Title

A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: retrospective training and validation of ResNet

Journal

European Spine Journal, 31(8)

ISSN

0940-6719

Authors

Yeh, Lee-Ren
Zhang, Yang
Chen, Jeon-Hor
et al.

Publication Date

2022-08-01

DOI

10.1007/s00586-022-07121-1

Peer reviewed

HHS Public Access

Author manuscript

Eur Spine J. Author manuscript; available in PMC 2022 December 20.

Published in final edited form as:

Eur Spine J. 2022 August ; 31(8): 2022–2030. doi:10.1007/s00586-022-07121-1.

A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: retrospective training and validation of ResNet

Lee-Ren Yeh[1], **Yang Zhang**[2], **Jeon-Hor Chen**[1,2], **Yan-Lin Liu**[2], **An-Chi Wang**[3], **Jie-Yu Yang**[3], **Wei-Cheng Yeh**[4], **Chiu-Shih Cheng**[1], **Li-Kuang Chen**[2], **Min-Ying Su**[2,5]

[1]Department of Radiology, E-Da Hospital and I-Shou University, Kaohsiung, Taiwan

[2]Department of Radiological Sciences, University of California, 164 Irvine Hall, Irvine, CA 92697-5020, USA

[3]Department of Radiology, Chi-Mei Medical Center, Tainan, Taiwan

[4]Department of Radiology, E-Da Cancer Hospital, Kaohsiung, Taiwan

[5]Department of Medical Imaging and Radiological Sciences, Kaohsiung Medical University, Kaohsiung, Taiwan

Abstract

Purpose—To improve the performance of less experienced clinicians in the diagnosis of benign and malignant spinal fracture on MRI, we applied the ResNet50 algorithm to develop a decision support system.

Methods—A total of 190 patients, 50 with malignant and 140 with benign fractures, were studied. The visual diagnosis was made by one senior MSK radiologist, one fourth-year resident, and one first-year resident. The MSK radiologist also gave the binary score for 15 qualitative imaging features. Deep learning was implemented using ResNet50, using one abnormal spinal segment selected from each patient as input. The T1W and T2W images of the lesion slice and its two neighboring slices were considered. The diagnostic performance was evaluated using tenfold cross-validation.

Results—The overall reading accuracy was 98, 96, and 66% for the senior MSK radiologist, fourth-year resident, and first-year resident, respectively. Of the 15 imaging features, 10 showed a significant difference between benign and malignant groups with $p < = 0.001$. The accuracy achieved by using the ResNet50 deep learning model for the identified abnormal vertebral segment

✉Jeon-Hor Chen, jeonhc@hs.uci.edu.
Lee-Ren Yeh and Yang Zhang equally contributed to the article.

Conflict of interest There was no conflict of interest/competing interest about this work. No relevant financial activities outside the submitted work.

Code availability Custom code can be shared on reasonable request.

Ethics approval This study was approved by the Institutional Review Board.

Consent to participate The informed consent was waived due to the retrospective analysis nature of this study.

Consent for publication Not Applicable. No subject identifiable information is published.

was 92%. Compared to the first-year resident's reading, the model improved the sensitivity from 78 to 94% ($p < 0.001$) and the specificity from 61 to 91% ($p < 0.001$).

Conclusion—Our deep learning-based model may provide information to assist less experienced clinicians in the diagnosis of spinal fractures on MRI. Other findings away from the vertebral body need to be considered to improve the model, and further investigation is required to generalize our findings to real-world settings.

Keywords

Automated differential diagnosis; Benign spinal fractures; Less experienced radiologists; Malignant spinal fractures

Introduction

Imaging plays an important role in the evaluation of spinal diseases. Benign and malignant vertebral fractures may present similar features and challenging for diagnosis in some cases, especially for inexperienced trainees. Studies have shown that the misdiagnosis rate of vertebral fractures can be as high as 20% [1]. Appropriate differentiation and staging of benign osteoporotic, traumatic, and malignant fractures are essential for therapy planning, especially in the acute and subacute stages. However, benign vertebral lesions occur in approximately one-third of cancer patients [1], and fractures from minor trauma are commonly seen in the elderly, which can further complicate the evaluation and diagnosis of malignant lesions.

In clinical settings, images acquired using various modalities are evaluated by radiologists and other clinicians, and the diagnostic accuracy depends on their specialty and level of experience [2, 3]. For diagnosis of fractures, orthopedics are known to be more accurate than general physicians, and among orthopedics, the accuracy is also dependent on their specialty training [4]. In detecting spinal abnormalities, neuroradiologists are in general more experienced than body radiologists [5].

For diagnosis of spinal lesions, MRI is the most helpful imaging modality. When the vertebral fat-containing yellow bone marrow is edematous or replaced by enough cancer cells, signal intensity changes can be seen on T1-weighted (T1W), T2-weighted (T2W), and fat-suppressed images acquired using short tau inversion recovery (STIR) [6, 7]. However, even after combining information from images acquired using all the sequences, accurate diagnosis remains challenging in patients with ambiguous features [8].

Recently, artificial intelligence (AI)-based imaging analysis has emerged as a popular method due to its potential to provide a comprehensive evaluation of imaging features, which can be used to aid in diagnosis of many diseases. Machine learning methods have been developed to anatomically localize and categorize vertebral fractures on radiographs [4, 9] and CT images [10]. The AI diagnostic tools can provide a decision support system, not only to improve diagnostic accuracy for less experienced radiologists, but also to improve workflow efficiency for all radiologists.

The purpose of this study is to apply an automatic deep learning with Residual Network-50 (ResNet50) algorithm [11] to distinguish between benign and malignant vertebral fractures on MRI. The results were compared to the diagnosis made by three radiologists with various level of training, including one experienced MSK attending and two residents, to investigate whether and how the AI model can assist less experienced radiologists. The MSK radiologist also assigned scores to a panel of imaging features, and these were used to characterize the wrong diagnoses made by trainees to investigate the features that may be emphasized in further training to improve their accuracy.

Materials and methods

Patients

The dataset was selected from the radiological reporting system in a period of 4 years, using the key words *fracture, vertebral collapse, pathological*, and *metastasis*. A total of 190 patients were included (mean age 66.5, range 23–95 years old), 140 with benign fractures (mean age 68.8) and 50 with malignant fractures (mean age 61.7). Malignant cases had either biopsy-proven cancer or a known history of primary tumor with progressive disease. The most common primary cancer came from the lung, followed by colon/rectum, breast, and prostate. All benign cases had no known cancer history and had been followed and confirmed with stable disease. This retrospective study was approved by the Institutional Review Board with waiver of informed consent.

MRI protocols

All subjects received MR imaging of the spine on a 1.5 T scanner (GE Signa Excite, Milwaukee, Wisconsin, USA). Imaging sequences included axial and sagittal non-fat-sat spin-echo T1-weighted, axial and sagittal non-fat-sat fast spin-echo T2-weighted, and coronal fast spin-echo T2-weighted fat-sat sequences. The imaging parameters of the two sequences used for analysis were as follows: sagittal spin-echo T1-weighted sequence with TR = 400 ms, TE = 15 ms, matrix size = 320 × 192, field of view = 30 cm, and slice thickness = 4 mm; and sagittal fast spin-echo non-fat-sat T2-weighted sequence with TR = 3200 ms, TE = 90 ms, matrix size = 448 × 224, field of view = 30 cm, and slice thickness = 4 mm. These images were reconstructed into a matrix of 512 × 512.

