Title

Deep Learning-based Algorithm Improved Radiologists' Performance in Bone Metastases Detection on CT

Abstract

Objectives: To develop and evaluate a deep learning-based algorithm (DLA) for automatic detection of bone metastases on CT.

Methods: This retrospective study included CT scans acquired at a single institution between 2009 and 2019. Positive scans with bone metastases and negative scans without bone metastasis were collected to train the DLA. Another 50 positive and 50 negative scans were collected separately from the training dataset and were divided into validation and test datasets at a 2:3 ratio. The clinical efficacy of the DLA was evaluated in an observer study with board-certified radiologists. Jackknife alternative free-response receiver operating characteristic analysis was used to evaluate observer performance.

Results: A total of 269 positive scans including 1375 bone metastases and 463 negative scans were collected for the training dataset. The number of lesions identified in the validation and test datasets was 49 and 75, respectively. The DLA achieved a sensitivity of 89.8% (44 of 49) with 0.775 false positives per case for the validation dataset and 82.7% (62 of 75) with 0.617 false positives per case for the test dataset. With the DLA, the overall performance of nine radiologists with reference to the weighted alternative free-response receiver operating characteristic figure of merit improved from 0.746 to 0.899 ($P < .001$). Furthermore, the mean interpretation time per case decreased from 168 to 85 s ($P = .004$).

Conclusion: With the aid of the algorithm, the overall performance of radiologists in bone metastases detection improved, and the interpretation time decreased at the same time.

Keywords

Bone Diseases; Neoplasm Metastasis; Multidetector Computed Tomography; Deep Learning; Radiographic Image Interpretation, Computer-Assisted

Key Points

- A deep learning-based algorithm for automatic detection of bone metastases on CT was developed.
- In the observer study, overall performance of radiologists in bone metastases detection improved significantly with the aid of the algorithm.
- Radiologists' interpretation time decreased at the same time.

Abbreviations

CAD	computer-aided detection
DLA	deep learning-based algorithm
CNN	convolutional neural network
DSC	Dice similarity coefficient
FP	false-positive
FN	false-negative
JAFROC	jackknife free-response receiver operating characteristic
wAFROC-FOM	weighted alternative free-response receiver operating characteristic figure of merit

Introduction

Early detection of bone metastases is a common and important task for radiologists. Bones are one of the most common sites of metastases, along with the lungs and liver [1]. Metastatic disease significantly affects staging and prognosis in cancer patients. Moreover, bone metastases often cause skeletal-related events, including cancer-related pain, pathologic fractures, spinal cord compression, and hypercalcemia [2, 3].

Currently, several imaging modalities, including CT, MRI, bone scintigraphy, and PET, can be used to examine bone metastases, with each showing distinct advantages [2, 4, 5]. Because of its high spatial resolution, CT can demonstrate small and slight bone abnormalities [4–6]. It also allows simultaneous evaluation of the primary and metastatic lesions [5, 7]. Furthermore, CT is readily available at a relatively low cost [8]. Thus, in the clinical setting, CT is the most frequently used modality for both initial staging and serial follow-up of cancer patients [5, 8, 9].

Nevertheless, bone metastases detection on CT is challenging and time-consuming for the following reasons: (i) since bones are present throughout the body, radiologists must scrutinize all slices; (ii) the radiologic appearance of bone metastases varies from sclerosing to lytic types [4, 5]; thus, no single window setting can properly depict all bone metastases [10, 11]; and (iii) benign mimickers such as bone islands, fractures, and degenerative changes often complicate diagnoses [4, 11]. An oversight can easily occur especially when the examination is not directly aimed to investigate bone metastases [12]. Therefore, there is a great demand for computer-aided detection (CAD) systems for bone metastases on CT.

The use of deep learning for various tasks in medical image analysis has gained much interest recently [13]. However, few reports have assessed automatic detection of bone metastases on CT [8, 14].

In this study, we attempted to develop a deep learning-based algorithm (DLA) for automatic detection of bone metastases on CT. An observer study with nine board-certified

radiologists was performed to evaluate the clinical efficacy of the algorithm.

Materials and Methods

This retrospective study was approved by our Ethics Committee. Formal consent was waived by the Ethics Committee due to the retrospective nature of data collection.

Data Acquisition

All images were obtained retrospectively from the clinical databases of a single institution acquired between 2009 and 2019. Both staging and follow-up studies of malignancy with both plain and intravenous contrast-enhanced CTs were included. For contrast-enhanced CTs, delayed phase images were used. Images with soft-tissue kernel reconstruction were included.

Training Dataset: The inclusion criteria for positive scans in the training dataset were as follows: (a) presence of at least one radiologically reported bone metastasis and (b) availability of thin-slice images of ≤1 mm thickness. To increase data volume, more than one scan was included from one patient if the radiological appearance of bone metastases changed substantially. Simultaneously, CT scans from patients without malignancy were collected as negative controls to allow the networks to identify normal bones. The patients included in the validation and test datasets described below were excluded from the training dataset.

Validation and Test Datasets: To evaluate the algorithm performance, 50 positive and 50 negative control scans were consecutively collected separately from the training dataset. Inclusion criteria for positive scans were as follows: (a) presence of at least one radiologically reported bone metastasis, (b) availability of thin-slice images of ≤1 mm thickness, (c) availability of one or more subsequent CT studies and one or more subsequent bone scintigraphy or fluorine-18-fluorodeoxyglucose (FDG)-PET studies, (d) presence of at least one lesion ≥10 mm, and (e) presence of fewer than six lesions per scan. Bone metastases were

diagnosed by continuous growth on subsequent CT images or substantial focal uptake on bone scintigraphy or FDG-PET. Bone metastases of <5 mm were not included because of the difficulty in confirming their diagnosis. Negative control cases with malignancy were selected to match the distribution of age, sex, and primary lesions. The 50 positive and 50 negative scans were divided into validation and test datasets randomly in a 2:3 ratio. A flowchart outlining data collection and division is presented in Figure 1.

Preparation of Ground Truth Labels

For the training dataset, ground truth labels were established with manual segmentation performed by a board-certified radiologist (S.N.: 7 years' experience, general radiologist). For the validation and test datasets, ground truth labels were established with manual segmentation, with consensus between two board-certified radiologists (S.N. and R.S.: 14 years' experience, general radiologist).

Deep Learning-based Algorithm

The algorithm consists of three convolutional neural networks (CNNs): (i) a 2D UNet-based network for bone segmentation, (ii) a 3D UNet-based network for candidate region segmentation, and (iii) a 3D ResNet-based network for false-positive (FP) reduction. UNet and ResNet are popular CNN architectures commonly used for segmentation and classification, respectively [15, 16]. The 3D modification of these networks was performed in the second and third steps of our framework to capture 3D anatomical information efficiently. A schematic of the algorithm is shown in Figure 2 and more detailed schematic is shown in Figure S1 in the supplementally material.

As the first step, the original image was fed into a 2D UNet-based bone segmentation network slice by slice. The network was developed in the previous study [17]. The result of

bone segmentation was used as a reference in the next step. After bone segmentation, isotropic resampling was performed to standardize each voxel to 1 × 1 × 1 mm. As the second step, resampled image was cropped into 96 × 96 × 96 voxel blocks with a stride of 48, and fed into a 3D UNet-based network to extract candidate regions of bone metastases. To reduce calculation time, only blocks including bone region (in reference to segmentation result) were fed into the network. The output blocks of the 3D UNet-based network were merged again into one image. In the merging process, overlapped areas of blocks were averaged according to the number of overlaps. After merging, candidate regions smaller than 100 voxels in volume were discarded. As the final step, the image was cropped into 32 × 32 × 32 voxel blocks with a stride of 16, and fed into 3D ResNet-based networks for FP reduction. Only blocks including candidate regions were fed into the networks. An ensemble of three 3D ResNet-based networks trained independently was employed to improve sensitivity. The networks predicted probability of bone metastasis for each block, and the probabilities were assembled for each candidate region. When at least one of the three networks predicted a probability exceeding the preset threshold, for at least one of the blocks including the relevant region, the region was included in the final output.

