A Fundamental Study Assessing the Diagnostic Performance of Deep Learning for a Brain Metastasis Detection Task

Tomoyuki Noguchi1,2,3,4*, Fumiya Uchiyama1, Yusuke Kawata1, Akihiro Machitori1, Yoshitaka Shida1, Takashi Okafuji1, Kota Yokoyama1, Yosuke Inaba5, and Tsuyoshi Tajima1

Purpose: Increased use of deep convolutional neural networks (DCNNs) in medical imaging diagnosis requires determinate evaluation of diagnostic performance. We performed the fundamental investigation of diagnostic performance of DCNNs using the detection task of brain metastasis.

Methods: We retrospectively investigated AlexNet and GoogLeNet using 3117 positive and 37961 negative MRI images with and without metastasis regarding (1) diagnostic biases, (2) the optimal \( K \) number of \( K \)-fold cross validations (\( K \)-CVs), (3) the optimal positive versus negative image ratio, (4) the accuracy improvement curves, (5) the accuracy range prediction by the bootstrap method, and (6) metastatic lesion detection by regions with CNNs (R-CNNs).

Results: Respectively, AlexNet and GoogLeNet had (1) 50 ± 4.6% and 50 ± 4.9% of the maximal mean ± 95% confidence intervals (95% CIs) measured with equal-sized negative versus negative image datasets and positive versus positive image datasets, (2) no less than 10 and 4 of \( K \) number in \( K \)-CVs fell within the respective maximum biases of 4.6% or 4.9%, (3) 74% of the highest accuracy with equal positive versus negative image ratio dataset and 91% of that with four times of negative-to-positive image ratio dataset, (4) the accuracy improvement curves increasing from 69% to 74% and 73% to 88% as positive versus negative pairs of the training images increased from 500 to 2495, (5) at least nine and six out of 10-CV result sets essential to predict the accuracy ranges by the bootstrap method, and (6) 50% and 45% of metastatic lesion detection accuracies by R-CNNs.

Conclusions: Our research presented methodological fundamentals to evaluate diagnostic features in the visual recognition of DCNNs. Our series will help to conduct the accuracy investigation of computer diagnosis in medical imaging.

Keywords: bias, brain neoplasms, learning curve, magnetic resonance imaging, neural networks (computer)
systems using DCNNs also requires the establishment of the methodology for evaluating the diagnostic performance.

We therefore performed basic verification experiments to investigate the diagnostic features of DCNNs using the detection task of the brain metastasis on MRI images.

Materials and Methods

Study design

This study was approved by our hospital’s Institutional Review Board, which waived the need for written informed consent from the patients.

Patients

As the brain metastases group, we enrolled 162 of consecutive 164 patients with brain metastases located in parenchyma or pia mater who underwent a contrast-enhanced three-dimensional T₁-weighted images with fat suppression in the period from April 2014 to November 2017 (male/female, 92/70; age range/average, 37–90/68.5 years). Their primary malignant neoplasms were lung cancer in 134 patients, breast cancer in 11, digestive cancer in 10, renal cancer in two, and one patient each with myeloma, breast lymphoma, mesothelioma, pheochromocytoma, and primary unknown. The other two patients were excluded because of suboptimal MRI results due to movement artifact.

As the control group, we selected 282 of consecutive 301 patients undergoing head MRI enrolled backwardly from November 2017 to October 2016 and being free from brain metastasis for at least subsequent 6 months confirmed on the electric medical chart (male/female, 163/119; age range/average, 20–91/69.4 years). We enrolled them so that the total slice number of their MRI images was 10 times as that of patients with brain metastases to match the Test 3 plus about two times of those allowing for flexibility of data analysis. The other 19 patients were excluded because of metal artifacts due to old brain vascular diseases in eight patients, dental instrumentation in four, and one patient each with suspected convexity meningioma, suspected acoustic schwannoma, suspected trigeminal schwannoma, basilar aneurysm, parietal venous malformation observed on MRI images.

Magnetic resonance imaging

We investigated MRI examinations performed on the three MRI units at our hospital, a MAGNETOM Avanto (Siemens AG, Erlangen, Germany), an MRT200SP5 (Canon Medical Co. Ltd., Otawara, Japan), and a MAGNETOM Verio (Siemens AG). We allowed to use MRI images obtained on three different MRI units because DCNNs should be trained so as to be tolerable for the difference in the image quality among MRI units. We used contrast-enhanced three-dimensional T₁-weighted images with fat suppression under following parameters which were optimized for MRI units in each. The detailed parameters were as followed: TR/TE (ms) = 3.92–15/1.44–4.19; flip angle (°) = 10–18; slice thickness-spacing between slices (mm) = 1.0–2.0/1.0–1.0; matrix = 188–256; pixel spacing (mm) = 0.5–1.0; number of slices = 144–288. The gadolinium (Gd) contrast agents that were intravenously injected with optimized volumes depending on the patient’s conditions were 0.1 mmol/kg Gd-DOTA-Butrol, 0.1 mmol/kg Gd-DTPA, or 0.1 or 0.2 mmol/kg Gd-HP-DOTA.