Visual assessment of MR images and diagnostic reading

The analysis flowchart is shown in Fig. 1. The first part was the visual reading by three radiologists. An MSK radiologist (LRY, with 28 years of experience) performed reading and gave a binary score to each of 15 qualitative features, as listed in Table 1. Based on all features, the radiologist gave a final diagnostic impression of benign versus malignant fracture for each patient. To compare the diagnosis performed by less experienced radiologists, two residents, one in the fourth year of training (ACW) and the other in the first year of training (JYY), were given the dataset to evaluate. For each patient, they gave a final diagnostic impression of benign or malignant.

Deep learning architecture

The second part of analysis in Fig. 1 was the AI analysis using deep learning, by using the most prominent abnormal vertebra in each patient as the input, marked by another experienced body radiologist (JHC). The abnormal region was first identified on sagittal T2W images. A square box containing the entire abnormal vertebra was generated and used as the input. The defined box was mapped onto T1W images using linear registration. The input of network included both T1W and T2W images of the identified slice with its two neighboring slices that also contained the lesion. Therefore, the total number of input channel was six. The bounding box was resized to 64 × 64 by linear interpolation. The intensities of each patch were normalized to a mean of zero and a standard deviation of one.

The ResNet50 architecture was applied to differentiate between benign and malignant groups, shown in Fig. 2. While convolutional neural networks (CNN) such as VGG or AlexNet learn features using large, convolutional network architectures [12], the ResNet can extract residual features, as subtraction of features learned from input of that layer, using “skip connections” [13]. The ResNet50 architecture contains one 3 × 3 convolutional layer, one max pooling layer, and 16 residual blocks. Each block contains one 1 × 1 convolutional layer, one 3 × 3 convolutional layer and one 1 × 1 convolutional layer. The residual connection is from the beginning to the end of the block. The output of the last block is connected to a fully connected layer with sigmoid function to provide the prediction. With ResNet, since it is pre-trained with photographs with RGB colors, only three sets of images can be used in input channel [13]. Thus, a convolutional layer with 1 × 1 filter was added to extract inter-channel features and transform from six channels to three channels.

Augmentation and training configurations

The following methods were used to compensate for the small case number and the imbalance between benign and malignant cases in the dataset. Each slice was used as an independent input. The benign dataset was augmented 20 times by using random affine transformations including translation, scaling, and rotation. To balance the fewer number of malignant cases, the malignant dataset was augmented 40 times. To control for overfitting, a L2 regularization term was added to the final loss function and, during the training process, early stop was applied based on the lowest validation loss to obtain the optimal model. The loss function was cross-entropy. The training was implemented using the Adam optimizer [14]. The learning rate was set to 0.0001 with momentum term β to 0.5 to stabilize training. Parameters were initialized using ImageNet [15]. The batch size was set to 32, and the number of epochs was set to 100.

Evaluation using cross-validation

The classification performance of ResNet50 was evaluated using tenfold cross-validation, and each case had only one chance to be included in the validation group. The prediction results based on 2D slices meant that each slice had its own diagnostic probability. For the per-patient diagnosis, the highest probability of malignancy among all slices of each patient was assigned to that patient. The malignancy probability obtained for each case was used to make the final diagnosis using the threshold of 0.5.

Statistical analysis

The diagnostic results of three radiologists and ResNet analysis were compared to the ground truth to determine true positive (TP), true negative (TN), false negative (FN), and false positive (FP) cases, and from these, the sensitivity, specificity, and overall accuracy were calculated. The diagnostic sensitivity and specificity between two readers or between each reader and the AI model were compared by the McNemar test. The 15 binary scores read by the senior MSK radiologist were analyzed by Fisher's exact test to examine the significance of the association (contingency) between benign and malignant groups with confidence interval of 0.95. Then, the scores were combined to develop a classification model using logistic regression, and the model accuracy was evaluated.

Results

Diagnostic performance based on radiologist's reading

The scores of 15 imaging features are shown in Table 1. Of these, 10 features showed significant differences between the two groups with $p < = 0.001$. The diagnostic results of three radiologists are listed in Table 2. The senior MSK radiologist's accuracy was 0.98. The fourth-year resident also had very high accuracy of 0.96, and not surprisingly, the first-year resident performed poorly with accuracy of 0.66.

Analysis of MSK radiologist's reading features

When individual scores of 15 features were used to build a logistic regression model, the diagnostic accuracy was 0.94. Intravertebral mass-like or nodular lesion, epidural/paraspinal soft tissue mass, and pedicle and posterior element involvement were rarely seen in benign cases, and thus, strongly indicated malignancy. Diffuse signal changes occurred more frequently in the malignant (44/50, 88%) than in the benign group (31/140, 22%). Homogeneous marrow signal (no marrow edema or infiltration) and intravertebral dark lines or bands (representing impaction of bone trabeculae) were present only in benign fractures and, thus, specific benign features. Those cases without marrow signal change were considered to be old or chronically healed fractures with resolution of marrow edema. Irregular dark patches in the vertebrae, on the other hand, were found in both groups with similar incidence (10% vs. 10%) and may represent osteoblastic change, chronic hemorrhage, fibrotic component in tumor, or sclerosis, fibrosis, cement (for vertebroplasty) in benign fracture.

ResNet50 diagnostic performance

When deep learning using ResNet50 was applied, the accuracy was 0.92 for per-patient diagnosis. There were 3 false negative diagnoses and 12 false positive cases, with sensitivity of 47/50 = 94% and specificity of 128/140 = 91%. Figure 3 shows two malignant cases correctly diagnosed as true positives. Figure 4 shows two benign cases correctly diagnosed as true negatives. Figure 5 shows two malignant cases misdiagnosed as benign, and Fig. 6 shows two benign cases misdiagnosed as malignant. These misdiagnosed cases by deep learning were all correctly diagnosed by the senior MSK radiologist, and the important features are described in the figure legends. The diagnostic sensitivity and specificity of the

ResNet model and the three readers are listed in Table 2, and the difference was compared using the McNemar test. The sensitivity of the MSK attending, R4, R1 resident, and ResNet model was 94, 94, 78, and 94%, respectively. The sensitivity of R1 was significantly worse compared to the others ($p < 0.001$). The specificity of the MSK attending, R4, R1 resident, and ResNet model was 99, 96, 61, and 910%, respectively. The performance of R1 was significantly worse compared to the others ($p < 0.001$), and also, the specificity of the ResNet model was worse compared to the MSK attending and the R4 resident ($p < 0.001$).

