Fully Automated MR Detection and Segmentation of Brain Metastases in Non-small Cell Lung Cancer Using Deep Learning

Stephanie T. Jünger, MD,1,2 Ulrike Cornelia Isabel Hoyer, MD,3 Diana Schaufler, MD,4 Kai Roman Laukamp, MD,3 Lukas Goertz, MD,3 Frank Thiele, MSc,3,5 Jan-Peter Grunz, MD,6 Marc Schlamann, MD,3 Michael Perkuhn, MD,3,5 Christoph Kabbasch, MD,3 Thorsten Persiegel, MD,3 Stefan Grau, MD,1,2 Jan Borggrefe, MD,3,7 Matthias Scheffler, MD,4 Rahil Shahzad, PhD,3,5 and Lenhard Pennig, MD3*

Background: Non-small cell lung cancer (NSCLC) is the most common tumor entity spreading to the brain and up to 50% of patients develop brain metastases (BM). Detection of BMs on MRI is challenging with an inherent risk of missed diagnosis.

Purpose: To train and evaluate a deep learning model (DLM) for fully automated detection and 3D segmentation of BMs in NSCLC on clinical routine MRI.

Study Type: Retrospective.

Population: Ninety-eight NSCLC patients with 315 BMs on pretreatment MRI, divided into training (66 patients, 248 BMs) and independent test (17 patients, 67 BMs) and control (15 patients, 0 BMs) cohorts.

Field Strength/Sequence: T1-/T2-weighted, T1-weighted contrast-enhanced (T1CE; gradient-echo and spin-echo sequences), and FLAIR at 1.0, 1.5, and 3.0 T from various vendors and study centers.

Assessment: A 3D convolutional neural network (DeepMedic) was trained on the training cohort using 5-fold cross-validation and evaluated on the independent test and control sets. Three-dimensional voxel-wise manual segmentations of BMs by a neurosurgeon and a radiologist on T1CE served as the reference standard.

Statistical Tests: Sensitivity (recall) and false positive (FP) findings per scan, dice similarity coefficient (DSC) to compare the spatial overlap between manual and automated segmentations, Pearson’s correlation coefficient (r) to evaluate the relationship between quantitative volumetric measurements of segmentations, and Wilcoxon rank-sum test to compare the volumes of BMs. A P value <0.05 was considered statistically significant.

Results: In the test set, the DLM detected 57 of the 67 BMs (mean volume: 0.99 ± 4.24 cm³), resulting in a sensitivity of 85.1%, while FP findings of 1.5 per scan were observed. Missed BMs had a significantly smaller volume (0.05 ± 0.04 cm³) than detected BMs (0.96 ± 2.4 cm³). Compared with the reference standard, automated
Lung cancer is the leading cause of cancer-related death among both men and women, and non-small cell lung cancer (NSCLC) constitutes the most prevalent subtype (around 85%). Because of its high incidence and metastatic pattern, NSCLC represents the most common tumor entity spreading to the brain and up to 50% of patients develop brain metastases (BMs). Therefore, current guidelines recommend MRI of the head to screen for BMs in advanced NSCLC.

Due to the rising number of MRI scans, the workload of radiologists has increased steadily, bearing the risk of missed findings with serious consequences, in particular regarding oncological patients especially in a screening setting. Given their inconsistent appearance with numerous scattered lesions of varying contrast enhancement, accurate detection of BMs in NSCLC is challenging and the satisfaction of search effect can lead to an oversight of small additional lesions. However, the exact assessment of the number of BMs is relevant because it guides the most suitable type of treatment, together with patient’s characteristics. Automated detection of BMs could provide a preselection of lesions and facilitate the assessment of the extent of disease for radiologists and treating physicians.

Alongside systemic antineoplastic therapy and surgical resection, stereotactic radiosurgery represents an important treatment modality for BMs, which requires a precise determination and delineation of lesion size. In clinical routine, definition of the target volume is performed manually by the radiosurgeon. However, manual segmentation is time-consuming, especially given the inconsistent appearance of BMs in NSCLC and a potentially high number of lesions. Further, manual segmentation bears the risk of inter- and intra-reader variabilities, consequently reducing reproducibility and increasing inaccuracies, which may affect the treatment outcome. In this context, automated segmentation could facilitate the clinical workflow and potentially improve the patient’s outcome by assessing tumor burden and potential response based on exact 3D assessment of tumor volume.

Previous studies have demonstrated the potential of deep learning models (DLMs), in particular deep convolutional neural networks (CNNs), for automatic detection, segmentation, and classification of various intracranial tumors. In this context, recent studies have introduced DL-based detection of BMs from different primary tumors, reporting promising results with sensitivities around 80% and higher. However, the often reported high number of false positive (FP) findings, the use of homogeneous MRI data, unavailable independent test sets, and the inability to segment lesions question their applicability in clinical routine.

Thus, the purpose of this study was to train a DLM for fully automated detection and segmentation of BMs in NSCLC and validate its performance on independent test and control sets using heterogeneous scanner data from multiple vendors and study centers.

**Materials and Methods**

The institutional review board approved this monocentric study (reference number: 19–1208). Given the retrospective design, the requirement for written informed consent was waived.

**Patient Population**

The authors reviewed the institutional image archiving system at a tertiary care university hospital between January 2012 and March 2020 for MRI examinations of the head of NSCLC patients receiving dedicated oncological treatment for BMs. Patient demographics and treatment of patients were retrieved from the patients’ medical charts.

Ninety patients were identified applying the following inclusion criteria: (I) MRI scans at primary diagnosis of BMs; (II) BM-specific therapy after diagnosis, eg, stereotactic radiosurgery or resection; and (III) a complete multiparametric MRI-dataset, consisting of unenhanced $T_1$-, $T_2$-, and $T_2$-weighted, fluid-attenuated inversion recovery (FLAIR) sequences. Seven patients were excluded due to the following criteria: (I) severe MRI artifacts reducing image quality; (II) insufficient application of intravenous (i.v.) contrast agent; (III) incomplete coverage of the brain in one or more aforementioned sequences. The included 83 patients (53 females, mean age of 57.4 ± 11.0 years at the time of MRI) were randomly allocated into a training set with 66 patients and a test set with 17 patients.

