0	1.7 ± 2.0
4 to 10	81 ± 19%	8.5 ± 9.8	4.4 ± 6.0
>10	76 ± 20%	13.1 ± 18.9	4.1 ± 10.3
All cases	83 ± 22%	8.3 ± 12.9	3.4 ± 7.0

AUC = area under the receiver operating characteristic (ROC) curve; FP = false positive.

*Sensitivity and specificity were determined by using the maximum value of Youden's index.

**The metrics were estimated using an optimal probability threshold of 0.93, as determined from the development set.

TABLE 4.

P-values Comparing Subgroups Using Wilcoxon Rank Sum Test

A: Voxel-by-voxel detection accuracy using ROC statistics*			
Subgroups	**AUC**	**Sensitivity**	**Specificity**
G1 vs. G2	**0.0131**	**0.0017**	**0.0496**
G1 vs. G3	**0.0024**	**0.0024**	**0.0038**
G2 vs. G3	0.4282	0.6794	**0.0421**
B: Detection and segmentation accuracy at an optimal probability threshold**			
Subgroups	**Dice**	**Recall**	**Precision**
G1 vs. G2	1.0000	0.5816	0.2557
G1 vs. G3	0.0629	0.1131	0.2557
G2 vs. G3	0.0915	**0.0230**	0.8633
C: Lesion-by-lesion detection accuracy at an optimal probability threshold**			
Subgroups	**Sensitivity**	**FP (no size limit)**	**FP (10 mm^3 size limit)**
G1 vs. G2	**0.0069**	**0.0158**	0.0829
G1 vs. G3	**0.0002**	**0.0139**	0.6352
G2 vs. G3	0.3952	0.7031	0.1178

G1 = subgroup having 1–3 metastases; G2 = subgroup having 4–10 metastases; G3 = subgroup having >10 metastases. Significant *P*-values are highlighted in bold. All *P*-values were measured using the Wilcoxon rank sum test.

*Sensitivity and specificity were determined using the maximum value of Youden's index.

**The metrics were estimated using an optimal probability threshold of 0.93, as determined from the development set.