Error analysis in R1 resident's diagnosis

In order to investigate the features in cases that the R1 resident gave wrong diagnosis, the feature scores determined by the senior MSK radiologist was used as references for comparison. In benign cases, the R1 reader had a high 54/140 (39%) false positive rate. The features were compared to those of 86 true negative (61%) cases. Of them, 3 features showed significant differences. The false positive cases were less likely to present "Homogeneous marrow signal" 6/54 (11%) vs. 42/86 (49%), $p < 0.01$ and more likely to present "Diffuse signal change of vertebral body > 3/4" 20/54 (37%) vs. 11/86 (13%), $p < 0.01$. In Table 1, "Homogeneous marrow signal" was found in 0% malignant lesion, and "Diffuse signal change of vertebral body > 3/4" was found in 88% malignant cases. For benign cases showing heterogeneous marrow or diffuse signal change, they were likely to be misinterpreted by R1 as false positive diagnosis, as shown in Fig. 7. Another feature likely to lead to false positive was "Band pattern bone marrow edema," 24/54 (44%) in FP vs. 20/86 (23%) in TN, $p = 0.01$.

Discussion

This study investigated the potential of deep learning to diagnose benign and malignant vertebral fractures based on T1W and T2W MRI, and compared the performance to the reading of three radiologists with different levels of experience. The motivation was to investigate whether an AI tool can assist less experienced radiologists to diagnose spinal fractures. The results showed that deep learning using ResNet50 achieved a satisfactory diagnostic accuracy of 92%. Although it was inferior to 98% made by a senior MSK radiologist and 96% made by an R4 resident, much higher compared to 66% made by an R1 resident. Compared to R1's reading, the AI model improved the sensitivity from 78 to 94% ($p < 0.001$) and the specificity from 61 to 91% ($p < 0.001$).

MRI features that help differentiate between benign and malignant vertebral fractures have been well studied [6, 7, 16]. Table 1 shows the evaluation of 15 features, several of which had good diagnostic implications. Detection of an epidural or paraspinal soft tissue mass, pedicle and posterior element involvement, intravertebral mass-like or nodular lesion, and coexisting skipped nodular bone marrow replacement in other vertebrae were found to be specific features of malignancy, while band pattern bone marrow lesions and intravertebral dark lines or bands suggested benign fractures [6, 16]. The trabeculae of malignant fractures were destroyed before the vertebra collapse and, theoretically, lost the chance for formation of an impacted trabecular band. Diffuse marrow replacement, anteroposterior protrusion of the vertebral body [6], and non-wedged collapse (central

concave deformity and compression of entire body) were features shared by both benign and malignant fractures. Fracture or collapse in other levels was not a specific sign since both osteoporotic fracture and malignant fracture could exist in the same patient, especially in the elderly. When the fractured vertebra showed equivocal features or both features of benign and malignant collapse, the diagnosis might be difficult and challenging, especially for less experienced radiologists.

The inferior performance of ResNet50 compared to experienced readers might be partly explained by the limited input information, since only a small bounding box covering the abnormal vertebral body was considered as the input. On the other hand, the radiologist could assess all information on the entire image, which included the epidural/paraspinal soft tissue mass, pedicle and posterior element involvement, and bone marrow replacement in other vertebrae that were considered as specific features related to malignancy. The 92% accuracy of the ResNet model was much better compared to that of the inexperienced R1 resident and may have a clinical value when there is a shortage of well-trained medical staff. Furthermore, there is room for further improvement to develop a more accurate AI model, e.g., by considering more inputs from adjacent tissues, more imaging sequences, more imaging planes, etc. As revealed by the experienced MSK radiologist, other features away from the vertebral body may provide very useful diagnostic information, and these need to be considered in future model development to improve the performance.

The development of AI-based methods, especially using fully automatic deep learning, has potential to assist radiologists in making accurate diagnoses with more confidence and also to be integrated into the clinical workflow and improve efficiency [17]. In this study, ResNet50 was used as the architecture of the convolutional neural network, and it was typically done using all 2D slices as individual inputs. The L2 norm regularization, dropout and data augmentation, were applied to control overfitting. In per-slice analysis using tenfold cross-validation, the AUC's were > 0.90 in all runs, suggesting that the trained model was robust and not over-fitted.

Lee et al. gave a general overview for the application of deep learning in medical imaging [18]. Several studies have applied this method for diagnosis of bone fractures on plain radiography [4, 19–21] as summarized in a recent review paper by Yang et al. [19]. Chung et al. [4] used a pre-trained ResNet152 to classify proximal humerus fractures using plain anteroposterior shoulder radiographs. Olczak et al. [20] analyzed wrist, hand and ankle radiographs using five different neural networks to detect body parts and fractures. Kitamura et al. [21] used another three different network architectures, including Inception V3, ResNet, and Xception, to differentiate abnormal from normal radiographs and reached the highest accuracy of 0.8. In addition, deep learning has also been applied to CT and MR images [10, 22–25]. Raghavendra et al. [22] applied deep learning to distinguish normal from thoracolumbar spine injuries. Tomita et al. [23] implemented a sophisticated CNN algorithm using pre-trained ResNet34 Network and Long Short-Term Memory (LSTM) to classify the osteoporotic vertebral fractures and normal subjects using CT scans and achieved 89.2% accuracy. Pedoia et al. [24] employed deep learning using DenseNet for the prediction of osteoarthritis on MRI. Bien et al. [25] developed a model using MRNet for diagnosis of abnormalities, anterior cruciate ligament (ACL) tears, and meniscal tears

on knee MRI. All these studies were designed for diagnosis of abnormalities. In the present study, we attempted to differentiate benign from malignant fractures using deep learning, which was much more challenging and has to our knowledge not yet been reported in literature, and thus, there are no results to compare with.

We compared reading made by 3 radiologists. Training new radiologists to interpret vertebral fracture is an effort that requires significant time and resources. Among the multiple barriers is the difficulty of recognizing, weighing, and synthesizing of features that favor malignant versus benign conditions. Frequently the signs of malignant and benign lesions coexist in the same patient or even in the same vertebra, and radiologists must therefore establish their own “weighting” system to determine the overall probability and establish a final diagnosis. This is a process with a steep and long learning curve that beginning radiologists usually become frustrated with.

There were a total of 140 benign cases, and the R1 resident only correctly diagnosed 86 of them (61%). In the four years residency program, usually the first-year residents have not had experience in interpretation of MR MSK images and that could explain the low accuracy. When examining the 54 wrong vs. 86 correct cases, it was found that FP cases were less likely to present “Homogeneous marrow signal” and more likely to present “Diffuse signal change of vertebral body > 3/4.” Homogeneous marrow signal is one of the reliable imaging features for the diagnosis of benign vertebral fracture. The less frequent appearance of this feature might mislead the R1 to falsely diagnose the case as malignancy. On the contrary, diffuse signal intensity changes of the vertebral body occurred more frequently in the malignant group (Table 1). The R1 might use this finding as the dominant imaging feature and gave wrong diagnosis. For the R4 resident, since he has completed one or two rotations at the MSK section and had enough knowledge in MR physics and signal intensity presentations, his diagnostic accuracy was much higher at 96%.