To evaluate the performance of the trained DLM in terms of FP findings in patients without BMs, a control set ($n = 15$ patients, 10 females, mean age of 61.2 ± 13.2 years at the time of MRI) was established by randomly selecting MRI scans of patients treated for NSCLC at the same tertiary care university hospital. These patients received head MRI to screen for BMs between January 2012 and March 2020 ($T_1$-, $T_2$-, and FLAIR sequences) but...
did not present BMs. While randomly selecting these scans, datasets were not chosen when they presented the aforementioned exclusion criteria.

There were no overlapping data between the three sets; the test and control sets were only used for independent evaluation of the DLM after training. After anonymization, MRI datasets (T1-/T2-weighted, T1 CE, and FLAIR sequences) were exported to Intellispace Discovery (ISD, v3.0.6, Philips Healthcare, Best, The Netherlands).

Imaging

MRI was performed at 1.0 (n = 6), 1.5 (n = 38), and 3.0 T (n = 54) using scanners from different vendors. All scans were conducted for clinical purposes only and consisted of T1-/T2-weighted, T1 CE, and FLAIR sequences. Sixty-five scans were acquired at our institution following a standardized protocol using i.v. administration of Gadolinium (Dotarem, Guerbet GmbH, Roissy, France; 0.5 mmol/mL, 1 mL = 279.3 mg gadoteric acid = 78.6 mg gadolinium) with a concentration of 0.1 mmol/kg body weight. The remaining 33 patients received MRI scans at referring constituions without standardized application of contrast agent.

Table 1 provides detailed data on MRI scanners in training, test, and control sets with Tables 2–4 listing the imaging parameters for each field strength of training, test, and control sets (including names of T1-/T2-weighted and T1 CE sequences).

Reference Standard

To establish the reference standard (RS), a neurosurgeon (S.T.J.) with 3 years and a radiologist (L.P.) with 4 years of experience in neuro-oncological MRI defined the reference standard of BMs on aforementioned sequences of BM positive and negative cases in consensus. To this end, they reviewed the original radiology report and performed a 2-fold assessment of the included MRI scans as well as prior/follow-up imaging. Additionally, the readers assessed the localization of BMs (parenchymal, pachymeningeal, supratentorial, and infratentorial) in consensus. In case of uncertainties, a neuroradiologist (J.B.) with 14 years of experience in neuro-oncological imaging was consulted.

By assessing T1-/T2-weighted, T1 CE, and FLAIR sequences on ISD, manual segmentation of BMs was performed by the aforementioned neurosurgeon and radiologist on T1 CE in consensus (including CE- and non-CE tumor parts). The initial segmentation was provided by the neurosurgeon and then presented to the radiologist to define the final segmentation of BMs. Segmentation of BMs was conducted interactively using a two-step semiautomatic approach. By 3D voxel-wise regional thresholding, a rough segmentation of tumor tissue was obtained, which was further edited manually using 2D editing tools if required.

Deep Learning Model

MODEL ARCHITECTURE. In this study, a 3D CNN based on DeepMedic (Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, UK) was employed.

The network comprises a deep 3D CNN architecture with two identical pathways. Three-dimensional image patches serve as an input to the two pathways. For the first pathway, original isotropic patches are employed, whereas for the second pathway, the patches are down-sampled to a third of their original size. This approach enables the increased depiction of contextual information. The deep CNN comprises 11 layers with kernels of size 3. The model consists of residual connections for layers 4, 6, and 8 with each layer followed by batch normalization and a parametric rectified linear unit as the activation function. In contrast, layers 9 and 10 are fully connected. The last prediction layer has a kernel size of 1 and employs sigmoid as the activation function.

PREPROCESSING. Before the multiparametric scans (T1-/T2-weighted, T1 CE, and FLAIR) of the BMs were passed to the DLM, automatic preprocessing of data was performed. The preprocessing pipeline included: (I) skull stripping using a brain mask, (II) co-registration of T1-/T2-weighted and FLAIR sequences to T1 CE, (III) bias field correction of all four sequences employing a proprietary method (Philips Healthcare) for MRI of the head, (IV) resampling to an isotropic resolution of 1 mm × 1 mm × 1 mm, and (V) intensity normalization to zero-mean with a standard deviation of one.

TRAINING. The training of the DLM was performed on the dedicated training set (n = 66 patients) using a 5-fold cross-validation approach without overlapping data, which resulted in five trained models. For training, multichannel 3D image patches with a size of 25³ were passed to the CNN. These image patches were extracted by employing a distribution of 50% between background and BMs to ensure class balance. In order to increase the number of training samples, image augmentation was used by randomly flipping the image patches along their axes. To compare the spatial overlap, the dice similarity coefficient (DSC) was used as the loss function and the root mean square propagation as the optimizer. Adaptive learning rate schedule was employed in which the initial learning rate was divided equally every time the accuracy did not improve for more than three epochs. Training batch size was set to 10 with the number of training epochs being fixed at 35.

EVALUATION. The evaluation of the DLM was performed on the independent test set (n = 17 patients) by applying the five individual models from the 5-fold cross-validation training approach. To this end, 3D image patches of size 45³ were extracted to reduce the time spent during inference on the test data. To reduce the detection of FP findings, the segmentation results from each of the five DLMs were fused using a majority voting scheme. To assess for FPs in BM negative cases, the fused DLM was additionally applied on the control set (n = 15 patients).

Automatically detected lesions <0.003 cm³ during inference on training, test, and control sets were, by default, regarded as image noise and discarded. This threshold was based on the resolution of T1 CE sequences (in which a volume of 0.003 cm³ is about two voxels). Given the limitation of scan resolution, readers cannot accurately segment lesions smaller than this volume.

TIME REQUIRED. The average time required including image preprocessing to run a complete pipeline on a dataset was about
8 minutes: (I) bias field correction <1 second, (II) co-registration and skull stripping <7 minutes, (III) image standardization <1 minute, and (IV) inference on the imaging data 10 seconds (using a Tesla-P100 GPU card (Nvidia, Santa Carla, USA)).

### Statistical Analysis
Statistical analysis was performed using JMP Software (release 14, SAS Institute, Cary, NC). Tumor volumes are presented in mean ± standard deviation and as the median with 10/90 percentile. The DSC is reported as the median with 10/90 percentile. The Wilcoxon rank-sum test was used to evaluate for statistical difference between the volumes of the BMs in training/test sets and of detected/missed lesions. The Wilcoxon signed-rank test was employed to assess for statistical difference of the volumes of detected BMs by the DLM and the reference standard. Statistical significance was set to $P < 0.05$.