Image data processing

We tested two classifiers with DCNN architecture, AlexNet and GoogLeNet, both of which were already trained on more than a million images from the ImageNet database (ImageNet. http://www.image-net.org) and distributed as add-on software of MATLAB (ver. 2017b; MathWorks Inc., Natick, MA, USA). We used a custom-build image processing computer (TEGARA Corporation, Shizuoka, Japan) containing a Quadro P2000 5 GB graphics processing unit (Nvidia Corporation, Santa Clara, CA, USA), an Intel Xeon E5-2680v4 2.40 GHz processor (Intel Corporation, Santa Clara, CA, USA), 1.0 TB of hard disk space, and 64 GB of RAM.

We chose 3117 positive images which included any cross-sections of enhanced tumor lesions regardless of its locations or lesion numbers extracted from the brain metastases group (Fig. 1). We allowed to use multiple slices of the same subject which may be higher correlation within the mutual slices, based on the following reasons: (1) It was difficult to select the most appropriate slice out of the whole brain images in patients with multiple brain metastases. (2) We did not adopt the data augmentation method in which the original image was rotated, cropped, or deformed differently, to increase the data size for training DCNNs allowing to have higher correlation between the original image and the augmented images. Instead, we expected multiple slices of the same subject having the favorable effect similar to the data augmentation method.

We extracted 37961 negative images including whole brain parenchyma from the control group which were about 12 times of the number of positive images as mentioned above. AlexNet and GoogLeNet used in the current study did not support Digital Imaging and Communications in Medicine (DICOM) format. Therefore, using the free software program Fiji (https://imagej.net/Fiji), we converted these images from DICOM format into two types of image formats with image sizes supported by DCNNs: Portable Network Graphics (PNG) with 16-bit with 227 × 227 pixels for use in AlexNet, and Joint Photographic Experts Group (JPEG) format with 224 × 224 pixels for use in GoogLeNet. In the processing of converting the image format, we adjusted the brightness and contrast of the original DICOM image using a function of look-up-table equipped in Fiji.

Two radiologists (F.U. and T.N.) extracted the negative and positive images and drew 9684 of free-sized bounding boxes around the metastatic lesions over all 3117 of the positive images using the annotation application running on MATLAB.

Since the CNNs pretrained with non-medical images could serve as an effective image recognition baseline for
classification of the medical images, we used the transfer learning method as follows. In both AlexNet and GoogLeNet, we separately displaced the last three fully-connected layers with three new layers consisting of a fully-connected layer, a softmax layer, and a classification output layer. The weights of the learning rate factor and the weight of the bias learning rate factor in the fully-connected layer were set to 20 and 20, respectively. The classification output layer had two outputs corresponding to the positive or negative categories. We used 0.0001 for the initial learn rate, 10 for the mini-batch size, and 100 for the max epochs as the parameters of the training options. The imaging data augmentation method was not used.

Statistical and data analysis
We performed the following tests.

Test (1): Bias in the diagnostic performance of the DCNNs
Before measuring the diagnostic ability to distinguish the positive images from the negative images, we investigated the potential bias of the diagnostic performance in DCNNs. We first prepared a set of the image data consisting of a pair of the 1558 equal-sized randomly-assigned negative images. Namely, this dataset was composed of two subsets in both of which only negative images were assigned. We then performed a standard K-fold cross-validation test (K-CV) under K numbers ranging from 3 to 20 as follows: (1) splitting the dataset into K equal parts, (2) setting the ratio of the data as (1/K), (1/K), and (1–2/K) for the testing, validation, and training image numbers, respectively, (3) training DCNNs with the training images, (4) validating DCNNs with the validation images to prevent the over training, (5) testing DCNNs with the testing images to evaluate the judgment ability, (6) repeating K times, each selecting a different testing dataset, and (7) finally determining the mean accuracy by averaging the K test results and 95% confidence interval (95% CI) by calculating the K test results with t-distributions. Since the mean accuracy should be 50% in the current tests, the deviation of the mean accuracy ± 95% CI from 50% could indicate the potential bias of the diagnostic performance of DCNNs.

In the same way, a set of the image data consisting of a pair of the 1558 equal-sized randomly-assigned positive images was subjected to the same investigation.