This study has several limitations. First, it is a pilot study to demonstrate feasibility, and the case number was relatively small. Second, only patients with metastatic cancer were included in the malignant group. Third, in the AI analysis, only one vertebral body segment in a patient was selected for analysis. In the future, a localization strategy including vertebra alignment segmentation and abnormality search can be integrated into the AI analysis. For automatic spinal segmentation, several studies have implemented CNN strategies and obtained satisfactory performance [26–29]. These methods may be implemented to first segment each vertebral body, and then, the imaging features inside and away from the vertebral body can be all integrated in an AI model to give a malignancy or abnormality probability. Lastly, we did not provide the model prediction results to the R1 resident, as done in Bien et al. [25], to investigate how the AI tool can be used as a decision support system to improve the diagnostic performance of human readers. When more cases are available, this can be tested in general physicians, radiologists, and orthopedic surgeons with different levels of training to investigate its clinical utility in real-world settings.

Conclusion

This study investigated the application of deep learning for the differential diagnosis of benign and malignant vertebral fracture on MRI. These results suggest that ResNet50 provides a feasible method to use T1-weighted and T2-weighted images on MRI to establish a diagnosis. The per-patient diagnostic accuracy was 92%, which was inferior to reading of radiologists who had sufficient training, but much higher than that of an inexperienced radiologist. The results suggest that the developed ResNet50 model may be used as an assisting decision support system in facilities lack of well-trained medical staff. With further technical improvement of the model, and specific refinement of the model in each clinical setting, this AI-based method has the potential to serve as a clinical tool to help less experienced readers and to improve workflow.

Funding

This study was supported by E-Da Hospital intramural seed grant EDAHM108003, NIH R01 CA127927, P30 CA062203 and the UC Irvine Comprehensive Cancer Center using UCI Anti-Cancer Challenge funds. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Chao Family Comprehensive Cancer Center.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Avellino AM, Mann FA, Grady MS et al. (2005) The misdiagnosis of acute cervical spine injuries and fractures in infants and children: the 12-year experience of a level I pediatric and adult trauma center. Childs Nerv Syst 21:122–127. 10.1007/s00381-004-1058-4 [PubMed: 15609065]
2. Casez P, Uebelhart B, Gaspoz J-M, Ferrari S, Louis-Simonet M, Rizzoli R (2006) Targeted education improves the very low recognition of vertebral fractures and osteoporosis management by general internists. Osteoporos Int 7(7):965–970. 10.1007/s00198-005-0064-z
3. Goradia D, Blackmore CC, Talner LB, Bittles M, Meshberg E (2005) Predicting radiology resident errors in diagnosis of cervical spine fractures. Acad Radiol 12(7):888–893. 10.1016/j.acra.2005.04.004 [PubMed: 16039542]
4. Chung SW, Han SS, Lee JW et al. (2018) Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 89:468–473. 10.1080/17453674.2018.1453714 [PubMed: 29577791]
5. Zhou AL, Bonham LW, Verde F (2018) Comparative analysis of body radiologist to neuroradiologist evaluation of the spine in trauma settings. J Am Coll Radiol 15(12):1687–1691. 10.1016/j.jacr.2018.03.002 [PubMed: 29804826]
6. Schwaiger BJ, Gersing AS, Baum T, Krestan CR, Kirschke JS (2016) Distinguishing benign and malignant vertebral fractures using CT and MRI. Semin Musculoskelet Radiol. 20(4):345–352. 10.1055/s-0036-1592433 [PubMed: 27842427]
7. Baker LL, Goodman SB, Perkash I, Lane B, Enzmann DR (1990) Benign versus pathologic compression fractures of vertebral bodies: assessment with conventional spin-echo, chemical-shift, and STIR MR imaging. Radiology 174:495–502. 10.1148/radiology.174.2.2296658 [PubMed: 2296658]
8. Diacinti D, Vitali C, Gussoni G et al. (2017) Misdiagnosis of vertebral fractures on local radiographic readings of the multicentre POINT (Prevalence of Osteoporosis in INTernal medicine) study. Bone 101:230–235. 10.1016/j.bone.2017.05.008 [PubMed: 28511873]

9. Gan K, Xu D, Lin Y et al. (2019) Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop 90(4):394–400. 10.1080/17453674.2019.1600125 [PubMed: 30942136]
10. Burns JE, Yao J, Summers RM (2017) Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiology 284:788–797. 10.1148/radiol.2017162100 [PubMed: 28301777]
11. Bengio Y (2009) Learning deep architectures for AI. Foundations and trends® in Machine Learning. FNT Machine Learn. 2:1–127. 10.1561/2200000006
12. LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. pp 3361:1995
13. He K, Zhang X, Ren S, Sun J editors (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
14. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint https://arxiv.org/pdf/1412.6980.pdf
15. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L, editors (2009) Imagenet: A large-scale hierarchical image database. Computer Vision and Pattern Recognition. In: 2009 IEEE conference on computer vision and pattern recognition (CVPR), pp 248–255. https://ieeexplore.ieee.org/abstract/document/5206848/
16. Takigawa T, Tanaka M, Sugimoto Y, Tetsunaga T, Nishida K, Ozaki T (2017) Discrimination between malignant and benign vertebral fractures using magnetic resonance imaging. Asian spine journal 11:478. 10.4184/asj.2017.11.3.478 [PubMed: 28670417]
17. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ (2018) Artificial intelligence in radiology. Nat Rev Cancer 18:500–510. 10.1038/s41568-018-0016-5 [PubMed: 29777175]
18. Lee J-G, Jun S, Cho Y-W et al. (2017) Deep learning in medical imaging: general overview. Korean J Radiol 18:570–584. 10.3348/kjr.2017.18.4.570 [PubMed: 28670152]
19. Yang S, Yin B, Cao W, Feng C, Fan G, He S (2020) Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. Clin Radiol 75:713.e17–713.e28. 10.1016/j.crad.2020.05.021
20. Olczak J, Fahlberg N, Maki A et al. (2017) Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 88:581–586. 10.1080/17453674.2017.1344459 [PubMed: 28681679]
21. Kitamura G, Chung CY, Moore BE (2019) Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imaging 32:672–677. 10.1007/s10278-018-0167-7 [PubMed: 31001713]
22. Raghavendra U, Bhat NS, Gudigar A, Acharya UR (2018) Automated system for the detection of thoracolumbar fractures using a CNN architecture. Futur Gener Comput Syst 85:184–189. 10.1016/j.future.2018.03.023
23. Tomita N, Cheung YY, Hassanpour S (2018) Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med 98:8–15. 10.1016/j.compbiomed.2018.05.011 [PubMed: 29758455]
24. Pedoia V, Lee J, Norman B, Link T, Majumdar S (2019) Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort. Osteoarthritis Cartilage 27:1002–1010. 10.1016/j.joca.2019.02.800 [PubMed: 30905742]
25. Bien N, Rajpurkar P, Ball RL et al. (2018) Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 15:e1002699. 10.1371/journal.pmed.1002699 [PubMed: 30481176]
26. Chen H, Shen C, Qin J, et al. editors (2015) Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In: 2015 international conference on medical image computing and computer-assisted intervention (MICCAI), pp 515–522
27. Whitehead W, Moran S, Gaonkar B, Macyszyn L, Iyer S editors (2018) A deep learning approach to spine segmentation using a feed-forward chain of pixel-wise convolutional networks. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI), pp 868–871
28. Sekuboyina A, Kukačka J, Kirschke JS, Menze BH, Valentinitsch A editors (2017) Attention-driven deep learning for pathological spine segmentation. In: international workshop and challenge

on computational methods and clinical applications in musculoskeletal imaging (MSKI), pp 108–119