Total image processing took around 3 min for a torso CT (chest to pelvis or neck to pelvis). Details of the development environment, network architecture, and hyperparameters are provided in the supplementary material. The code is available at https://github.com/snkp/bmd.

Measurement of Algorithm Performance

The Dice similarity coefficient (DSC), a commonly used spatial overlap index, was employed to evaluate the accuracy of the final output of the DLA [18]. The DSCs of the ground truth labels and candidate regions were calculated in a 3D volumetric manner. Ground truth labels with a DSC of ≥0.3 were counted as true positive, and those with a DSC of <0.3 were counted as false negative (FN). Candidate regions showing no spatial overlap with ground truth labels were counted as FP. Lesion-based sensitivity and FP counts per case were calculated based on these

values and were defined as the main outcomes to assess the performance of the DLA.

Observer Study

An observer study with nine board-certified radiologists was conducted to evaluate the clinical efficacy of the algorithm. The observers evaluated 60 CT scans of the test dataset without and with the DLA and marked the locations of suspicious lesions on the images by rating the likelihood (1–100) of bone metastasis.

A medical monitor (Radiforce RX440; EIZO Corporation) and an in-house dedicated image viewer (Figure 3) with multi-planar reconstruction and window level/width modification functions were offered to view the CT images. To control practice effects, the observers trained usage of the viewer prior to the actual study. The observers were blinded to all clinical data except for the age and sex of each patient. The interpretation time for each scan was recorded automatically by the viewer.

Five of the nine observers interpreted each scan first without and then with the DLA. The other four interpreted the scans first with and then without the DLA. The interval between the sessions with and without the DLA was ≥28 days. The order of scans was randomized for each observer.

After completion of all assessments, the marked lesion locations were compared with the ground truth labels for lesion identification. Lesions with a likelihood rating of 51–100 (upper half of 1–100) were considered as positive in lesion-based analysis. Cases with at least one positive lesion were considered positive in case-based analysis.

Statistical Analysis

To evaluate the overall performance of the radiologists, the jackknife free-response receiver operating characteristic (JAFROC) analysis was performed with random-reader and random-

case models. A weighted alternative free-response receiver operating characteristic figure of merit (wAFROC-FOM) was defined as the main evaluation index of the observer study. Simply put, the wAFROC curve is a variant of ROC curve that supports data with multiple lesions per case, and the wAFROC-FOM is identical to the area under the wAFROC curve [19].

Lesion-based sensitivity, FP count per case, case-based sensitivity, case-based specificity, and interpretation time were compared between sessions without and with DLA using the Wilcoxon signed rank test. R (Version 4.02) and RJafroc package (Version 2.0.1) [20, 21] were used for statistical analyses, and $P < .05$ was considered to indicate a significant difference.

Based on the results of the previous study [22] and the sample size calculation conducted with RJafroc package, 60 cases with nine readers were estimated as a suitable sample size for the current study.

Results

Demographics of the Three Datasets

A total of 269 positive scans, including 1375 bone metastases, and 463 negative scans were collected for the training dataset. The validation dataset consisted of 20 positive scans including 49 lesions and 20 negative scans, and the test dataset consisted of 30 positive scans including 75 lesions and 30 negative scans. Case demographics, scan conditions, and lesion characteristics are summarized in Tables 1 and 2. The details of image acquisition are also presented in the supplementary material.

Algorithm Performance

The overall results of the DLA for the validation and test datasets are given in Table 3. The

algorithm outputs different results according to the preset threshold; thus, the results for each threshold from 0.1 to 0.9 are presented. The lesion-based sensitivity increased with lowering the threshold. However, the FP count per case also increased accordingly. Based on the results for the validation dataset, a threshold of 0.6 was defined as the standard value, since too many FP outputs may impede clinical assessments. At a threshold of 0.6, the lesion-based sensitivity was 89.8% (44 of 49) with 0.775 FPs per case for the validation dataset and 82.7% (62 of 75) with 0.617 FPs per case for the test dataset. The case-based sensitivity and specificity were 100% (20 of 20) and 70.0% (14 of 20) for the validation dataset and 100% (30 of 30) and 80.0% (24 of 30) for the test dataset.

The results of the bone metastases detection stratified according to lesion characteristics are presented in Table 4. Representative images of true-positive lesions with various appearances and locations are shown in Figure 4, and representative images of FN lesions and FP regions are shown in Figure 5. The sensitivity improved with increasing diameter; for the test dataset, it reached 87.8% (58 of 66) in lesions ≥10 mm and 94.1% (16 of 17) in lesions ≥30 mm. However, the sensitivity in lesions of <10 mm was limited to 44.4% (4 of 9). In terms of location, sensitivities were high for lesions located on the vertebrae and pelvis (89.7% and 92.9%, respectively) and low for lesions located on the scapulae and limbs (50.0% and 60.0%, respectively). For the test dataset, sensitivities for sclerotic lesions and mixed lesions were lower than those for lytic lesions (83.9%, 73.7% vs. 92.0%); however, this was not the case for the validation dataset (85.7%, 100% vs. 89.5%).

The number of FNs for the test dataset was 13 of 75 lesions (ribs, 4; vertebrae, 3; scapulae, 2; limbs, 2; pelvis, 1; sternum, 1). Of 13 FNs, 5 lesions were <10 mm and 12 lesions were <20 mm. The only FN lesion ≥20 mm is shown in Figure 5c. The lesion was detected by the DLA but was counted as FN since the DSC was <0.3.

The FP count for the test dataset was 37 per 60 cases (pelvis, 11; ribs, 7; scapulae, 6; vertebrae, 5; other bones, 3; and outside bones, 5). Typical patterns of FPs included

osteoarthritic changes (3 FPs), spinal compression fractures or rib fractures (5 FPs), and sclerotic changes due to other reasons (4 FPs).

Observer Study

The experience of the nine observers in diagnostic radiology ranged from 6 to 18 years, and their areas of expertise varied (breast radiology: 2, genitourinary radiology: 2, nuclear medicine: 2, gastrointestinal radiology: 1, neuroradiology: 1, general radiology: 1). All of them had interpreted torso CT scans on a daily basis.

Table 5 shows the main results for image interpretation, and Figure 6 shows the average free-response receiver operating characteristic curves of the nine radiologists without and with the DLA. In comparison with interpretation without the DLA, DLA-assisted interpretations were associated with a higher mean wAFROC-FOM (0.746 [95% CI: 0.690, 0.802] vs. 0.899 [95% CI 0.865, 0.932]; $P < .001$), mean sensitivity in lesion-based analysis (51.7% [38.8 of 75] vs. 71.7% [53.8 of 75]; $P = .004$), and mean sensitivity in case-based analysis (74.4% [22.3 of 30] vs. 91.1% [27.3 of 30]; $P = .004$). In addition, the mean interpretation time per case decreased when using the DLA (168 s [95% CI: 125, 210] vs. 85 s [95% CI: 59, 110]; $P = .004$). No significant inter-session difference was observed in the mean FP count per case (0.237 vs. 0.157; $P = .57$) and mean specificity for case-based analysis (95.2% [28.6 of 30] vs. 96.2% [28.9 of 30]; $P = .69$).

The results of the radiologists' interpretations stratified according to lesion characteristics are presented in Table 4. Improvement in sensitivity with the DLA was observed regardless of the location, appearance, and diameter of lesions.