Test (2): Variation of K number in the K-fold cross-validation method
As K number in K-CVs increases, the reliability of the accuracy improves, and the accuracy itself also escalates due to the increase in the image number of training data. For the assessment of the variance of the diagnostic accuracy, we performed standard K-CVs using 3117 pairs of positive and negative images under K numbers were set from 3 to 20. To reduce K number effects in K-CVs, we also performed modified K-CVs in which the training images were fixed to 1000 pairs of positive and negative images in the ratio of (1/K), (1/K), and (1–2/K) for the testing, validation, and training image numbers. We also measured the mean accuracy ± 95% CI in each of K-CVs and determined the optimal K-value in light of the bias obtained in Test 1.

Test (3): The negative-to-positive image data ratio
A difference in the negative-to-positive image data ratio can affect the diagnostic performance. We prepared 10 datasets of the images consisting of the positive images with the number of constant 3000 and negative images with the number of 750, 1000, 1500, 2000, 3000, 6000, 12000, 18000, 24000 or 30000 in a random extraction method. Namely, the numbers of negative images were assigned at one-fourth, one-third, one-half, two-thirds, one, two, four, six, eight, and 10 times of the positive images. We tested these 10 datasets by 10-CV with a modified testing procedure in which the 300 pairs of negative and positive images were constantly assigned to each of the validation images and the testing images instead of using the ratio of (1/K), (1/K), and (1–2/K) for the testing, validation, and training image numbers, respectively, to avoid
the inconsistency of the number of significant figures in the results. We here adopted $K$-CV with 10 of $K$ number based on the result of the optimal $K$ number investigation determined in Test 2. We measured mean accuracies, area under the curves (AUCs), sensitivities, and specificities.

**Test (4): The accuracy improvement curves**

As the number of the training images to train DCNNs increases, the mean accuracy obtained with 10-CV increases. Although it is difficult to estimate the amount of the training images enough to get the best performance of DCNNs, we might judge whether or not the current number of the training images is sufficient to fully demonstrate the ability of DCNNs. We prepared five datasets of 624, 1250, 1832, 2362, 3117 pairs of the randomly extracted positive and negative images, which were adjusted to meet 500, 1000, 1500, 2000, and 2495 pairs of the training images with the ratio of $(1/K)$, $(1/K)$, and $(1–2/K)$ for the testing, verification, and training image numbers for each of the datasets. We then performed 10-CVs to obtain the mean accuracies. Finally, we plotted them to obtain the accuracy improvement curves. Based on the result in Test 2, we adopted the number of the training images as the index of increase of the dataset and K-CV with 10 of $K$ number. Here, the results of the paired 1000 and 2495 training images by 10-CV were reused from those of standard 10-CVs and modified 10-CVs in Test 2, respectively.

**Test (5): Accuracy prediction by the bootstrap method**

$K$-CV is to repeat $K$ times to train and test DCNNs, which needs the long-term possession of the computational resources. If some of those operations can be skipped, it may save the resource-consuming process in $K$-CV. Here, we explored the time- and resource-saving validation method for 10-CV.

We created nine combinational groups comprising $10C_1$, $10C_2$, $10C_3$, …, and $10C_9$ (i.e., 10, 45, 120, …, and 10) datasets out of 10 result sets obtained in Test 2, and those simple combinational groups were hereinafter referred to as $10C_1$, $10C_2$, $10C_3$, …, and $10C_9$ groups, respectively. We then calculated 95% CI of each of the groups by the bootstrap method. When given $10C_N$ group had no $< 95\%$ of the number rates of 95% CI ranges including the mean accuracy determined by 10-CV in Test 2, given $N$ datasets out of 10 results sets were judged to predict the range of the accuracy for 10-CV.

Meanwhile, given $N$ datasets out of 10 results sets can be propagated to $N$ subgroups comprising $NC_1$, $NC_2$, …, and $NC_N$ of datasets. We first calculated the accuracy in each of those subgroups by the bootstrap method, and next sorted those subgroups by the accuracy, and then extracted the median numbers from the subgroups. We finally replaced the initial $10C_1$, $10C_2$, $10C_3$, …, and $10C_9$ datasets with those median numbers, respectively. Those propagated combinational groups were hereinafter called as the median $10C_1$, median $10C_2$, …, and median $10C_9$ groups, respectively. Then, we calculated 95% CIs of those groups by the bootstrap method and scored their number rates as mentioned above.

We also compared the mean processing times per one operation of 10-CV to those per one range of 95% CIs obtained from the simple and propagated combinational groups by the bootstrap method for AlexNet and GoogLeNet, respectively, to confirm if the bootstrap method could reduce the processing time.

In the calculation of the accuracies and 95% CIs by the bootstrap method in the current test, the bias-corrected and accelerated bootstrap interval method was adopted.