29. Lu J-T, Pedemonte S, Bizzo B et al. (2018) Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. arXiv preprint https://arxiv.org/abs/1807.10215

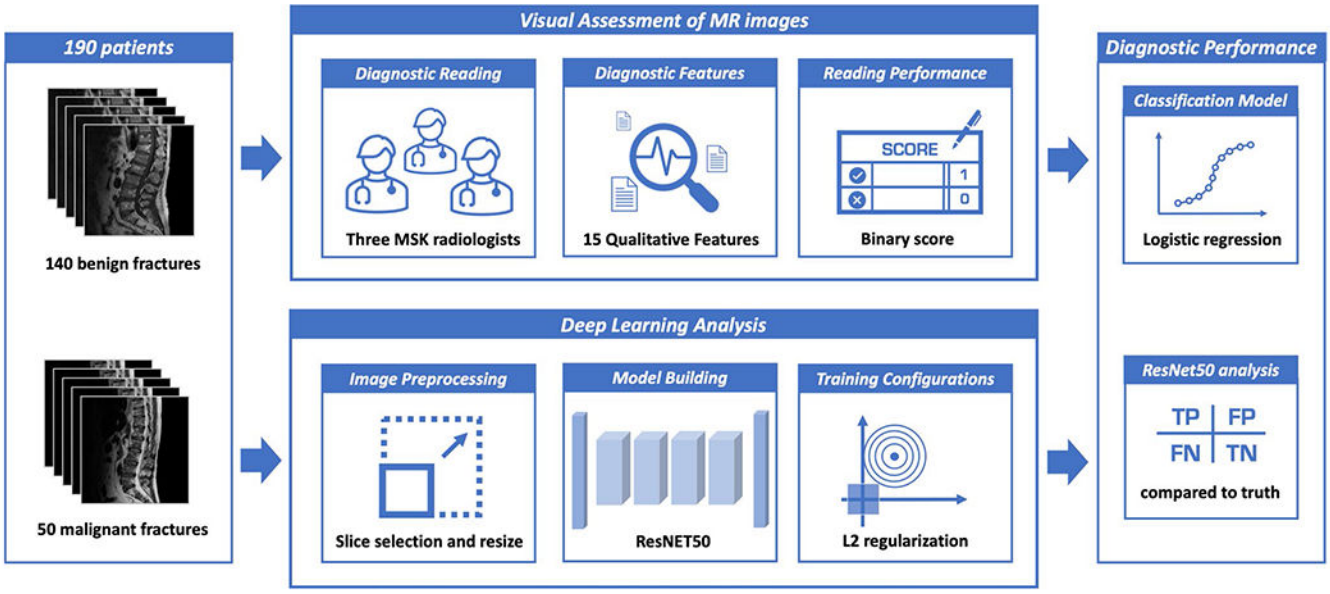

Fig. 1.

The overall analysis flowchart. The cases include 50 malignant and 140 benign patients, each with T1W and T2W MR images. The visual reading is performed by 3 radiologists, to give a diagnostic impression of malignant or benign for each case. The deep learning analysis is performed using the ResNet50 architecture, evaluated by tenfold cross-validation. The results are compared

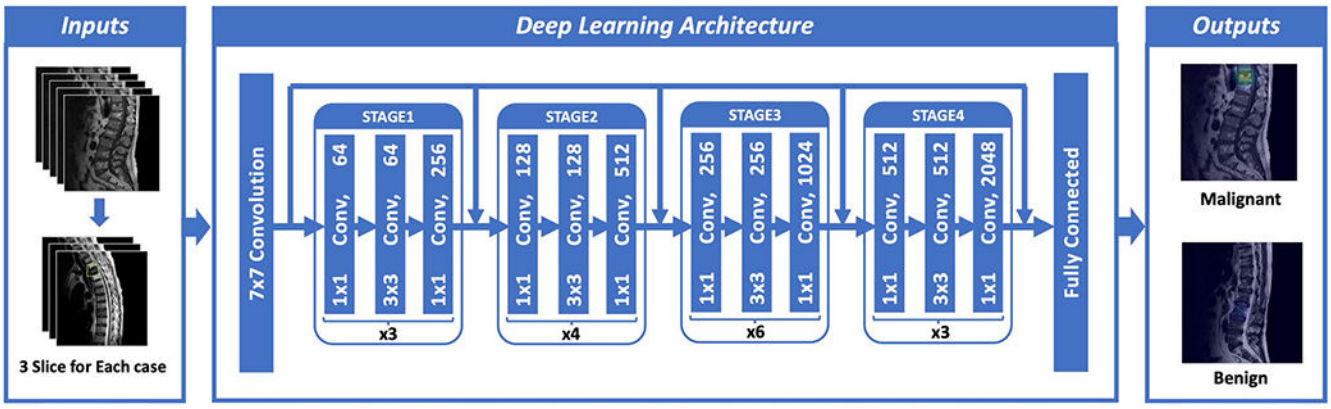

Fig. 2.

Architecture of ResNet50, containing 16 residual blocks. Each residual block begins with one 1 × 1 convolutional layer, followed by one 3 × 3 convolutional layer and ends with another 1 × 1 convolutional layer. The output is then added to the input via a residual connection. The total input number is 6: T1W and T2W of the slice with its two neighboring slices, so one convolutional layer with 1 × 1 filter is added before ResNet to extract inter-channel features and transform from 6 to 3 channels as input