After the observer study, all the candidate lesions indicated by the DLA or marked by the observers were reviewed, and two true lesions—missing from the ground truth labels—were newly found. The summary of these lesions is presented in Figure S4 in the supplementary material. Only one or two of nine radiologists detected each lesion, and the DLA did not detect

either of them. These two lesions were not accounted for in the sensitivity calculation and the markings on them were treated as FPs, because they were not included in the pre-established ground truth labels.

Discussion

Bone metastases detection on CT is a common but challenging task for radiologists. In the current study, we employed deep learning to develop a supportive algorithm to overcome this problem. The proposed algorithm achieved a lesion-based sensitivity of 89.8% with 0.775 false positives per case for the validation dataset, and 82.7% with 0.617 false positives per case for the test dataset. With the algorithm, the weighted alternative free-response receiver operating characteristic figure of merit, which indicated the overall performance of the radiologists, improved from 0.746 to 0.899. Furthermore, the mean interpretation time per case decreased from 168 to 85 s.

In the subgroup analysis of DLA performance, diameter showed the clearest correlation with sensitivity. The smaller the lesion, the less image features it contained, making extraction more difficult. In terms of location, the sensitivity for lesions located on the vertebrae and pelvis was high and that for lesions located on the scapulae and limbs was low. This can be explained by the difference in the number of training data: 620 and 412 vertebral and pelvic lesions and only 38 and 32 scapula and limb lesions, respectively. The small number of scapula or limb lesions in the training dataset was because of the low frequency of bone metastases to such locations [23, 24]. The difference in sensitivity by the radiologic appearance of lesions was not significant.

In the observer study, the sensitivity of radiologists in the detection of bone metastases improved with the aid of the DLA, from 51.7% to 71.7% in lesion-based analysis and 74.4% to 91.1% in case-based analysis. The number of FNs (i.e., overlooked lesions or cases) decreased from 36.2 to 21.2 in lesion-based analysis and 7.7 to 2.7 in case-based analysis, which indicates

the clinical usefulness of the algorithm (Figure S5 in the supplementary material). FP count per case was not increased with DLA use. The case-based specificity was high enough without using DLA; thus, there was no significant difference between the two sessions.

Several attempts have been made to develop a CAD system for detecting bone metastases on CT. Before the era of deep learning, Burns et al. used a combination of a watershed segmentation algorithm and a support vector machine classifier, and Hammon et al. used three consecutive random forest classifiers, each processing local image features, to assess candidate regions for bone metastases [9, 12]. Roth et al. applied a deep CNN to the output of a pre-existing CAD system and demonstrated its efficacy in FP reduction [14]. Chmelik et al. used a deep CNN for voxel-wise segmentation, followed by a random forest classifier for FP reduction [8]. However, all these studies focused only on spinal lesions. Although the spine is the most frequent site of bone metastases, metastases can occur at any site in the entire skeleton [23, 24]. Our algorithm can detect bone metastases in all scanned areas, making it more clinically useful. From a technical point of view, the previous studies applied deep CNN in combination with traditional machine learning algorithms processing handcrafted image features [8, 14]. Our algorithm is the first to apply deep CNN to both candidate region detection and FP reduction. Comparison to the previous studies is summarized in Table S1 in the supplementary material.

This is the first study that evaluated the clinical usefulness of a deep learning-based CAD system for bone metastases detection on CT through an observer study with radiologists. We believe that this will be a significant step toward the development of general-purpose CAD [25]. Many reports have described deep learning-based CAD system for lung nodules or liver tumors detection, and some are already in practical use [26–32]. Combining them with the proposed algorithm may yield a general-purpose CAD that can be widely used for staging and follow-up of cancer patients.

This study had several limitations. First, our datasets included images from only a

single institution and a single scanner vendor. The generalizability of the algorithm needs to be assessed with a multi-institution, multi-vendor external dataset. Second, data were collected retrospectively; therefore, selection bias may have influenced our findings and the prevalence of disease may not be same as in a real clinical situation. Ideally, the test dataset should be corrected prospectively. These two limitations are the most important factors we should address in our next study. Third, lesions <5 mm and scans with only lesions <10 mm were not included in the validation and test datasets. The usefulness of the algorithm in such extremely difficult cases was not evaluated. Fourth, the ground truth labels were established carefully by the experts, but were determined to be not truly perfect; there were two lesions newly found after the observer study, causing a slight bias in the results. Fifth, the lesions were not confirmed by histological analysis, because biopsies are rarely performed for suspected bone metastases [33]. As an alternative, subsequent CT images and other imaging modalities were referenced to confirm the diagnosis. Sixth, the algorithm assumed using thin-slice images of 1 mm or less. This requirement can be a limitation to clinical implementation.

In conclusion, we successfully developed a deep learning-based algorithm for automatic detection of bone metastases on CT. The results of the observer study indicate the clinical efficacy of the algorithm.

Acknowledgments

The authors are grateful to Editage (http://www.editage.com) for their assistance in language editing.

References

1. Coleman RE (2001) Metastatic bone disease: clinical features, pathophysiology and treatment strategies. Cancer Treat Rev 27:165–176
2. Macedo F, Ladeira K, Pinho F, et al. (2017) Bone metastases: an overview. Oncol Rev 11:321
3. D'Oronzo S, Coleman R, Brown J, Silvestris F (2019) Metastatic bone disease: Pathogenesis and therapeutic options: up-date on bone metastasis management. J Bone Oncol 15:100205
4. O'Sullivan GJ, Carty FL, Cronin CG (2015) Imaging of bone metastasis: an update. World J Radiol 7:202-211
5. Heindel W, Gübitz R, Vieth V, Weckesser M, Schober O, Schäfers M (2014) The diagnostic imaging of bone metastases. Dtsch Arztebl Int 111:741–747
6. Kalogeropoulou C, Karachaliou A, Zampakis P (2009) Radiologic evaluation of skeletal metastases: role of plain radiographs and computed tomography. In: Cancer Metastasis – Biology and Treatment, 12:119–136. Springer, Dordrecht
7. Groves AM, Beadsmoore CJ, Cheow HK, et al. (2006) Can 16-detector multislice CT exclude skeletal lesions during tumour staging? Implications for the cancer patient. Eur Radiol 16:1066–1073
8. Chmelik J, Jakubicek R, Walek P, et al. (2018) Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in

3D CT data. Med Image Anal 49:76–88

9. Hammon M, Dankerl P, Tsymbal A, et al. (2013) Automatic detection of lytic and blastic thoracolumbar spine metastases on computed tomography. Eur Radiol 23:1862–1870
10. Vandemark RM, Shpall EJ, Affronti M Lou (1992) Bone metastases from breast cancer: value of CT bone windows. J Comput Assist Tomogr 16:608–614
11. Pomerantz SM, White CS, Krebs TL, et al. (2000) Liver and bone window settings for soft-copy interpretation of chest and abdominal CT. Am J Roentgenol 174:311–314
12. Burns JE, Yao J, Wiese TS, Muñoz HE, Jones EC, Summers RM (2013) Automated detection of sclerotic metastases in the thoracolumbar spine at CT. Radiology 268:69–78
13. Choy G, Khalilzadeh O, Michalski M, et al. (2018) Current applications and future impact of machine learning in radiology. Radiology 288:318–328
14. Roth HR, Lu L, Liu J, et al. (2016) Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans Med Imaging 35:1170–1181
15. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: MICCAI 2015. Lecture Notes in Computer Science, 9351:234–241. Springer, Cham
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) p770–778. Las Vegas
17. Noguchi S, Nishio M, Yakami M, Nakagomi K, Togashi K (2020) Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques. Comput Biol Med 121:103767
18. Zou KH, Warfield SK, Bharatha A, et al. (2004) Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 11:178–189

19. Chakraborty DP, Zhai X (2016) On the meaning of the weighted alternative free-response operating characteristic figure of merit. Med Phys 43:2548–2557

20. Chakraborty DP (2017) Observer performance methods for diagnostic imaging: foundations, modeling, and applications with R-based examples. CRC Press, Boca Raton