**Test (6): Metastatic lesion detection by R-CNN**

Regions with CNN (R-CNN) is one of the applications derived from DCNN and proposes a ROI. We assessed the metastatic lesion detection abilities of AlexNet- and GoogLeNet-mounted-R-CNNs with 10 result sets determined in Test 2 by 10-CV.

Because we could not define the ratio of validation and training image numbers which R-CNNs automatically handled, the positive training and validation images were used for training R-CNNs in each of those datasets in conjunction with coordinate data of the true bounding boxes as mentioned above. The positive and negative testing images were used for the evaluation of the metastatic lesion detection ability of R-CNNs. Namely, the ratio of $(1/K)$ and $(1–1/K)$ for the testing and training image numbers was adopted for each of the datasets.

We performed two types of tests; the metastatic lesion detection tests of R-CNNs and the metastatic lesion detection tests in conjunction with the primary selection test. In the latter, according to the judgement of the primary selection test adopted from the results obtained in Test 2, the positively judged testing images were further evaluated by R-CNNs, and the negatively judged testing images were determined to have no bounding box predicted by R-CNNs.

For the positive testing images, we scored the number rates of the true bounding box covered by the entire part of the entire predictive bounding boxes out of the number of the true bounding box to avoid missing brain metastases in spite of the considerable increase in false positives. For the negative testing image, zero or one out of one point was scored if any or no predictive bounding boxes were proposed by R-CNNs, respectively. We adopted this scoring system in which the higher scores were obtained the more bounding boxes were placed because we gave greater importance to the sensitivity for the metastatic lesion detection than the specificity and R-CNNs proposed the bounding boxes according to their computational calculation. Representative judgement examples were shown in Fig. 2.

**Results**

**Diagnostic bias in the DCNNs**

Figure 3 illustrated the results of diagnostic bias in AlexNet and GoogLeNet, respectively. The mean accuracies ± 95% CIs in classifying the equal-sized negative image dataset held within 50 ± 3.9% in AlexNet and 50 ± 4.9% in GoogLeNet in...
**Fig. 2** Representative examples of the judgement of lesion detection by regions with convolutional neural network (R-CNN) in Test (6). The two yellow bounding boxes (1 and 2) in a positive image indicate two lesions of brain metastases (a), and the two of green areas in the same positive image reflect the predicted lesions by R-CNN (b). Bounding box (1) is covered by one of the two green areas, which gets 1 of 1 point. Bounding box (2) is not covered by either of green areas, which gets 0 of 1 point. In the negative testing image (c), 0 out of 1 point was scored although two predictive green areas were proposed by R-CNN (d).

**Fig. 3** The variation of mean accuracy in classification of the equal-sized negative (a) and positive image datasets (b) judged by AlexNet (Δ) and GoogLeNet (•). The maximal biases from the true error rate of 50% in negative and positive image datasets are 4.6% in AlexNet with the positive image dataset and 4.9% in GoogLeNet with the negative image dataset. CI, confidence interval.

any K numbers of K-CVs (Fig. 3a). The mean accuracies ± 95% CIs in classifying the positive image dataset were within 50 ± 4.6% in AlexNet and 50 ± 4.4% in GoogLeNet compared to those in the positive image dataset (Fig. 3b).

**The optimal K number in K-fold cross-validation method**

Figure 4 illustrated the mean accuracies and 95% CIs of K-CVs in AlexNet and GoogLeNet. When K number was increased from 3 to 20, AlexNet showed that the mean accuracy of standard K-CVs increased from 69% to 75% ($y = 1.15 \times 10^{-3}x + 0.73$ in single regression analysis), while it demonstrated that the mean accuracy of modified K-CVs with 1000-fixed training images changed less from 73% to 72% ($y = -3.60 \times 10^{-3}x + 0.72$). GoogLeNet showed that the mean accuracy increased from 78% to 88% in standard K-CVs ($y = -5.97 \times 10^{-3}x + 0.83$) and held from 80% to 80% in modified K-CVs ($y = -0.170 \times 10^{-3}x + 0.80$) (Fig. 4a).

95% CIs of standard and modified K-CVs for AlexNet and GoogLeNet tended to diminish as K number increased (Fig. 4b). K-values with 10 or more in AlexNet or four or more in GoogLeNet had 95% CIs of K-CVs that fell within the respective maximum bias of 4.6% or 4.9% determined in Test 1.

**Ratio of the negative-to-positive image dataset**

The results of the variation of statistical values in the ratio of the negative-to-positive image dataset in AlexNet and GoogLeNet were provided in Fig. 4. As the negative image data increased,
A Fundamental Assessment of Deep Learning

the specificity increased but the sensitivity decreased, and then they intersected at the boundary around the same size of the negative images at which mean accuracies, AUCs, sensitivities, and specificities were 74%, 0.841, 70%, and 78% in AlexNet, and 87%, 0.961, 90%, and 84%, in GoogLeNet, respectively.