#### Table 1. MRI Scanner Models (Manufactured by Philips Healthcare, Best, The Netherlands, GE Healthcare, Chicago, IL, USA, and Siemens Healthineers, Erlangen, Germany), Field Strength, Examination Location, and 3D/2D Acquisition of T1-Weighted Contrast-Enhanced (CE) and FLAIR (Fluid-Attenuated Inversion Recovery) Sequences in Training, Test, and Control Sets

| Scanner model       | Training Set ($n = 66$) | Test Set ($n = 17$) | Control Set ($n = 15$) |
|---------------------|-------------------------|---------------------|------------------------|
|                     | Number of Patients      | Number of Patients  | Number of Patients     |
| Philips Achieva     | 33                      | 3                   | 8                      |
| GE Optima           | -                       | 1                   | -                      |
| Philips Ingenia     | 11                      | 3                   | 4                      |
| Philips Intera      | 9                       | 5                   | 1                      |
| Philips Panorama    | 2                       | -                   | 1                      |
| Siemens Aera        | 3                       | 2                   | -                      |
| Siemens Amira       | 1                       | -                   | -                      |
| Siemens Avanto      | 5                       | -                   | -                      |
| Siemens Espree      | -                       | 1                   | -                      |
| Siemens Skyra       | -                       | 1                   | -                      |
| Siemens Symphony   | 1                       | 1                   | 1                      |
| Siemens Prisma      | 1                       | -                   | -                      |
| Field strength      |                         |                     |                        |
| 1.0 T               | 4 (6.1%)                | 1 (5.9%)            | 1 (6.7%)               |
| 1.5 T               | 23 (34.8%)              | 11 (64.7%)          | 4 (26.7%)              |
| 3.0 T               | 39 (59.1%)              | 5 (29.4%)           | 10 (66.7%)             |
| Examination location|                         |                     |                        |
| Local               | 43 (65.2%)              | 8 (47.1%)           | 14 (93.3%)             |
| Referring           | 23 (34.8%)              | 9 (52.9%)           | 1 (6.7%)               |
| Sequence            |                         |                     |                        |
| T1 CE               |                         |                     |                        |
| 3D acquisition      | 42 (63.6%)              | 11 (64.7%)          | 9 (60.0%)              |
| 2D acquisition      | 24 (36.4%)              | 6 (35.3%)           | 6 (40.0%)              |
| FLAIR               |                         |                     |                        |
| 3D acquisition      | 5 (7.6%)                | 3 (17.6%)           | 2 (13.3%)              |
| 2D acquisition      | 61 (92.4%)              | 14 (82.4%)          | 13 (86.7%)             |
| Parameter                                      | T<sub>1</sub> | T<sub>1</sub> CE | T<sub>2</sub> | FLAIR   |
|-----------------------------------------------|---------------|----------------|-------------|---------|
| **3.0 T**                                     |               |                |             |         |
| Slice thickness (mean, mm ± SD)               | 4.8 ± 0.7     | 2.5 ± 1.8      | 4.9 ± 0.3   | 4.1 ± 1.4 |
| In-plane resolution (mean, mm<sup>2</sup> ± SD) | 0.5 ± 0.1     | 0.5 ± 0.1      | 0.4 ± 0.1   | 0.5 ± 0.1 |
| Acquisition matrix                            | 256–404 × 200–333 | 256–460 × 184–460 | 280–512 × 221–418 | 224–280 × 168–224 |
| Repetition time* (ms)                         | 8–910         | 8–730          | 708–4000    | 4800–12,000 |
| Echo time (ms)                                | 2–13          | 3–13           | 16–107      | 81–340   |
| Inversion time (ms)                           | -             | -              | -           | 1650–2850 |
| Flip angle (°)                                | 10–90         | 8–90           | 18–150      | 90–150   |
| Sequence name (number of scans)               | 39            | 39             | 39          | 39      |
| Gradient-echo                                 | 6 (15.4%)     | 26 (66.7%)     | 5 (12.8%)   | -       |
| Gradient-echo with IR                         | -             | 1 (2.6%)       | -           | -       |
| Spin-echo                                     | 33 (84.6%)    | 12 (30.8%)     | 34 (87.2%)  | -       |
| Spin-echo with IR                             | -             | -              | -           | -       |
| **1.5 T**                                     |               |                |             |         |
| Slice thickness (mean, mm ± SD)               | 5.0 ± 1.4     | 3.4 ± 2.1      | 5.2 ± 0.9   | 5.2 ± 0.7 |
| In-plane resolution (mean, mm<sup>2</sup> ± SD) | 0.7 ± 0.2     | 0.7 ± 0.2      | 0.6 ± 0.2   | 0.8 ± 0.2 |
| Acquisition matrix                            | 248–320 × 193–310 | 192–388 × 174–309 | 256–556 × 200–339 | 232–320 × 171–256 |
| Repetition time* (ms)                         | 6–792         | 6–678          | 3300–8923   | 4800–11,000 |
| Echo time (ms)                                | 2–15          | 2–17           | 85–130      | 86–343   |
| Inversion time (ms)                           | -             | -              | -           | 1660–2800 |
| Flip angle (°)                                | 8–90          | 8–90           | 90–170      | 90–180   |
| Sequence name (number of scans)               | 23            | 23             | 23          | 23      |
| Gradient-echo                                 | 4 (17.4%)     | 11 (47.8%)     | -           | -       |
| Gradient-echo with IR                         | -             | 2 (8.7%)       | -           | -       |
| Spin-echo                                     | 19 (82.6%)    | 10 (43.8%)     | 23 (100.0%) | -       |
| Spin-echo with IR                             | -             | -              | -           | -       |
| **1.0 T**                                     |               |                |             |         |
| Slice thickness (mean, mm ± SD)               | 5.8 ± 0.4     | 3.6 ± 2.0      | 6.0         | 5.8 ± 0.4 |
| In-plane resolution (mean, mm<sup>2</sup> ± SD) | 0.9 ± 0.1     | 0.6 ± 0.2      | 0.7 ± 0.2   | 0.9 ± 0.1 |
| Acquisition matrix                            | 256–276 × 205–218 | 256–336 × 205–336 | 272–328 × 208–263 | 256–280 × 161–173 |
Table 2. Continued