21. Chakraborty DP (2021) The RJafroc Book. Available via https://dpc10ster.github.io/RJafrocBook/. Accessed 24 Dec 2021

22. Sakamoto R, Yakami M, Fujimoto K, et al. (2017) Temporal subtraction of serial CT images with large deformation diffeomorphic metric mapping in the identification of bone metastases. Radiology 285:629–639

23. Nakamoto Y, Osman M, Wahl RL (2003) Prevalence and patterns of bone metastases detected with positron emission tomography using F-18 FDG. Clin Nucl Med 28:302–307

24. Kakhki VRD, Anvari K, Sadeghi R, Mahmoudian AS, Torabian-Kakhki M (2013) Pattern and distribution of bone metastases in common malignant tumors. Nucl Med Rev 16:66–69

25. Kobatake H (2007) Future CAD in multi-dimensional medical images: - project on multi-organ, multi-disease CAD system -. Comput Med Imaging Graph 31:258–266

26. Liu K, Li Q, Ma J, et al. (2019) Evaluating a fully automated pulmonary nodule detection approach and its impact on radiologist performance. Radiol Artif Intell 1:e180084

27. Xie H, Yang D, Sun N, Chen Z, Zhang Y (2019) Automated pulmonary nodule detection in CT images using deep convolutional neural networks. Pattern Recognit 85:109–119

28. Pehrson LM, Nielsen MB, Lauridsen CA (2019) Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: A systematic review. Diagnostics 9:29

29. Chlebus G, Schenk A, Moltz JH, van Ginneken B, Hahn HK, Meine H (2018) Automatic

liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. Sci Rep 8:15497

30. Vorontsov E, Cerny M, Régnier P, et al. (2019) Deep learning for automated segmentation of liver lesions at ct in patients with colorectal cancer liver metastases. Radiol Artif Intell 1:e180014

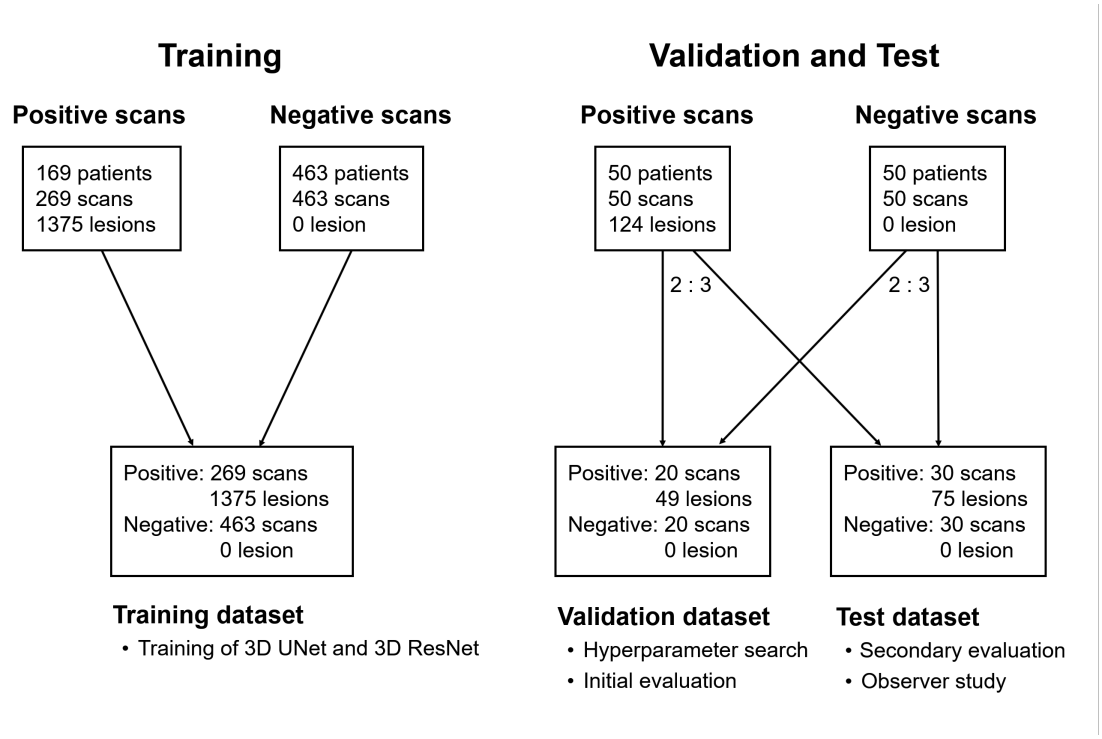
31. Azer SA (2019) Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: a systematic review. World J Gastrointest Oncol 11:1218–1230

32. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M (2021) Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 31:3797–3804

33. Çiray I, Åström G, Sundström C, Hagberg H, Ahlström H (1997) Assessment of suspected bone metastases: CT with and without clinical information compared to CT-guided bone biopsy. Acta radiol 38:890–895