Afterward, AlexNet showed the decrease of the mean accuracy, sensitivity, and mean AUC, but not specificity (Fig. 5a). GoogLeNet demonstrated 91%, 0.983, 82%, and 99%, respectively, at four times of negative-to-positive image ratio dataset with the highest mean accuracy and AUC (Fig. 5b).

**Fig. 4** The mean accuracies (a) and 95% confidence intervals (95% CIs) (b) of K-fold cross-validation tests (K-CVs) in AlexNet and GoogLeNet. AlexNet (a: Δ with a black line) and GoogLeNet (a: • with a black line) show that the mean accuracy of standard K-CVs increase, while those of modified K-CVs with 1000-fixed training images less change (a: Δ with a gray line for AlexNet and • with a gray line for GoogLeNet), as K number increases from 3 to 20. Meanwhile, 95% CIs of standard and modified K-CVs for AlexNet (b: Δ with respective black and gray lines) and GoogLeNet (b: • with respective black and gray lines) tend to diminish as K number increases. AlexNet and GoogLeNet have no less than 10 and four of K number in K-CVs fell within the respective maximum biases of 4.6% or 4.9%.

**Fig. 5** The variation of the statistical values of 10 datasets of the image data consisting of the number of negative images randomly assigned at one-fourth, one-third, one-half, two-thirds, one, two, four, six, eight, and 10 times of 3000 positive images judged by AlexNet (a) and GoogLeNet (b). As the negative image data increases, the specificity increases but the sensitivity decreases, and then they intersect at the boundary around the same size of the negative images. Afterward, GoogLeNet demonstrates the highest mean accuracy and AUC at four times of negative-to-positive image ratio dataset. AUC, area under the curve.
The accuracy improvement curve

Figure 6 showed the results of the accuracy improvement curves in 1-CV in AlexNet and GoogLeNet. When the number of training images were changed to 500, 1000, 1500, 2000, and 2495, AlexNet showed the increased mean accuracies of 69%, 72%, 70%, 78%, and 74% with the increasing rate of 3.3% per a positive versus negative pair of 1000 training images \( (y = 3.33 \times 10^{-5}x + 0.67) \). GoogLeNet demonstrated the higher mean accuracies of 73%, 80%, 83%, 86%, and 88%, in conjunction with a higher increasing rate of 8.3% per a positive versus negative pair of 1000 training images \( (y = 8.31 \times 10^{-5}x + 0.69) \) compared with AlexNet.

Accuracy prediction by the bootstrap method

Table 1 summarized the results of the range prediction of the mean accuracy of 10-CV determined in Test 2 in AlexNet and GoogLeNet using the bootstrap method. None of the simple combinational groups in AlexNet or GoogLeNet reached more than 95% of the number rates of 95% CI ranges including the mean accuracy of 10-CV in Test 2. In the propagated combinational groups, however, the median \( 10C_{1-9} \) group in AlexNet and the median \( 10C_{1-6}, 10C_{1-7}, 10C_{1-8}, \) and \( 10C_{1-9} \) groups in GoogLeNet showed more than 95% of the number rates.

About 393 and 3502 s of the mean processing times per one operation of 10-CV for AlexNet and GoogLeNet, respectively, were longer than 0.3 and 0.4 s or 125 and 166 s of the maximum of the mean processing times per one range of 95% CIs obtained from simple combinational groups for AlexNet and GoogLeNet, respectively.

Metastatic lesion detection by R-CNN

Table 2 provided the results of the metastatic lesion detection by R-CNN in AlexNet and GoogLeNet. The best metastatic lesion detection ability was observed in GoogLeNet-mounted-R-CNN with the primary selection test by GoogLeNet at which the mean accuracy, sensitivity, and specificity, positive predictive value, and negative predictive value were 50%, 28%, 95%, 92%, and 39%, respectively. The second-best lesion detection ability was by the AlexNet-mounted-R-CNN with the primary selection test by AlexNet having 45%, 27%, 83%, 77%, and 35%, respectively. The detectability performance of AlexNet-mounted-R-CNNs without the primary selection test showed 31%, 33%, 29%, 49%, and 17%, respectively, which was similar to that of the GoogLeNet-mounted-R-CNN without the primary selection test with 31%, 29%, 35%, 48%, and 19%, respectively.

Discussion

We performed the fundamental investigations of DCNNs regarding diagnostic biases, the optimal negative-to-positive image ratio, the optimal \( K \) number of \( K \)-fold cross validation, the accuracy improvement curve, the accuracy range prediction by the bootstrap method, and lesion detection by R-CNNs using the detection task of brain metastasis on MRI images.