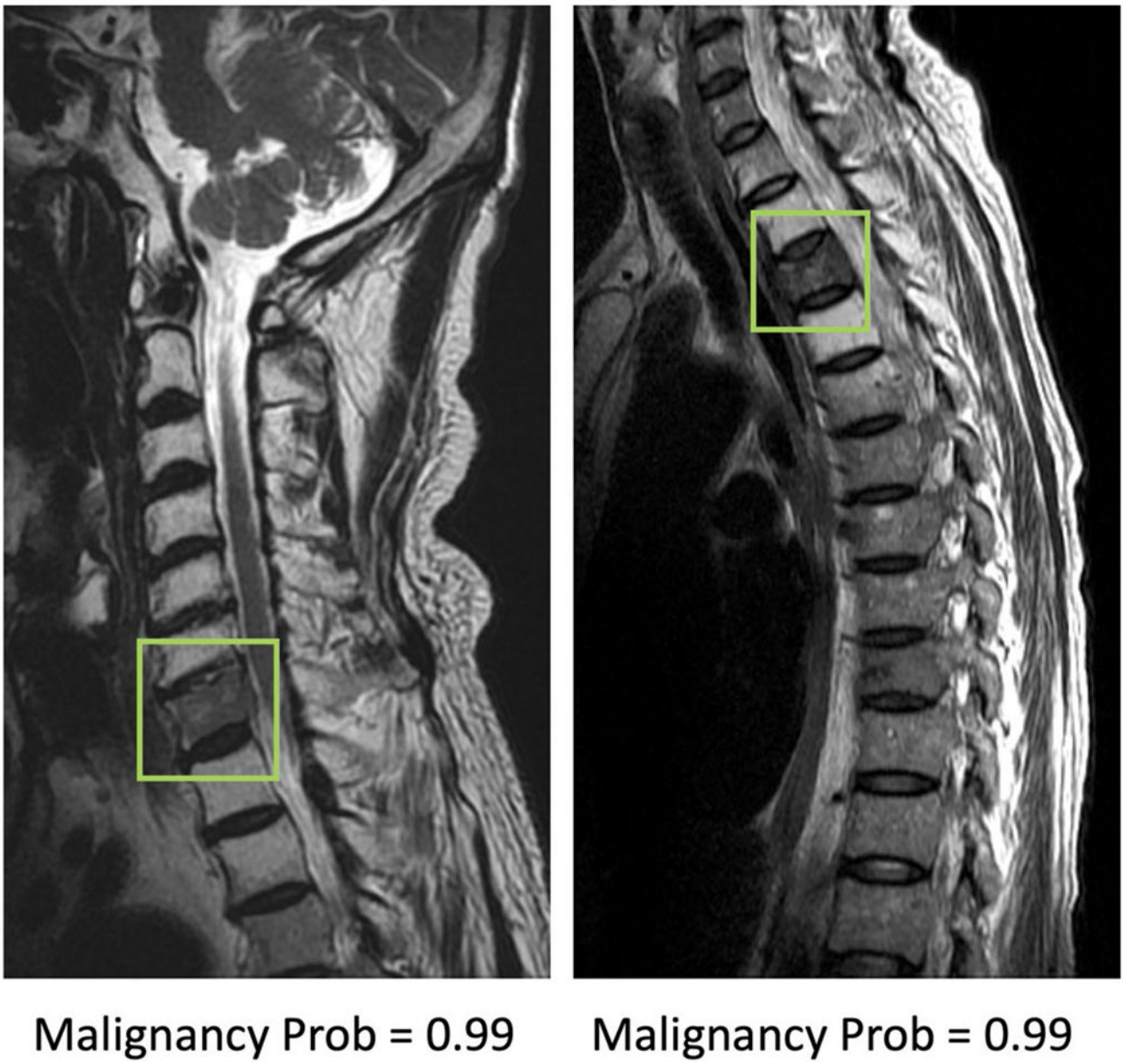

Fig. 3.

Two true positive malignant cases. The image at left panel shows diffuse tumor infiltration at the 7th cervical (C7) vertebral body with posterior cortical destruction and no apparent collapse. The image at right panel shows diffuse tumor infiltration at third thoracic (T3) vertebra with anterior wedge deformity. The fatty change of other cervical vertebrae in the left panel and T2/T4 vertebrae in right panel is post-radiation effect

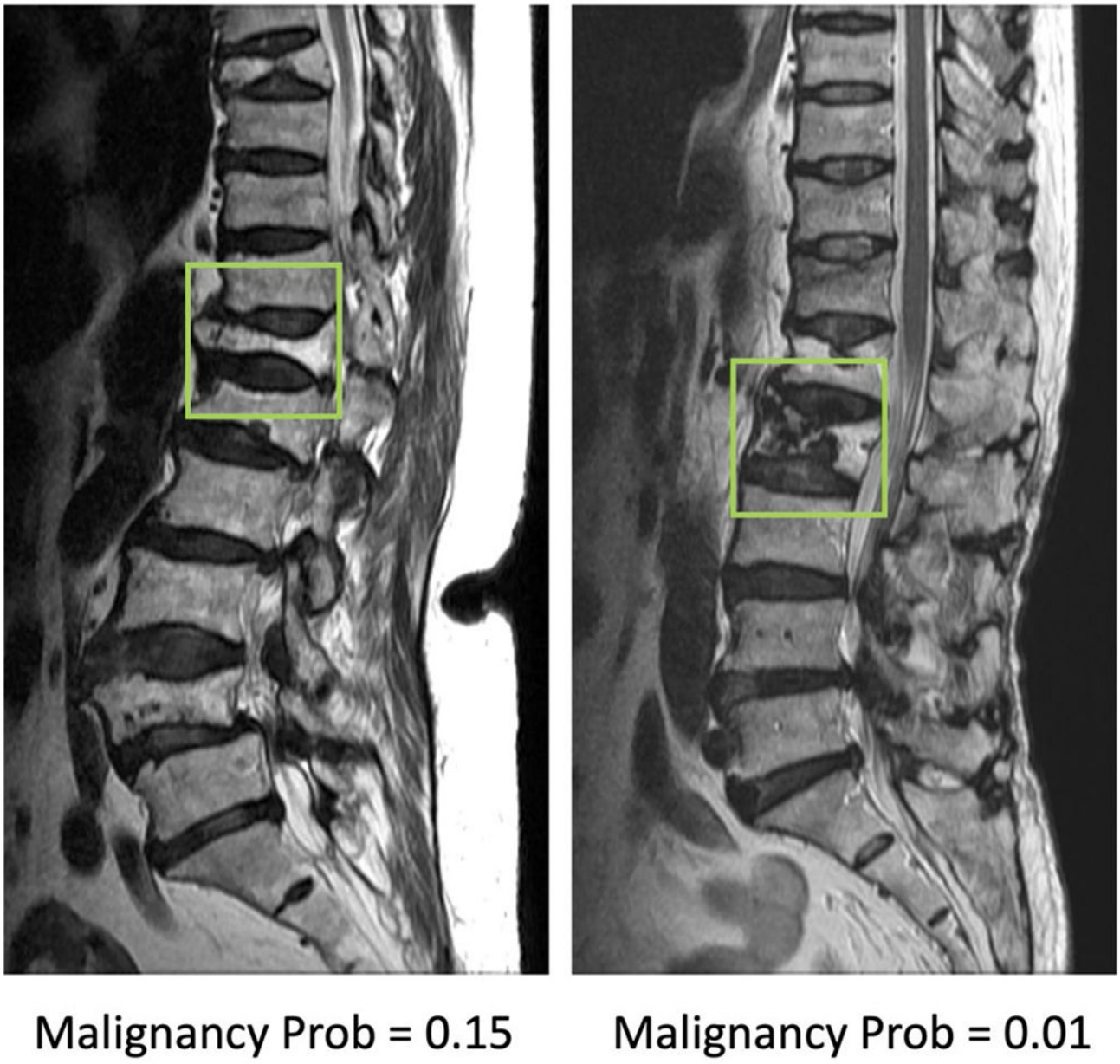

Fig. 4.