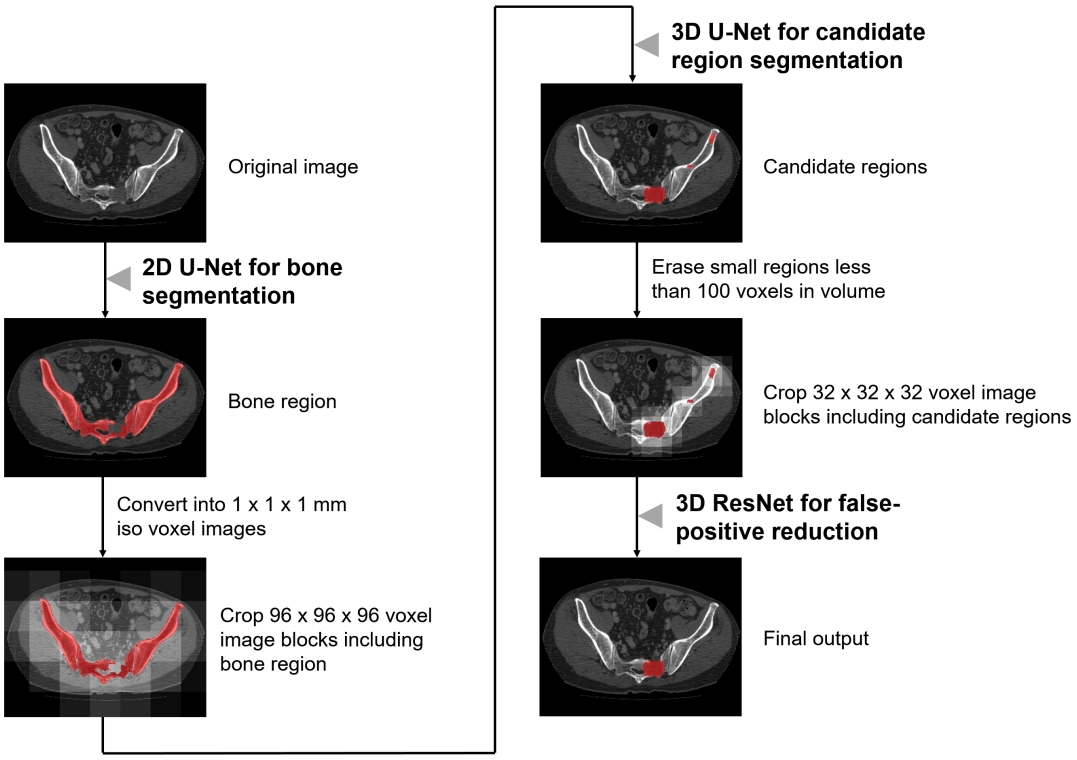
Figures

Figure 1. Flowchart of data collection and division



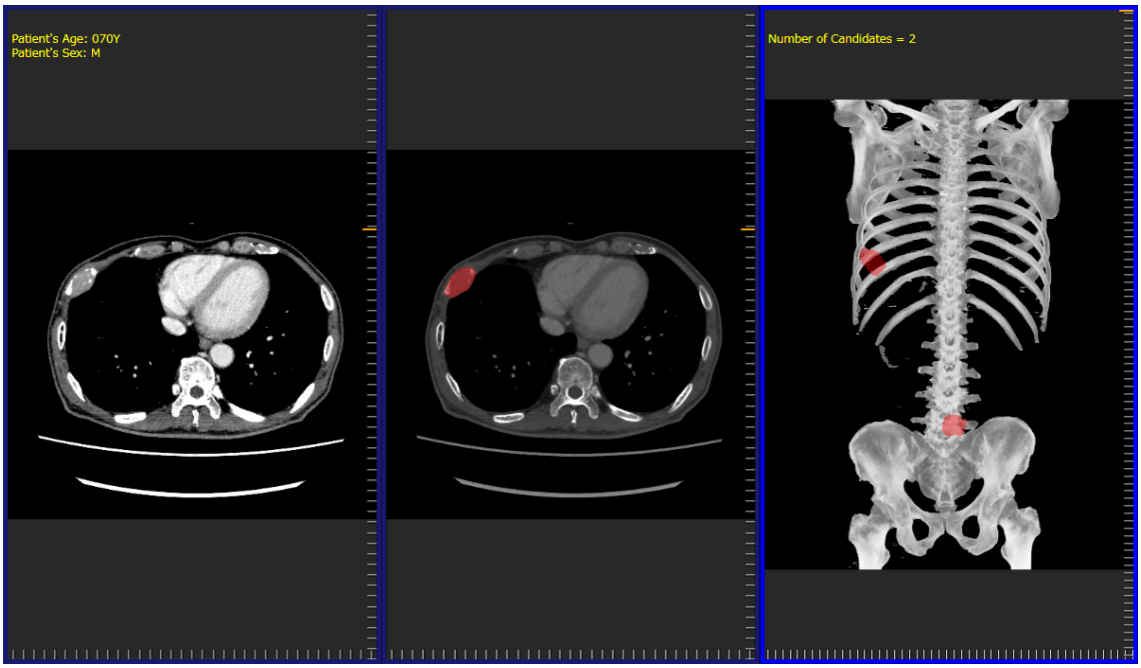
All scans were collected retrospectively from the clinical databases of a single institution.

Figure 2. Schematic of the proposed algorithm



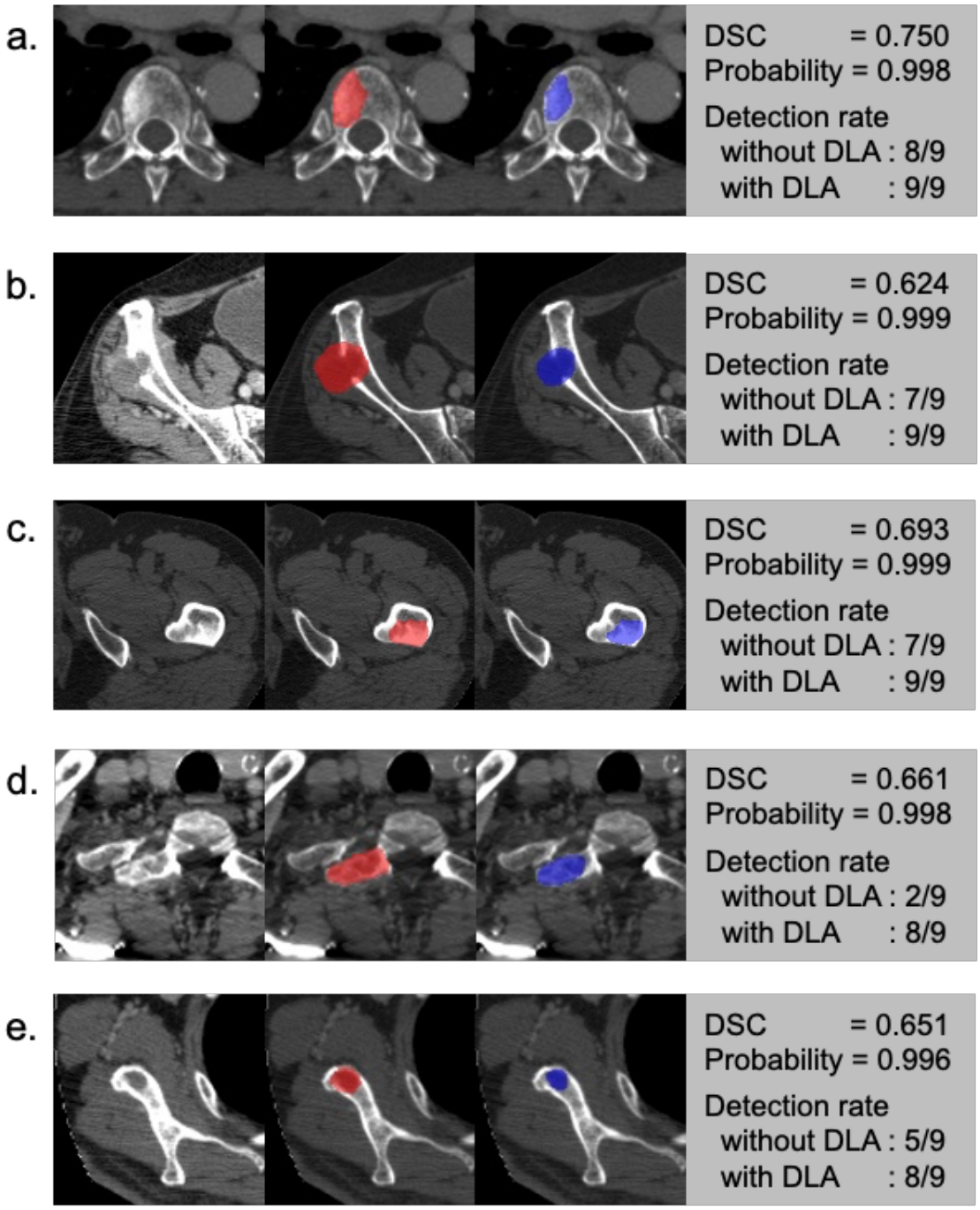
Schemas are shown in 2D planar images for simplicity. In truth, most of the processes are operated in a 3D volumetric manner, except for 2D UNet for bone segmentation. In this case, a candidate region on the left half of the sacrum was included in the final output, and two candidate regions on the left ileum were discarded.