Diagnostic bias in the DCNNs

Ideally, in the testing for diagnostic bias in DCNNs using an equal-sized positive versus positive or negative versus negative image dataset, 50% accuracy should be established. In practice, however, the maximum error rates of 4.6% in AlexNet and 4.9% in GoogLeNet were derived in our datasets, regardless of \( K \) number in \( K \)-CVs. Although these error rates might depend on the features of the image datasets, such potential biases should be taken into account when the diagnostic performance by DCNNs is investigated.

The optimal \( K \) number in \( K \)-fold cross-validation method

Both AlexNet and GoogLeNet showed increased mean accuracies with standard \( K \)-CVs as \( K \) number increased from 3 to 20. This is because the number of training images assigned a ratio of \((1–2/K)\) increases as \( K \) number increases. In practice, the number of training images increased from a positive versus negative pair of 1039–2807 images as the \( K \) increased from 3 to 20. In addition, the results in modified \( K \)-CVs with 1000-fixed training images showed no pronounced fluctuation in accuracy. This result suggested that the number of training images as well as sample images should be taken into account when the diagnostic performances of different DCNNs are compared.

The optimal \( K \) number for \( K \)-CV has been discussed for a long time. In almost cases, five- or 10-CVs were used to validate the machine learning only because these work well
in practice, although $K$-values are arbitrary. When we applied the maximum range of 4.6% in AlexNet and 4.9% in GoogLeNet revealed by Test (1) as the quality limitation, the optimal $K$ number of 10 or more in AlexNet and four or more in GoogLeNet were acceptable in our study.

### Ratio of the negative-to-positive image dataset

Imbalanced datasets are one of the common but essential problems in real-world settings. Although there are many ways to solve imbalanced dataset problems, the effects and tendencies of dataset balance should be investigated. In our present testing for the effect of the diagnostic performance based on the ratio of the negative-to-positive image dataset, the specificity and sensitivity curves intersected at the boundary around the same size of the negative images in both AlexNet and GoogLeNet, as the negative image data increased. Interestingly, GoogLeNet demonstrated the highest mean accuracy at four times of negative-to-positive image ratio dataset. GoogLeNet might gain diagnostic performance power from the sample number rather than the balance of the samples under experimental conditions. GoogLeNet might have unknown mechanisms extracting the abnormal features from the images that we consider normal, which could be less susceptible to the modulation of diagnostic performance by the imbalanced datasets. Anyway, our results suggested that equal-sized positive versus negative image data would be desirable to obtain statistical results in a well-balanced manner although there might be some kinds of variation in statistical values depending on characteristic of DCNNs or datasets, or both.

### The accuracy improvement curve

In the investigation of the accuracy improvement curves, AlexNet and GoogLeNet demonstrated the increasing rates

### Table 1  The range of mean accuracy prediction by the bootstrap method

| Type of group       | AlexNet                      | GoogLeNet                    |
|---------------------|------------------------------|------------------------------|
|                     | Number rate of 95% CI ranges including the mean accuracy of 10-CV (%) | Mean processing time to obtain one of 95% CIs (s) |
|                     |                              |                              |
| Simple combinational group |                              |                              |
| $10C_1$             | 30 (3/10)                    | 0.30                         |
| $10C_2$             | 44 (20/45)                   | 0.21                         |
| $10C_3$             | 44 (53/120)                  | 0.18                         |
| $10C_4$             | 49 (103/210)                 | 0.17                         |
| $10C_5$             | 50 (125/252)                 | 0.18                         |
| $10C_6$             | 59 (123/210)                 | 0.19                         |
| $10C_7$             | 67 (80/120)                  | 0.21                         |
| $10C_8$             | 67 (30/45)                   | 0.23                         |
| $10C_9$             | 60 (6/10)                    | 0.25                         |

| Propagated combinational group |                              |                              |
| Median $10C_1$             | 30 (3/10)                    | 0.30                         |
| Median $10C_1$–2           | 44 (20/45)                   | 0.64                         |
| Median $10C_1$–3           | 58 (69/120)                  | 1.27                         |
| Median $10C_1$–4           | 64 (134/210)                 | 2.59                         |
| Median $10C_1$–5           | 69 (174/252)                 | 5.51                         |
| Median $10C_1$–6           | 73 (153/210)                 | 12.07                        |
| Median $10C_1$–7           | 83 (100/120)                 | 26.77                        |
| Median $10C_1$–8           | 87 (39/45)                   | 58.82                        |
| Median $10C_1$–9           | 100 (10/10)                  | 125.30                       |

|                | Mean processing time to obtain one of 95% CIs (s) |
|----------------|--------------------------------------------------|
| Simple combinational group |                              |
| $10C_1$             | 0.30                              |
| $10C_2$             | 0.21                              |
| $10C_3$             | 0.18                              |
| $10C_4$             | 0.17                              |
| $10C_5$             | 0.18                              |
| $10C_6$             | 0.19                              |
| $10C_7$             | 0.21                              |
| $10C_8$             | 0.23                              |
| $10C_9$             | 0.25                              |