Two true negative benign cases. The left case is a chronic benign osteoporotic fracture with resolution of bone marrow edema. Although with severe collapse, the height of posterior vertebral body is still preserved. The right case is a chronic osteoporotic fracture with prior vertebroplasty. The irregular dark patch in the vertebra represents the cement material of vertebroplasty. Both cases show fractures in several other vertebrae

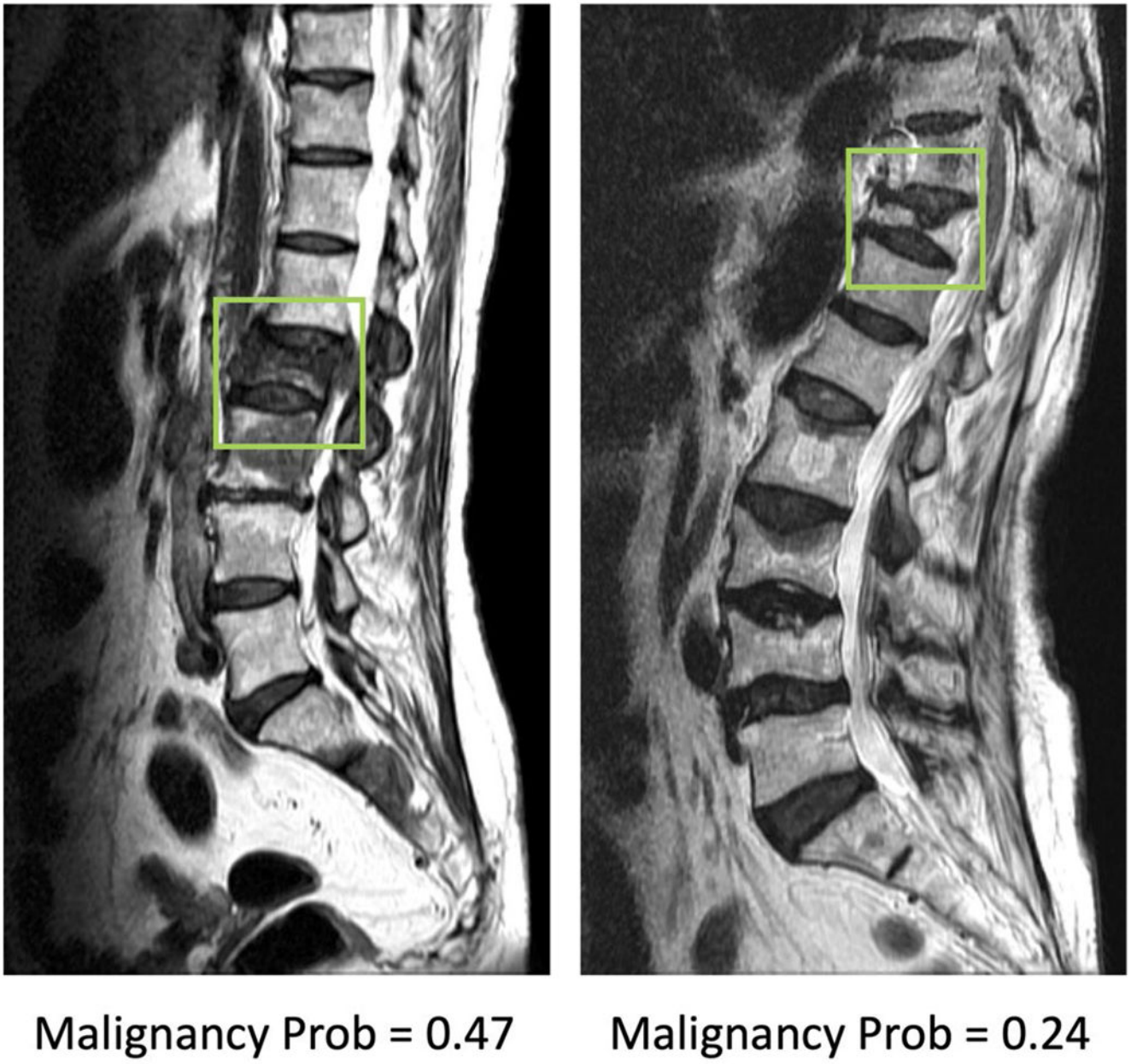

Fig. 5.

Two false negative cases, malignant fracture misdiagnosed as benign. The image at left panel shows diffuse signal change and paravertebral soft tissue mass at L2 vertebra. The coexisted metastatic mass at L3 and S2 vertebrae are also noted. The right case shows diffuse tumor infiltration, necrotic cleft, central concave collapse, and paravertebral soft tissue mass