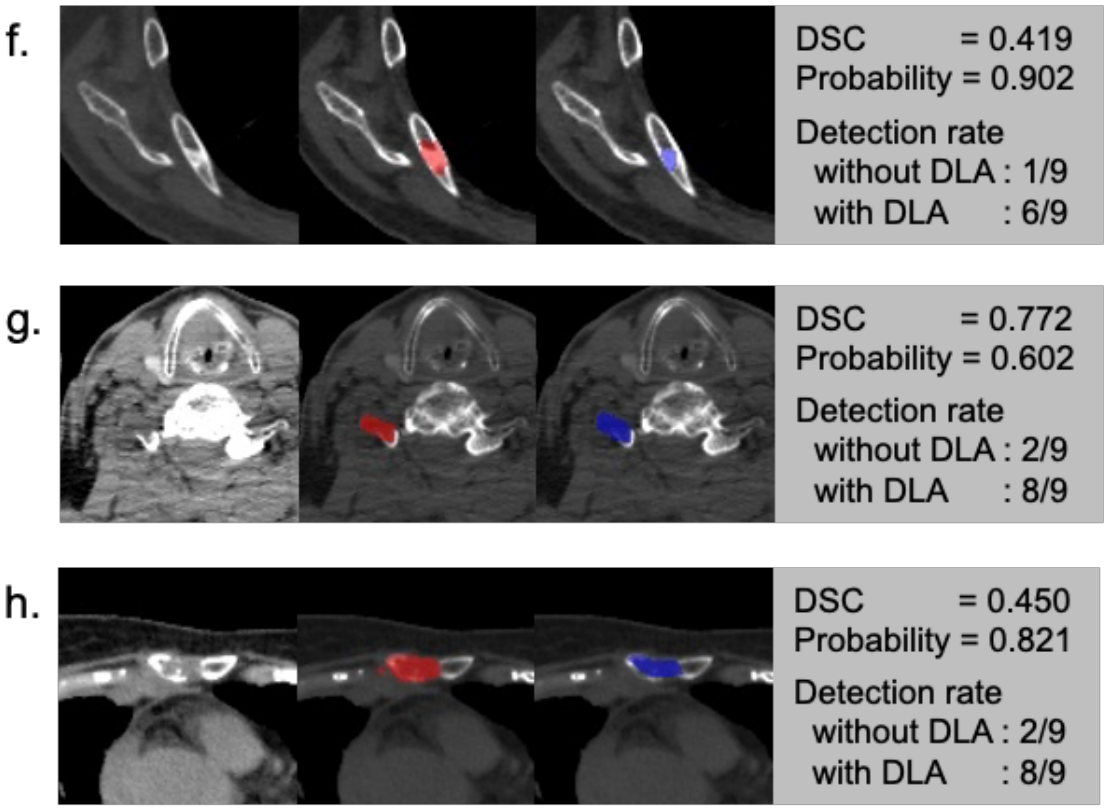
Figure 3. Screenshot of the image viewer for the observer study



From left to right, the three images in a row are the original image, overlaid image of the original image and the candidate region output from the proposed algorithm, and maximum intensity projection of the bone region overlaid with the candidate region. In this case, two candidate regions located on the right rib and lumbar spine are presented. Patient age, sex, and number of candidate regions were also displayed. When an observer clicks on a suspicious lesion, a dialog box for rating the likelihood (1–100) of bone metastasis appears.

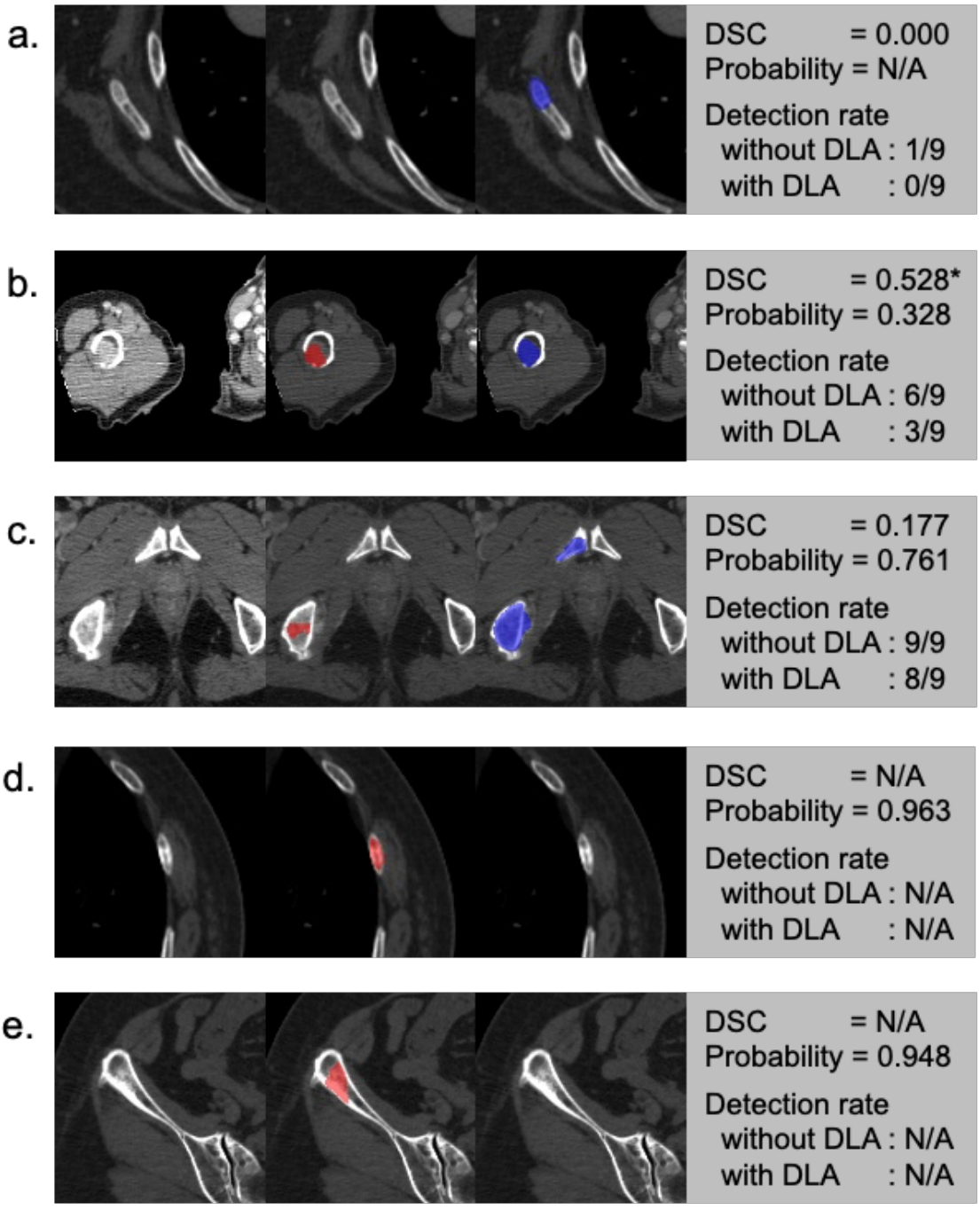
Figure 4. Representative images of true-positive lesions with various appearances and locations



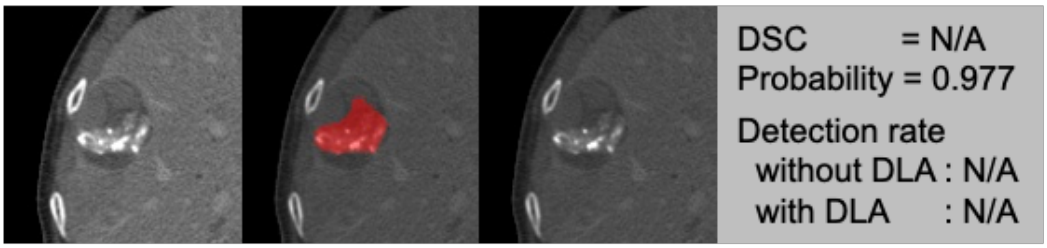


From left to right, the three images in a row are the original image, candidate region output from the DLA (red), and the ground truth label (blue). The DSC of the candidate region and ground truth label, predicted probability for the candidate region, and detection rate by radiologists without and with the DLA in the observer study are shown in the right table. (a) Sclerotic bone metastasis on the vertebra. (b) Expansile lytic bone metastasis in the right iliac bone of the pelvis. (c) Sclerotic bone metastasis in the left femur. (d) Mixed sclerotic and lytic bone metastasis on the right transverse process of the vertebra. (e) Lytic bone metastasis in the right scapula. (f) Small sclerotic bone metastasis in the right rib. (g) Small lytic bone metastasis on the right transverse process of the vertebra. (h) Lytic bone metastasis in the sternum. Abbreviations: DLA, deep learning-based algorithm; DSC, Dice similarity coefficient.

Figure 5. Representative images of false-negative lesions (a–c) and false-positive regions (d–f).