| Parameter                  | T<sub>1</sub> | T<sub>1</sub> CE | T<sub>2</sub> | FLAIR |
|----------------------------|--------------|-----------------|---------------|-------|
| Repetition time* (ms)      | 168–653      | 25–542          | 772–6262      | 6000  |
| Echo time (ms)             | 2.2–15       | 2–15            | 21–120        | 120   |
| Inversion time (ms)        | -            | -               | -             | 2000  |
| Flip angle (°)             | 69–80        | 30–80           | 18–90         | 90    |
| Sequence name (number of scans) | 4        | 4               | 4             | 4     |
| Gradient-echo              | 2 (50.0%)    | 3 (75.0%)       | 2 (50.0%)     | -     |
| Gradient-echo with IR      | -            | -               | -             | -     |
| Spin-echo                  | 2 (50.0%)    | 1 (25.0%)       | 2 (50.0%)     | -     |
| Spin-echo with IR          | -            | -               | -             | -     |

SD: standard deviation; CE: contrast-enhanced; FLAIR: fluid-attenuated inversion recovery; IR: inversion recovery.

*The repetition time (TR) has been extracted as listed in the respective Digital Imaging and COmmunications in Medicine (DICOM) Tag. Since MRI datasets from different vendors have been included in this work (with different definitions of the TR), there is a broad range of TRs among examinations.

Table 3. MRI Scan Parameters and Sequence Names for Each Field Strength in the Test Set

| Parameter                  | T<sub>1</sub> | T<sub>1</sub> CE | T<sub>2</sub> | FLAIR |
|----------------------------|--------------|-----------------|---------------|-------|
| 3.0 T                      |              |                 |               |       |
| Slice thickness (mean, mm ± SD) | 5.0      | 2.7 ± 1.9       | 4.8 ± 0.4     | 4.2 ± 1.6 |
| In-plane resolution (mean, mm<sup>2</sup> ± SD) | 0.5 ± 0.1 | 0.6 ± 0.2       | 0.4 ± 0.1     | 0.5 ± 0.1 |
| Acquisition matrix         | 256–404 × 200–333 | 256–412 × 205–414 | 280–512 × 221–418 | 224–384 × 211–223 |
| Repetition time* (ms)      | 319–2000     | 8–1650          | 3000–6790     | 4800–12,000 |
| Echo time (ms)             | 5–13         | 3–13            | 80–112        | 81–340 |
| Inversion time (ms)        | -            | -               | -             | 1650–2850 |
| Flip angle (°)             | 80–150       | 15–90           | 90–140        | 90–150 |
| Sequence name (number of scans) | 5        | 5               | 5             | -     |
| Gradient-echo              | 1 (20.0%)    | 3 (40.0%)       | -             | -     |
| Gradient-echo with IR      | -            | -               | -             | -     |
| Spin-echo                  | 3 (60.0%)    | 1 (20.0%)       | 5 (100.0%)    | -     |
| Spin-echo with IR          | 1 (20.0%)    | -               | -             | -     |
| 1.5 T                      |              |                 |               |       |
| Slice thickness (mean, mm ± SD) | 4.9 ± 1.4 | 2.4 ± 1.8       | 5.4 ± 0.5     | 4.7 ± 1.6 |
| In-plane resolution (mean, mm<sup>2</sup> ± SD) | 0.8 ± 0.2 | 0.8 ± 0.2       | 0.6 ± 0.2     | 1.0 ± 0.4 |
To evaluate the detection performance of the DLM, sensitivity (recall), precision (positive predictive value), and the F1-score were computed. Automated detection of a BM was achieved if the DLM obtained a spatial overlap of at least two voxels with the RS. Additionally, subanalysis of the detection performance of the fused DLM was performed in terms of MRI specifics (examination location, field strength) and localization of BMs (parenchymal, pachymeningeal, supratentorial, and infratentorial).

To investigate the segmentation performance of the DLM, automatically obtained segmentations were compared with the manual annotations by computing the overlap measure between the segmentations using the DSC. To compare quantitative volumetric measurements of both manual and automatic segmentations, the Pearson’s correlation coefficient (r) was calculated and Bland–Altman analysis was performed.

Furthermore, an additional subanalysis was performed to evaluate the detection and segmentation performance of the fused DLM for lesions >0.15 cm³.

Table 3. Continued

| Parameter | T₁ | T₁ CE | T₂ | FLAIR |
|-----------|----|-------|----|-------|
| Acquisition matrix | 232–320 × 180–256 | 224–328 × 144–270 | 256–556 × 200–512 | 216–320 × 172–260 |
| Repetition time* (ms) | 176–1800 | 25–1800 | 2700–7807 | 4800–9000 |
| Echo time (ms) | 3–20 | 2–10 | 85–115 | 86–335 |
| Inversion time (ms) | - | - | - | 1660–2500 |
| Flip angle (°) | 8–150 | 8–120 | 90–160 | 90–180 |
| Sequence name (number of scans) | 11 | 11 | 11 | - |
| Gradient-echo | 1 (9.1%) | 6 (54.5%) | - | - |
| Gradient-echo with IR | 1 (9.1%) | 2 (18.2%) | - | - |
| Spin-echo | 9 (81.8%) | 3 (27.3%) | 11 (100.0%) | - |
| Spin-echo with IR | - | - | - | - |
| 1.0 T | | | | |
| Slice thickness (mm) | 5.0 | 5.0 | 5.0 | 5.0 |
| In-plane resolution (mm²) | 0.5 | 0.5 | 0.5 | 0.5 |
| Acquisition matrix | 272 × 218 | 352 × 282 | 320 × 247 | 256 × 196 |
| Repetition time* (ms) | 678 | 199 | 5055 | 6000 |
| Echo time (ms) | 15 | 2 | 100 | 120 |
| Inversion time (ms) | - | - | - | 2000 |
| Flip angle (°) | 62 | 60 | 90 | 90 |
| Sequence name (number of scans) | 1 | 1 | 1 | - |
| Gradient-echo | - | 1 (100.0%) | - | - |
| Gradient-echo with IR | - | - | - | - |
| Spin-echo | 1 (100.0%) | - | 1 (100.0%) | - |
| Spin-echo with IR | - | - | - | - |