| Propagated combinational group |                              |
| Median $10C_1$             | 0.30                              |
| Median $10C_1$–2           | 0.64                              |
| Median $10C_1$–3           | 1.27                              |
| Median $10C_1$–4           | 2.59                              |
| Median $10C_1$–5           | 5.51                              |
| Median $10C_1$–6           | 12.07                             |
| Median $10C_1$–7           | 26.77                             |
| Median $10C_1$–8           | 58.82                             |
| Median $10C_1$–9           | 125.30                            |

CV, cross validation; CI, confidence interval.

### Table 2  Lesion detection by R-CNNs

| Statistical parameter | R-CNN (%) | R-CNN with primary screening test by DCNN (%) |
|-----------------------|-----------|---------------------------------------------|
|                       | AlexNet   | GoogLeNet                                   |
|                       | AlexNet   | GoogLeNet                                   |

| Accuracy               | 31        | 31                                           |
| Sensitivity            | 33        | 33                                           |
| Specificity            | 29        | 35                                           |
| PPV                    | 49        | 48                                           |
| NPV                    | 17        | 35                                           |
of 3.3% and 8.3% per a positive versus negative pair of 1000 training images, respectively, which were calculated from the single regression analyses. When 98% or more of the accuracy is needed, more than 9309 and 3490 pairs of training images in AlexNet and GoogLeNet, respectively, are needed according to the regression equations. Since the hypothetical accuracy improvement curve does not necessarily show a linear relationship,13 those increasing rates cannot be directly applied to the prediction of the sample sizes. However, our testing procedure might be generally used to judge whether or not the current sample size is enough to fully demonstrate the ability of DCNNs.

Accuracy prediction by the bootstrap method
Although K-CV is the most widely used method for estimating prediction error,26 the use of K-CV consumes a considerable amount of time and computational resources. The bootstrap method can be used instead of estimation based on parametric assumption in cases where the hypothesized distribution is suspicious or parametric assumption is impossible or requires very complicated calculation.20,27–29 Our result using the bootstrap method in conjunction with the data extraction and propagation procedure might suggest a solution of the disadvantages of K-CV. That is, the range of the mean accuracy calculated by 10-CV can be predicted using nine in AlexNet or six in GoogLeNet, respectively, out of 10 result sets in 10-CV with a probability of 95% or more in the current study. We are planning further investigations to explore its theoretical background.

Metastatic lesion detection by R-CNN
Although an R-CNN with additional auxiliary system as the primary selection test might be effective to improve the ability of the metastatic lesion detection with 50% of the best accuracy, R-CNN used in our present series showed room for improvement. Even 9684 bounding boxes might be insufficient for training an R-CNN to perform effectively. In addition, the evolution of artificial intelligence (AI) algorithms is very rapid32–35 and we can expect the future development of novel lesion detection systems. Especially, as shown in Fig. 2d, some improper bounding boxes were placed the extra-axial structures such as vessels, ocular globes, and so on. The appropriate preparation processes such as the skull-stripping might improve the performance of DCNNs.

Limitations
Our study has several limitations. We did not use data augmentation methods and, therefore, the amount of image data was insufficient to derive the full performance of both DCNNs. It is unconfirmed whether the transfer learning method used in the current study is available to other pre-trained models or superior to the full training method. AlexNet and GoogLeNet are relatively established DCNNs, and there might be some discrepancy if our results are applied to newly developed DCNNs. In addition, AlexNet and GoogLeNet require the conversion of the DICOM format to PNG and JPEG formats, respectively, which might leave room for improving to effectively convolute the image information. There might be some bias in the datasets because positive and negative images were ‘optimally’ positive and negative, respectively. Namely, we excluded patients with non-metastatic brain tumors, cerebrovascular diseases, or examination artifacts from the control group and removed the two positive cases with movement artifacts from the brain metastases group. We obtained the current results from a single dataset instead of multiple datasets under the retrospective study design, which might have some bias in nature, such as the selection bias, observation bias, or multiple testing problem. To accurately clarify and confirm our present findings, prospective studies with larger amounts of data must be conducted.