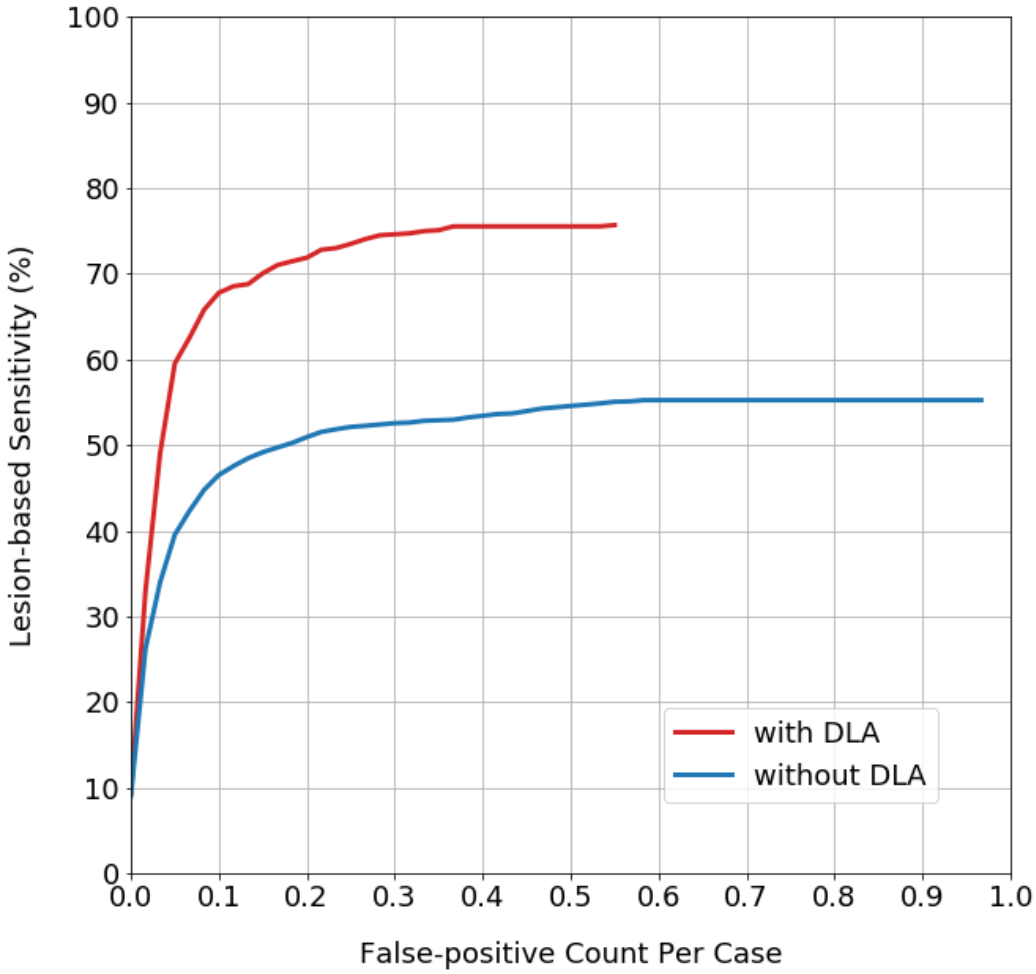


f.



(a) Small sclerotic bone metastasis on the right scapula. It appears to be too small and faint for the DLA to detect. (b) Lytic bone metastasis on the right humerus. Note that the red region on the middle image is a candidate region before thresholding, and DSC (*) was calculated on this region. With a threshold of 0.6, the region was deleted since its probability was 0.328, which is <0.6. Therefore, this lesion was counted as false-negative. (c) Mixed sclerotic and lytic bone metastasis on the right ischial and pubic bones of the pelvis. The lesion was detected by the DLA but was counted as false-negative since the DSC was <0.3. (d) False-positive region due to an old rib fracture. (e) False-positive region due to non-specific inhomogeneous density of the right iliac bone of the pelvis. (f) False-positive region located outside the bone due to post-therapeutic changes of the liver tumor. Such errors occurred occasionally, because the DLA focuses only on local image features and does not take the holistic anatomical information into account. Abbreviations: DLA, deep learning-based algorithm; DSC, Dice similarity coefficient; N/A, not applicable.

Figure 6. The average free-response receiver operating characteristic curves of the nine radiologists without and with the DLA



The overall performance of radiologists improved significantly with the aid of the DLA.

Abbreviations: DLA, deep learning-based algorithm

Tables

Table 1. Demographics of the cases in the three datasets

		Training		Validation		Test	
		Positive	Negative	Positive	Negative	Positive	Negative
Per Patient	Age (years)	66.5 ± 12.8 (28-86)	63.2 ± 15.9 (1-92)	66.6 ± 13.0 (34-84)	68.6 ± 9.7 (43-81)	67.3 ± 9.8 (45-86)	67.9 ± 9.1 (46-87)
	Sex						
	Male	100 (59)	225 (49)	13 (65)	12 (60)	20 (67)	17 (57)
	Female	69 (41)	238 (51)	7 (35)	8 (40)	10 (33)	13 (43)
	Primary Lesion						
	Lungs	57 (34)	0* (0)	6 (30)	6 (30)	8 (27)	8 (27)
	Prostate	33 (20)	0* (0)	4 (20)	5 (25)	6 (20)	6 (20)
	Breast	25 (15)	0* (0)	3 (15)	5 (25)	6 (20)	6 (20)
	Others	54 (32)	0* (0)	7 (35)	4 (20)	10 (33)	10 (33)
	Total Number of Patients	*169***	*463*	*20*	*20*	*30*	*30*
Per Scan	Use of Contrast Media						
	Plain	153 (57)	237 (51)	8 (40)	9 (45)	12 (40)	16 (53)
	Contrast-enhanced	116 (43)	226 (49)	12 (60)	11 (55)	18 (60)	14 (47)
	Slice Thickness						
	1.0 mm	266 (99)	417 (90)	20 (100)	20 (100)	30 (100)	30 (100)
	0.5 mm	3 (1)	46 (10)	0 (0)	0 (0)	0 (0)	0 (0)
	Scanner Model						
	Aquilion Prime	124 (46)	191 (41)	9 (45)	7 (35)	6 (20)	10 (33)
	Aquilion One	117 (43)	157 (34)	3 (15)	6 (30)	5 (17)	10 (33)
	Aquilion	24 (9)	105 (23)	8 (40)	7 (35)	19 (63)	10 (33)
	Aquilion Precision	4 (1)	10 (2)	0 (0)	0 (0)	0 (0)	0 (0)
	Scan Coverage						
	Neck to Abdomen	133 (49)	329 (71)	12 (60)	8 (40)	14 (47)	13 (43)
	Chest to Abdomen	102 (38)	20 (4)	4 (20)	8 (40)	13 (43)	9 (30)
	Neck to Chest	2 (1)	1 (0)	1 (5)	1 (5)	0 (0)	1 (3)
	Chest	27 (10)	9 (2)	2 (10)	3 (15)	3 (10)	4 (13)
	Abdomen	4 (1)	41 (9)	1 (5)	0 (0)	0 (0)	3 (10)
	Brain	1 (0)	44 (10)	0 (0)	0 (0)	0 (0)	0 (0)
	Neck	0 (0)	19 (4)	0 (0)	0 (0)	0 (0)	0 (0)
	Total Number of Scans	*269***	*463*	*20*	*20*	*30*	*30*

For patient age, the mean age and standard deviation are presented, with range of values in parentheses. For other data, the number of patients or scans are presented, with percentages in parentheses.

*Negative scans of the training dataset were acquired from patients without malignancy.

**For positive cases in the training dataset, the total number of patients and scans were not equal, because more than one scan was included from one patient if the radiological appearance of bone metastases had changed substantially.