SD: standard deviation; CE: contrast-enhanced; FLAIR: fluid-attenuated inversion recovery; IR: inversion recovery.
*The repetition time (TR) has been extracted as listed in the respective Digital Imaging and COmmunications in Medicine (DICOM) Tag. Since MRI datasets from different vendors have been included in this work (with different definitions of the TR), there is a broad range of TRs among examinations.
| Parameter                                      | T₁       | T₁ CE    | T₂       | FLAIR    |
|-----------------------------------------------|----------|----------|----------|----------|
| **3.0 T**                                     |          |          |          |          |
| Slice thickness (mean, mm ± SD)               | 5.0      | 2.0 ± 1.5| 5.0      | 3.8 ± 1.4|
| In-plane resolution (mean, mm² ± SD)          | 0.5 ± 0.1| 0.5 ± 0.1| 0.4 ± 0.1| 0.5 ± 0.1|
| Acquisition matrix                            | 256–404 × 272–333 | 256–412 × 272–414 | 280–436 × 223–341 | 224–280 × 211–224 |
| Repetition time* (ms)                         | 316–730  | 9–756    | 878–3826 | 4800–12,000 |
| Echo time (ms)                                | 5–13     | 4–13     | 16–80    | 140–253  |
| Inversion time (ms)                           | -        | -        | -        | 1650–2850|
| Flip angle (°)                                | 80–90    | 30–90    | 18–90    | 90       |
| Sequence name (number of scans)               | 10       | 10       | 10       | -        |
| Gradient-echo                                 | 2 (20.0%)| 8 (80.0%)| 1 (10.0%)| -        |
| Gradient-echo with IR                         | -        | -        | -        | -        |
| Spin-echo                                     | 8 (80.0%)| 2 (20.0%)| 9 (90.0%)| -        |
| Spin-echo with IR                             | -        | -        | -        | -        |
| **1.5 T**                                     |          |          |          |          |
| Slice thickness (mean, mm ± SD)               | 5.8 ± 0.4| 4.7 ± 1.8| 5.8 ± 0.4| 5.8 ± 0.4|
| In-plane resolution (mean, mm² ± SD)          | 0.8 ± 0.2| 0.8 ± 0.2| 0.6 ± 0.2| 0.8 ± 0.2|
| Acquisition matrix                            | 256–280 × 192–205 | 244–280 × 192–246 | 256–384 × 200–308 | 256 × 172–203 |
| Repetition time* (ms)                         | 400–791  | 25–791   | 3890–7807| 6000–9000|
| Echo time (ms)                                | 8–15     | 8–15     | 90–110   | 100–120  |
| Inversion time (ms)                           | -        | -        | -        | 2000–2500|
| Flip angle (°)                                | 8–150    | 30–90    | 90–150   | 90–150   |
| Sequence name (number of scans)               | 4        | 4        | 4        | -        |
| Gradient-echo                                 | -        | 1 (25.0%)| -        | -        |
| Gradient-echo with IR                         | -        | -        | -        | -        |
| Spin-echo                                     | 4 (100.0%)| 3 (75.0%)| 4 (100.0%)| -        |
| Spin-echo with IR                             | -        | -        | -        | -        |
| **1.0 T**                                     |          |          |          |          |
| Slice thickness (mm)                           | 6.0      | 6.0      | 6.0      | 6.0      |
| In-plane resolution (mm²)                     | 0.9      | 0.9      | 0.9      | 0.9      |
| Acquisition matrix                            | 276 × 218| 280 × 224| 280 × 224| 280 × 172|
| Repetition time* (ms)                         | 653      | 653      | 774      | 6000     |
Patient Characteristics

Three-hundred-fifteen BMs were segmented on ISD and comprised the RS. Thirty-one patients presented with a single BM. The highest number of BMs per patient was \( n = 29 \) in the training and \( n = 15 \) in the test set. Supratentorial (81\% and 82\% in training and test sets, respectively) was more common than posterior fossa localization. For treatment of BMs, radiotherapy was performed in 77\% and 82\% of patients, respectively. Table 5 provides detailed information regarding patient demographics as well as treatment, number, and metastatic pattern of the BMs in training and test sets.

Evaluation of the DLM on the Independent Test Set

Based on manual segmentations, the reference standard consisted of 248 lesions with a volume of 0.99 ± 4.24 cm\(^3\) (median: 0.06 cm\(^3\), 10/90 percentiles: 0.01–1.70 cm\(^3\)) in the training set and of 67 BMs with a volume of 0.96 ± 2.44 cm\(^3\) (0.12 cm\(^3\), 0.02–2.54 cm\(^3\); \( P < 0.05 \)) in the test set. Detailed results on the performance of each DLM and their fusion on the test set are provided in Table 6.

DETECTION PERFORMANCE. The fused DLM detected 57 of the 67 BMs, representing a sensitivity of 85.1\%. The missed BMs yielded a volume of 0.05 ± 0.04 cm\(^3\) (median: 0.03 cm\(^3\), 10/90 percentiles: 0.01–0.14 cm\(^3\)), whereas the detected BMs presented a volume of 0.96 ± 2.4 cm\(^3\) (0.15 cm\(^3\), 0.02–3.20 cm\(^3\); \( P < 0.05 \)). The individual DLMs showed findings of 3.3–5.7 FPs per scan. By employing the majority voting scheme, the fusion of the five DLMs reduced the FP findings to 1.5 per scan. Consequently, fusion of the DLMs increased the precision (68.7\%) and the F1-score (0.76) compared with the individual DLMs (39.0–50.8\% and 0.55–0.64, respectively). Illustrative examples of detected and missed BMs as well as FP lesions by the DLM are given in Figs. 1–3.

A subanalysis of the detection performance of the fused DLM in terms of MRI specifics and localization of BMs is provided in Table 7.

SEGMENTATION PERFORMANCE. For the detected BMs, the fused DLM achieved a median DSC of 0.72 (10/90 percentiles: 0.20–0.92) compared with manual segmentations. When comparing the volumetric assessment for detected BMs by automated (0.81 ± 1.96 cm\(^3\); median: 0.12 cm\(^3\), 10/90 percentiles: 0.01–1.70 cm\(^3\)) and manual (1.11 ± 2.61 cm\(^3\); 0.15 cm\(^3\), 0.02–3.2 cm\(^3\); \( P < 0.05 \)) segmentations, a correlation coefficient of 0.95 was observed. Figure 4 displays the volumetric correlation between manual and automated segmentations using Pearson’s correlation on a lesion level. Further, Bland–Altman analysis indicated a good agreement between manual and automated segmentations with a mean difference of 0.31 cm\(^3\) between the DLM and the RS per lesion (Figure 4). Figure 5 shows a box plot comparing the DSC as well as the number of FP and false negative findings for the five DLMs using the 5-fold cross-validation and the combined DLM employing the majority-voting scheme.