Conclusion
In conclusion, our investigation revealed that, respectively, AlexNet and GoogLeNet had (1) maximal bias of 4.6% and 4.9%, (2) 95% CIs in K-CV within the respective maximum bias of 4.6% or 4.9% when the K-value was no <10 or four, (3) the highest accuracy of 74% with an equal positive versus negative ratio and 91% at four times of negative-to-positive image ratio dataset, (4) the accuracy improvement curves ranging from 69% to 74% and from 73% to 88% with a positive versus negative pair of the training image number ranging from 500 to 2495 images, (5) the range of mean accuracy possibly predicted by the bootstrap method using nine or six result sets out of 10 result sets of 10-CV, and (6) metastatic lesion detection accuracy at 50% and 45% by R-CNN in conjunction with the primary selection test. Our series will help to conduct the accuracy investigation of computer diagnosis in medical imaging.

Funding
This work was supported in part by Grants-in-Aid for Scientific Research from Japan Society for the promotion of science under Grant Number 16K10333 and Japan Agency for Medical Research and Development (AMED) under Grant Number JP18lk1010028.

Conflicts of Interest
No potential conflicts of interest were reported by the authors.
References

1. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology 2018; 286:887–896.

2. Nakao T, Hanako S, Nomura Y, et al. Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. J Magn Reson Imaging 2018; 47:948–953.

3. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017; 284:574–582.

4. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN Architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016; 35:1285–1298.

5. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther 2015; 8:2015–2022.

6. Noguchi T, Higa D, Asada T, et al. Artificial intelligence using neural network architecture for radiology (AINNAR): classification of MR imaging sequences. Jpn J Radiol 2018; 36:691–697.

7. Ueda D, Yamamoto A, Nishimori M, et al. Deep learning for MR angiography: automated detection of cerebral aneurysms. Radiology 2019; 290:194–195.

8. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems, Volume 1, Curran Associates Inc., Nevada, 2012; 1097–1105.

9. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Boston, 2015; 1–9. doi: 10.1109/CVPR.2015.7298594

10. Ueda D, Shimazaki A, Miike Y. Technical and clinical overview of deep learning in radiology. Jpn J Radiol 2019; 37:15–33.

11. Tajbakhsh N, Shin JY, Guruud SR, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 2016; 35:1299–1312.

12. Cheng PM, Malhi HS. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. J Digit Imaging 2017; 30:234–243.

13. Hastie T, Tibshirani R, Friedman J. 7.10.1 K-fold cross validation, In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009; 241–245.

14. Stone M. Cross-validatory choice and assessment of statistical predictions. J Royal Stat Soc 1974; 36:111–133.

15. Breiman L, Spector P. Submodel selection and evaluation in regression. The X-random case. Int Stat Rev 1992; 60:291–319.

16. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Volume 2, Morgan Kaufmann Publishers Inc., Quebec, 1995; 1137–1143.

17. Geisser S. The predictive sample reuse method with applications. J Am Stat Assoc 1975; 70:320–328.

18. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: a review. GESTS Int Trans Comput Sci Eng 2006; 30:25–36.

19. Guo X, Yin Y, Dong C, Yang G, Zhou G. On the class imbalance problem. 2008 Fourth International Conference on Natural Computation, IEEE Computer Society, Jinan, 2008; 192–201.

20. Efron B. Better bootstrap confidence intervals. J Am Stat Assoc 1987; 82:171–185.

21. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Computer Vision and Pattern Recognition 2014, IEEE Computer Society, Columbus, 2014; 580–587.

22. Hastie T, Tibshirani R, Friedman J. 7.11.1 Example (Continued), In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009; 252–254.

23. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation: rejoinder. J Am Stat Assoc 1987; 82:548–550.

24. Hussain Z, Gimenez F, Yi D, Rubin D. Differential data augmentation techniques for medical imaging classification tasks. AMIA Annu Symp Proc 2017; 2017:979–984.

25. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems, Volume 2, MIT Press, Montreal, 2014; 2672–2680.

26. Sakai K, Yamada K. Machine learning studies on major brain diseases: 5-year trends of 2014–2018. Jpn J Radiol 2019; 37:34–72.

27. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997; 92:548–560.

28. Hastie T, Tibshirani R, Friedman J. 7.11 Bootstrap methods, In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009; 249–252.

29. Efron B. Bootstrap methods: another look at the jackknife. Ann Stat 1979; 7:1–26.

30. Charron O, Lallemand A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. Comput Biol Med 2018; 95:43–54.

31. Perez-Ramirez U, Arana E, Moratal D. Computer-aided detection of brain metastases using a three-dimensional template-based matching algorithm. Conf Proc IEEE Eng Med Biol Soc 2014; 2014:2384–2387.
32. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation, In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Springer, Cham: Switzerland, 2015; 9351:234–241.

33. Girshick R. Fast R-CNN. Computer Vision (ICCV), IEEE Computer Society, Santiago, 2015; 1440–1448.

34. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. Computer Vision and Pattern Recognition (CVPR) 2016, IEEE Computer Society, Las Vegas, 2016; 779–788.

35. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 2017; 39:1137–1149.