Table 2. Characteristics of lesions in the three datasets

	Training	Validation	Test
Location			
Vertebra	620 (45)	22 (45)	29 (39)
Pelvis	412 (30)	15 (31)	14 (19)
Rib	228 (17)	9 (18)	18 (24)
Scapula	38 (3)	1 (2)	4 (5)
Limb	32 (2)	0 (0)	5 (7)
Sternum	30 (2)	2 (4)	5 (7)
Clavicle	11 (1)	0 (0)	0 (0)
Skull	4 (0)	0 (0)	0 (0)
Appearance			
Sclerotic	709 (52)	21 (43)	31 (41)
Lytic	518 (38)	19 (39)	25 (33)
Mixed	148 (11)	9 (18)	19 (25)
Diameter			
≥50 mm	109 (8)	2 (4)	4 (5)
≥30 mm to <50 mm	263 (19)	11 (22)	13 (17)
≥10 mm to <30 mm	896 (65)	31 (63)	49 (65)
≥5 mm to <10 mm	107 (8)	5 (10)	9 (12)
Total Number of Lesions	*1375*	*49*	*75*

Data are the number of lesions for each category, with percentages in parentheses.

Table 3. Performance of the DLA according to the preset threshold

		Lesion-based analysis					Case-based analysis					
	Threshold	TP	FN	FP	Sensitivity (%)	FP per case	TP	FN	TN	FP	Sensitivity (%)	Specificity (%)
Validation dataset (20 positive cases with 49 lesions and 20 negative cases)	0.9	40	9	21	81.6	0.525	19	1	17	3	95.0	85.0
	0.8	41	8	25	83.7	0.625	20	0	17	3	100.0	85.0
	0.7	42	7	26	85.7	0.650	20	0	16	4	100.0	80.0
	0.6	**44**	**5**	**31**	**89.8**	**0.775**	**20**	**0**	**14**	**6**	**100.0**	**70.0**
	0.5	44	5	35	89.8	0.875	20	0	14	6	100.0	70.0
	0.4	44	5	41	89.8	1.025	20	0	13	7	100.0	65.0
	0.3	45	4	47	91.8	1.175	20	0	13	7	100.0	65.0
	0.2	45	4	66	91.8	1.650	20	0	12	8	100.0	60.0
	0.1	45	4	91	91.8	2.275	20	0	10	10	100.0	50.0
Test dataset (30 positive cases with 75 lesions and 30 negative cases)	0.9	55	20	24	73.3	0.400	30	0	27	3	100.0	90.0
	0.8	58	17	32	77.3	0.533	30	0	26	4	100.0	86.7
	0.7	60	15	35	80.0	0.583	30	0	26	4	100.0	86.7
	0.6	**62**	**13**	**37**	**82.7**	**0.617**	**30**	**0**	**24**	**6**	**100.0**	**80.0**
	0.5	62	13	43	82.7	0.717	30	0	22	8	100.0	73.3
	0.4	62	13	52	82.7	0.867	30	0	20	10	100.0	66.7
	0.3	65	10	60	86.7	1.000	30	0	16	14	100.0	53.3
	0.2	66	9	72	88.0	1.200	30	0	11	19	100.0	36.7
	0.1	67	8	90	89.3	1.500	30	0	9	21	100.0	30.0

The results for each threshold from 0.1 to 0.9 are presented. Sensitivities are indicated as percentages. FP per case indicates the average number of FP counts per a case. Based on the results for the validation dataset, a threshold of 0.6 was defined as the standard value for the algorithm (given in bold numbers). Note that TNs for lesion-based analysis are omitted because TN lesion is undefinable for a data that contains multiple lesions in one scan and contains location information, unlike typical diagnostic test with a binary outcome (e.g., presence or absence). Abbreviations: DLA, deep learning-based algorithm; TP, true-positive; FN, false-negative; TN, true-negative; FP, false-positive.

Table 4. Sensitivities of the DLA and nine radiologists without and with the DLA, stratified according to lesion characteristics

	Validation	Test		
	DLA	DLA	Radiologists without DLA	Radiologists with DLA
Location				
Vertebra	90.9 (20/22)	89.7 (26/29)	60.2 (17.4/29)	78.5 (22.8/29)
Pelvis	93.3 (14/15)	92.9 (13/14)	60.3 (8.4/14)	87.3 (12.2/14)
Rib	77.8 (7/9)	77.8 (14/18)	35.8 (6.4/18)	54.9 (9.9/18)
Scapula	100.0 (1/1)	50.0 (2/4)	33.3 (1.3/4)	47.2 (1.9/4)
Limb	(0/0)	60.0 (3/5)	57.8 (2.9/5)	62.2 (3.1/5)
Sternum	100.0 (2/2)	80.0 (4/5)	44.4 (2.2/5)	77.8 (3.9/5)
Appearance				
Sclerotic	85.7 (18/21)	83.9 (26/31)	45.5 (14.1/31)	66.7 (20.7/31)
Lytic	89.5 (17/19)	92.0 (23/25)	64.4 (16.1/25)	84.0 (21.0/25)
Mixed	100.0 (9/9)	73.7 (14/19)	45.0 (8.6/19)	63.7 (12.1/19)
Diameter				
≥50 mm	100.0 (2/2)	75.0 (3/4)	86.1 (3.4/4)	94.4 (3.8/4)
≥30 mm to <50 mm	90.9 (10/11)	100.0 (13/13)	70.1 (9.1/13)	88.0 (11.4/13)
≥10 mm to <30 mm	96.8 (30/31)	85.7 (42/49)	50.6 (24.8/49)	73.0 (35.8/49)
≥5 mm to <10 mm	60.0 (3/5)	44.4 (4/9)	16.0 (1.4/9)	30.9 (2.8/9)
Total	89.8 (44/49)	82.7 (62/75)	51.7 (38.8/75)	71.7 (53.8/75)

Sensitivities are indicated as percentages, with actual numbers in parentheses. For numerators of radiologists' sensitivities, averages of nine radiologists are presented. Abbreviations: DLA, deep learning-based algorithm.

Table 5. Interpretation results of nine radiologists without and with the DLA

	wAFROC-FOM		Lesion-based sensitivity (%)		False-positives per case		Case-based sensitivity (%)		Case-based specificity (%)		Interpretation time per case (s)	
Radiologist	wo	w	wo	w	wo	w	wo	w	wo	w	wo	w
1	0.828	0.899	64.0	69.3	0.333	0.083	86.7	96.7	96.7	100.0	204	108
2	0.743	0.924	45.3	72.0	0.083	0.050	70.0	86.7	100.0	100.0	119	43
3	0.714	0.933	40.0	78.7	0.100	0.050	60.0	90.0	100.0	93.3	144	80
4	0.802	0.901	65.3	78.7	0.833	0.183	90.0	96.7	90.0	93.3	257	62
5	0.769	0.914	54.7	73.3	0.083	0.150	76.7	93.3	93.3	100.0	148	127
6	0.754	0.936	65.3	74.7	0.217	0.283	86.7	93.3	86.7	96.7	152	82
7	0.743	0.904	50.7	70.7	0.383	0.267	76.7	93.3	90.0	90.0	214	66
8	0.783	0.888	52.0	76.0	0.050	0.300	70.0	93.3	100.0	93.3	196	140
9	0.575	0.791	28.0	52.0	0.050	0.050	53.3	76.7	100.0	100.0	75	54
Average	0.746	0.899*	51.7	71.7*	0.237	0.157	74.4	91.1*	95.2	96.2	168	85*

Sensitivity and specificity are indicated as percentages, and interpretation times are indicated in seconds. Asterisks indicate a significant difference between the two sessions. Abbreviations: DLA, deep learning-based algorithm; wo, without DLA; w, with DLA; wAFROC-FOM, weighted alternative free-response receiver operating characteristic figure of merit.