For lesions >0.15 cm\(^3\) (diameter ~6.6 mm), all 29 BMs were detected by the fused DLM, while 0.1 FPs/scan, a DSC of 0.82 (0.26–0.93), and a volumetric correlation of 0.94 were observed.

Table 4. Continued

| Parameter                  | \( T_1 \) | \( T_1 \) CE | \( T_2 \) | FLAIR |
|----------------------------|-----------|--------------|-----------|-------|
| Echo time (ms)             | 15        | 2            | 21        | 120   |
| Inversion time (ms)        | -         | -            | -         | 2000  |
| Flip angle (°)             | 69        | 69           | 18        | 90    |
| Sequence name              | 1         | 1            | 1         | -     |
| (number of scans)          |           |              |           |       |
| Gradient-echo              | -         | 1 (100.0%)   | 1 (100.0%)| -     |
| Gradient-echo with IR      | -         | -            | -         | -     |
| Spin-echo                  | 1 (100.0%)| -            | -         | -     |
| Spin-echo with IR          | -         | -            | -         | -     |

SD: standard deviation; CE: contrast-enhanced; FLAIR: fluid-attenuated inversion recovery; IR: inversion recovery.

*The repetition time (TR) has been extracted as listed in the respective Digital Imaging and COmmunications in Medicine (DICOM) Tag. Since MRI datasets from different vendors have been included in this work (with different definitions of the TR), there is a broad range of TRs among examinations.

Results

Patient Characteristics

Three-hundred-fifteen BMs were segmented on ISD and comprised the RS. Thirty-one patients presented with a single BM. The highest number of BMs per patient was \( n = 29 \) in the training and \( n = 15 \) in the test set. Supratentorial (81\% and 82\% in training and test sets, respectively) was more common than posterior fossa localization. For treatment of BMs, radiotherapy was performed in 77\% and 82\% of patients, respectively. Table 5 provides detailed information regarding patient demographics as well as treatment, number, and metastatic pattern of the BMs in training and test sets.

Evaluation of the DLM on the Independent Test Set

Based on manual segmentations, the reference standard consisted of 248 lesions with a volume of 0.99 ± 4.24 cm\(^3\) (median: 0.06 cm\(^3\), 10/90 percentiles: 0.01–1.70 cm\(^3\)) in the training set and of 67 BMs with a volume of 0.96 ± 2.44 cm\(^3\) (0.12 cm\(^3\), 0.02–2.54 cm\(^3\); \( P < 0.05 \)) in the test set. Detailed results on the performance of each DLM and their fusion on the test set are provided in Table 6.

DETECTION PERFORMANCE. The fused DLM detected 57 of the 67 BMs, representing a sensitivity of 85.1\%. The missed BMs yielded a volume of 0.05 ± 0.04 cm\(^3\) (median: 0.03 cm\(^3\), 10/90 percentiles: 0.01–0.14 cm\(^3\)), whereas the detected BMs presented a volume of 0.96 ± 2.4 cm\(^3\) (0.15 cm\(^3\), 0.02–3.20 cm\(^3\); \( P < 0.05 \)). The individual DLMs showed findings of 3.3–5.7 FPs per scan. By employing the majority voting scheme, the fusion of the five DLMs reduced the FP findings to 1.5 per scan. Consequently, fusion of the DLMs increased the precision (68.7\%) and the F1-score (0.76) compared with the individual DLMs (39.0–50.8\% and 0.55–0.64, respectively). Illustrative examples of detected and missed BMs as well as FP lesions by the DLM are given in Figs. 1–3.

A subanalysis of the detection performance of the fused DLM in terms of MRI specifics and localization of BMs is provided in Table 7.

SEGMENTATION PERFORMANCE. For the detected BMs, the fused DLM achieved a median DSC of 0.72 (10/90 percentiles: 0.20–0.92) compared with manual segmentations. When comparing the volumetric assessment for detected BMs by automated (0.81 ± 1.96 cm\(^3\); median: 0.12 cm\(^3\), 10/90 percentiles: 0.01–1.70 cm\(^3\)) and manual (1.11 ± 2.61 cm\(^3\); 0.15 cm\(^3\), 0.02–3.2 cm\(^3\); \( P < 0.05 \)) segmentations, a correlation coefficient of 0.95 was observed. Figure 4 displays the volumetric correlation between manual and automated segmentations using Pearson’s correlation on a lesion level. Further, Bland–Altman analysis indicated a good agreement between manual and automated segmentations with a mean difference of 0.31 cm\(^3\) between the DLM and the RS per lesion (Figure 4). Figure 5 shows a box plot comparing the DSC as well as the number of FP and false negative findings for the five DLMs using the 5-fold cross-validation and the combined DLM employing the majority-voting scheme.

For lesions >0.15 cm\(^3\) (diameter ~6.6 mm), all 29 BMs were detected by the fused DLM, while 0.1 FPs/scan, a DSC of 0.82 (0.26–0.93), and a volumetric correlation of 0.94 were observed.
Evaluation of the DLM on the Independent Control Set

On the control set, the fused DLM showed 27 FP findings, which translate to 1.8 FPs per scan. Illustrative examples of FP lesions by the DLM are given in Fig. 6. FP findings were, as in the test set, mostly related to other contrast-enhancing intracranial structures, eg, the choroid plexus and blood vessels as well as to artifacts at the skull base.

Discussion

In this study, we trained a DLM for automated detection and segmentation of BMs from NSCLC and validated its performance on independent test and control sets. Despite the small size and inconsistent appearance of the BMs, the trained DLM achieved a high detection sensitivity and a good spatial overlap with manual segmentations on heterogeneous scanner data while obtaining a low number of FPs/scan in BM positive and negative cases.

Following classical machine learning, different DL-based techniques were recently proposed for fully automated detection of BMs, which included CNNs (DeepMedic,24,32 U-Net,25,28 GoogLe-Net,26,29 CropNet,27 AlexNet,29 and Faster region-based31) and single-shot detector algorithms.30 However, the majority of studies have not focused on a single tumor entity.24–27,29–31 Nevertheless, BMs comprise the most frequent malignant brain tumor.35 Regarding their appearance on MRI, BMs substantially differ between the individual primary tumors and within the same tumor entity (especially in NSCLC).7 When applying a recently proposed DeepMedic DLM, which was trained on BMs in malignant melanoma (MM,32), to the BMs of NSCLC of training and test sets of the present study, a noticeably decreased detection and segmentation performance could be observed. More details are provided in the Supplementary Information. In the current study, in which the DeepMedic architecture received a dedicated training on BMs from NSCLC rather than from MM, the sensitivity and median DSC were higher at a similar FP rate. These differences are most likely due to the distinct appearance of BMs in MM (often showing hyperintense signal on unenhanced T1-weighted sequences and a strong homogenous contrast-enhancement) compared with NSCLC (often necrotic and of various enhancement patterns).7,36 Consequently, BMs from different primary tumors should rather be considered as individual entities and therefore, in order to obtain a sufficient detection and segmentation performance, dedicated DLMs are required (eg, for NSCLC or MM).

In the present study, the DLM provided a high detection sensitivity (85%) on an independent test set, which was similar (eg, Bousabarah et al: 77–82%,25 Zhou et al 81%,30 Grovik et al: 83%,26 Dikici et al: 90%,27 and Charron et al: 93%24) or superior (eg, Deike-Hofmann et al, who investigated the detection of BMs in MM: 73%28) to previously reported results. Of note, the average size of the BMs (0.96 cm3) was comparable to (eg, Bousabarah et al: 1.3–1.9 cm325) or smaller than in recent studies (eg, Charron et al: 2.4 cm324), which most likely explains the slightly...
higher sensitivity of the latter study. Furthermore, the lack of a dedicated test set in the studies by Dikici et al. and Grøvik et al. needs to be considered when analyzing the reported sensitivities. Interestingly, the DLM of the present study showed a higher detection sensitivity on MRI datasets acquired at our institution, which might in part be explained by the standardized imaging protocol considering in-house examinations and in turn a potentially lower image quality at referring sites. Furthermore, the DLM obtained a higher sensitivity for parenchymal and supratentorial BMs compared with pachymeningeal and infratentorial localizations where automated detection is impaired by the high signal intensity of the dura mater and the presence of artifacts at the skull base, respectively.

In order to achieve a sufficient detection of lesions, the resulting high number of FP findings represents a major drawback in automated segmentation of BMs and has been reported to be as high as 6 (Zhou et al.), 8 (Charron et al., Grøvik et al.), 9 (Dikici et al.), 5–13 (Deike-Hofmann et al.), or 20 (Zhang et al.) per scan. In contrast, 1.5 FPs/scan in the test set were observed in the present study by combining the results from the trained five DLMs using the majority-voting scheme. Further, when setting a threshold for a lesion size of 0.15 cm³ (which translates to a diameter of 6.6 mm), the sensitivity improved to 100% while FP findings were reduced to 0.1 per scan. Zhou et al. reported a comparable sensitivity of 98% for BMs larger than 6 mm, whereas the number of FPs was higher (3–4 per scan). Additionally,
the DLM showed a low number of FPs/scan in the control set without BMs, therefore emphasizing the model’s utility for general diagnostic usage as a screening tool. Of note, the FP findings were associated with obvious contrast-enhancing structures that are generally straightforward to identify by an experienced radiologist.

During the past decades, the incidence of NSCLC has steadily increased. At the same time, the survival rates have improved due to advances in anticancer treatments, resulting in increased numbers of primary and follow-up imaging for assessment of BMs. The evaluation of these scans is tiresome and bears an inherent risk of missed diagnosis, in particular for subtle lesions. In this context, the proposed DLM can provide assistance to the physicians by automated detection of BMs leading to a reduced workload and counteract human errors that result from fatigue and the satisfaction of search effect phenomenon.

A large number of recently proposed deep learning algorithms enabling automated detection of BMs cannot generate segmentations of tumor tissue. In contrast, the DLM of the present study provides 3D voxel-wise segmentations of BMs with a performance (DSC of 0.72), which is in line with the present literature (0.67–0.79) and comparable to DL-based segmentation of larger malignant tumors, eg, glioblastoma (0.62–0.86) or primary central nervous system lymphoma (0.73–0.76). In this context, automated segmentation of BMs might be useful for lesion contouring in stereotactic radiosurgery, leading to a reduction of time effort as well as of intra- and inter-rater variabilities. Further, the automated segmentations may facilitate the accurate assessment of tumor burden by providing volumetric data of BMs, which are superior to conventional linear measurements of lesions since these are not entirely spherical in shape. Thereby, automated segmentations have the potential to facilitate treatment decisions and to improve the outcome of the patient.

Automated detection and segmentation of BMs in NSCLC pose the following challenges: (I) multifocality, (II) heterogeneity of BMs in terms of size and appearance due to the underlying mutation of the primary tumor, its stage, previously administered treatments, and (III) inhomogeneous imaging data. In this context, the majority of recent studies investigating deep learning-based detection of BMs used homogenous imaging data, mostly a standardized protocol consisting of a distinct 3D T1CE sequence at a single institution for planning of radiosurgery, which limits their generalizability and questions the usefulness in clinical routine. In contrast, the DLM applied in the present study provided a high detection sensitivity on heterogeneous “real-life” imaging data acquired on scanners from different vendors, generations, and study centers with resulting divergent scan parameters and unstandardized application of contrast media.

The findings of this study indicate that the application of the DeepMedic network, which was originally developed for other neuro-oncological entities (gliomas), may prove to be successful on BMs, but preceding training for the individual primary tumor is mandatory.

In the past, the combined approach of radiomics and machine learning based on manual segmentations of BMs in NSCLC has shown potential for the evaluation of the epidermal growth factor receptor status of the primary tumor, which may provide a clinical benefit since the application of targeted therapy in these patients has shown promising results. However, the manual segmentation of tumor tissue for radiomics analysis in aforementioned studies suffers from a low reproducibility as well as inter- and intra-rater...
variabilities.\textsuperscript{15,16} In this context, feature extraction from automated segmentations of BMs might lead to an increased robustness of tumor characterization and should be investigated further.

**Limitations**

Given its retrospective design, an evaluation of whether detection and segmentation performances are sufficient for clinical needs was not possible. Additionally, this study neither investigated the potential long-term effects of detection by human vs. human and artificial intelligence nor if the DLM results influence the treatment of patients. However, a short processing time in combination with a high sensitivity and a low number of FP detections may facilitate clinical implementation and the conduction of prospective studies assessing the direct influence on patient care. It should also be noted that while this study included imaging data from referring institutions, it does not approximate to a true multicenter

| Table 7. Detection Performance of the Fused DLM in the Test Set Stratified by MRI Specifics and Localization of Brain Metastases (BM) |
| Detected/Missed BMs | Sensitivity (%) |
|---------------------|----------------|
| **Examination location** | |
| Local \((n = 8)\) | 23/1 | 95.8 |
| Referring \((n = 9)\) | 34/9 | 79.1 |
| **Field strength** | |
| 1.0 T \((n = 1)\) | 1/0 | 100.0 |
| 1.5 T \((n = 11)\) | 38/10 | 79.2 |
| 3.0 T \((n = 5)\) | 18/0 | 100.0 |
| **Localization of BMs** | |
| Parenchymal | 50/7 | 87.7 |
| Pachymeningeal | 7/3 | 70.0 |
| Supratentorial | 36/5 | 90.0 |
| Infratentorial | 21/5 | 80.8 |
approach. Besides, after excluding 8% of scans for the abovementioned criteria (e.g., insufficient administration of contrast agent), the true number of patients not suitable for the application of the DLM may be underestimated. Additionally, the small number of patients and BMs in the test set limits the generalizability of the presented results, including the performed subanalysis of BM detection. Further, in this study, only MRI data at primary diagnosis of the BMs were included with unknown performance of the DLM after therapy. However, the DeepMedic architecture has already shown its potential for application in longitudinal tumor imaging in primary central nervous system lymphoma.20 Finally, compared with other deep learning based approaches to automatically detect BMs,25,30 the DeepMedic network requires multiparametric MRI datasets (FLAIR, T1-/T2-weighted, and T1CE). Nevertheless, these sequences are part of clinical MRI tumor assessment across most institutions.

Conclusion

Despite their small size, the proposed DLM detects BMs in NSCLC with high detection sensitivity at a low false positive rate while obtaining a good segmentation performance.

Acknowledgments

Clinician Scientist position supported by the Deans Office, Faculty of Medicine, University of Cologne.

Conflict of Interest

Frank Thiele, Michael Perkuhn, and Rahil Shahzad: Employees, Philips Healthcare. Jan Borggrefe: Speaker’s bureau, Philips Healthcare. Lenhard Pennig: Research
support, Philips Healthcare. The remaining authors disclose no relevant relationships.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61(2):69-90.

2. Kromer C, Xu J, Ostrom QT, et al. Estimating the annual frequency of synchronous brain metastasis in the United States 2010–2013: A population-based study. J Neurooncol 2017;134(1):55-64.

3. Nolan C, Deangelis LM. Overview of metastatic disease of the central nervous system. Handb Clin Neurol 2018;149:3-23.

4. Taboure E, Chinot O, Metellus P, Tallet A, Viens P, Gonçalves A. Recent trends in epidemiology of brain metastases: An overview. Anti-cancer Res 2012;32(1):4655-4662.

5. Postmus PE, Kerr KM, Oudkerk M, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 2017;28(suppl_4): iv1-iv21.

6. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. Radiographics 2015;35(6):1668-1676.

7. Jena A, Taneja S, Talwar V, Sharma JB. Magnetic resonance (MR) patterns of brain metastasis in lung cancer patients: Correlation of imaging findings with symptom. J Thorac Oncol 2008;3(2):140-144.

8. Berbaum KS, Franken EA, Dorfman DD, et al. Satisfaction in search in diagnostic radiology. Invest Radiol 1990;25(2):133-140.

9. Brady AP. Error and discrepancy in radiology: Inevitable or avoidable? Insights Imaging 2017;8(1):171-182.

10. Yamamoto M, Serizawa T, Shuto T, et al. Stereotactic radiosurgery for brain metastases to the brain on MRI. J Magn Reson Imaging 2020;52(4):1227-1236.

11. Mok TS, Wu Y-L, Ahn M-J, et al. Osimertinib or platinum-based chemotherapy for non-small-cell lung cancer (NSCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 2017;28(suppl_4): iv1-iv21.

12. Sankey EW, Tsvankin V, Grabowski MM, et al. Operative and peri-operative considerations in the management of brain metastasis. Cancer Med 2019;5(3):90.

13. Mok TS, Wu Y-L, Ahn M-J, et al. Osimertinib or platinum–Pemetrexed in EGFR T790M-positive lung cancer. N Engl J Med 2017;376(7):629-640.

14. Hanna GG, Hounsell AR, O Neill AE, et al. Computer-aided detection and segmentation of brain metastases to the brain on CTA using deep learning. Sci Rep 2020;10(1):21799.

15. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning enables automatic detection and segmentation of brain metastases on multisegmentation MRI. J Magn Reson Imaging 2019;51(1):175-182.

16. Grevik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisegmentation MRI. J Magn Reson Imaging 2019;51(1):175-182.

17. Dikici E, Ryu JL, Demirer M, et al. Automated brain metastases detection framework for T1-weighted contrast-enhanced 3D MRI. IEEE J Biomed Heal Informatics 2020;24(10):2883-2893.

18. Deike-Hofmann K, Vawbcke W, Pach D, et al. Pre-examinations improve automated metastases detection on cranial MRI. Invest Radiol 2020;56(5):320-327.

19. Noguchi T, Uchiyama F, Kawata Y, et al. A fundamental study assessing the diagnostic performance of deep learning for a brain metastasis detection task. Magn Reson Med Sci 2019;19(3):184-194.

20. Zhou Z, Sanders JW, Johnson JM, et al. Computer-aided detection of brain metastases in T1-weighted MRI for stereotactic radiosurgery using deep learning single-shot detectors. Radiology 2020;295(2):407-415.

21. 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(4):1227-1236.

22. Pennig L, Shahzad R, Caldeira L, et al. Fully automated detection and segmentation of brain metastases on contrast-enhanced T1-weighted image radiomics of brain metastases may predict EGFR mutation status in primary lung cancer. Sci Rep 2020;10(1):8905.