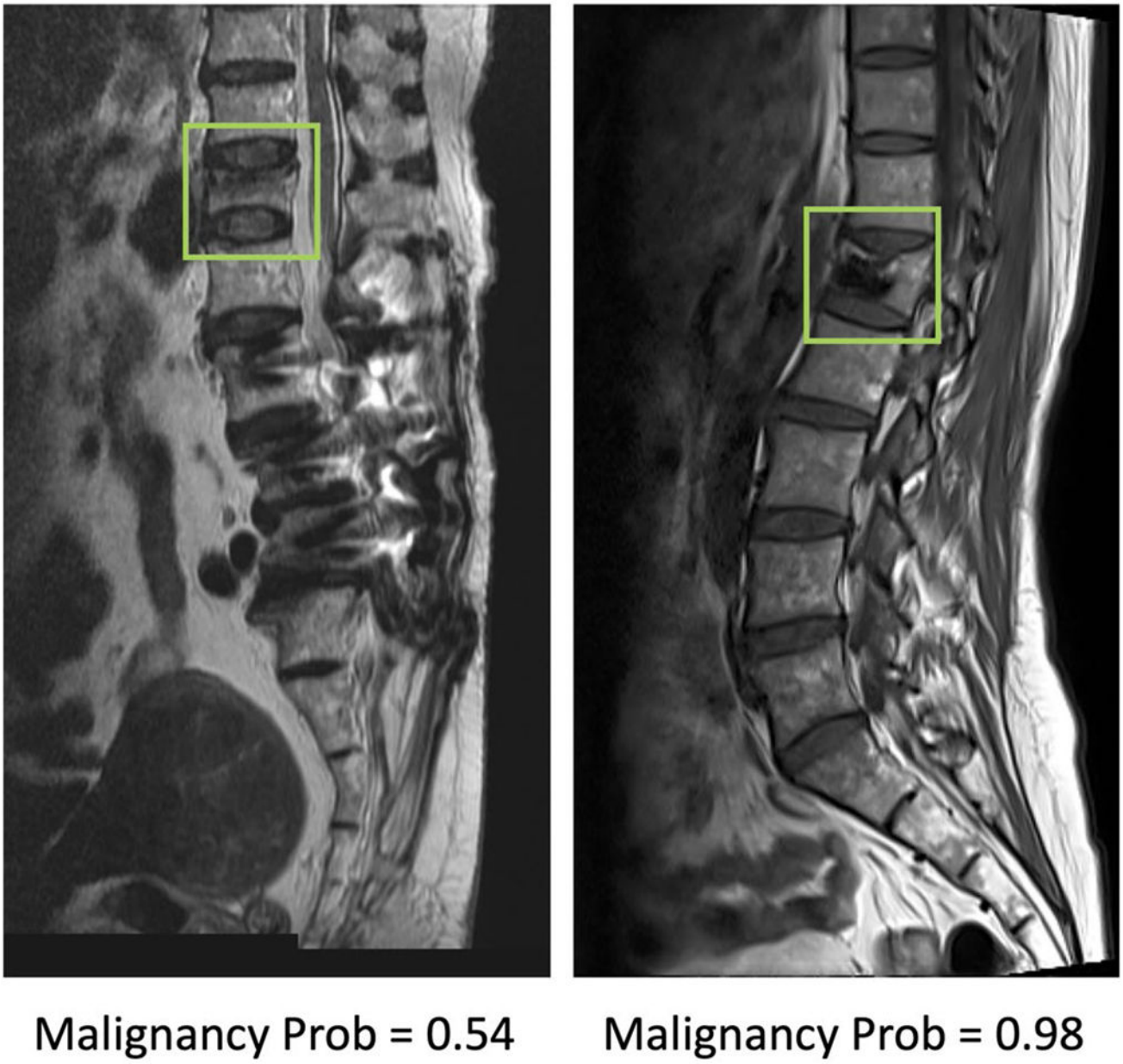

Fig. 6.

Two false positive cases, benign fracture misdiagnosed as malignant. The left case is a recent benign fracture with typical band pattern marrow edema. The right case is a benign fracture post cement vertebroplasty

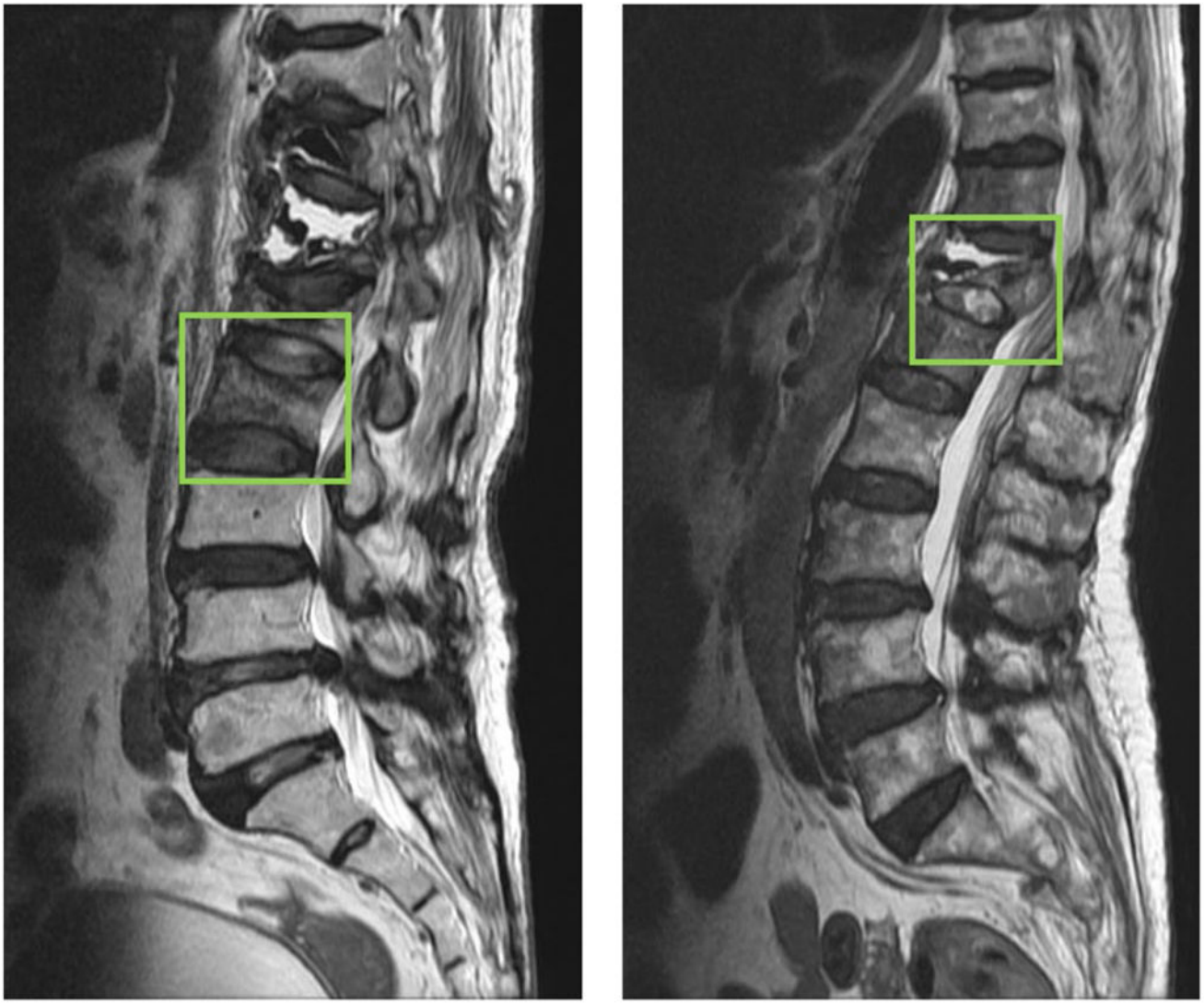

Fig. 7.

Example of two false positive cases diagnosed by the first-year inexperienced resident. Left case: The presence of heterogeneous marrow signal is falsely diagnosed as malignancy. Right case: The presence of diffuse signal intensity change (marrow edema or replacement) of vertebral body > 3/4 is wrongly diagnosed as malignancy

Table 1

Qualitative features evaluated by an experienced radiologist

Feature Name	Malignant *N* = 50	Benign *N* = 140	*P*-Value
Absence of collapse	17/50 (34%)	3/140 (2%)	< 0.001
Anterior wedge deformity (preserved posterior vertebral height)	7/50 (14%)	77/140 (55%)	< 0.001
Compression of entire body	25/50 (50%)	54/140 (39%)	0.18
Central concave deformity	14/50 (28%)	59/140 (42%)	0.09
Homogeneous marrow signal (No marrow edema or infiltration)	0/50 (0%)	48/140 (41%)	< 0.001
Intravertebral dark line, band	0/50 (0%)	37/140 (26%)	< 0.001
Band pattern bone marrow edema	2/50 (4%)	44/140 (31%)	< 0.001
Intravertebral dark patch	5/50 (10%)	14/140 (10%)	1
Fluid or necrotic cleft	1/50 (2%)	8/140 (6%)	0.045
Diffuse signal change (marrow edema or replacement) of vertebral body > 3/4	44/50 (88%)	31/140 (22%)	< 0.001
Intravertebral mass-like or nodular lesion	11/50 (22%)	0/140 (0%)	< 0.001
Anterior/posterior protrusion of vertebral body	16/50 (32%)	19/140 (14%)	0.07
Epidural/paraspinal soft tissue mass	22/50 (44%)	1/140 (1%)	< 0.001
Pedicle and posterior element involvement	5/50 (10%)	0/140 (0%)	0.001
Coexisted skipped nodular lesion or mass-like bone marrow replacement in other vertebra	39/50 (78%)	8/140 (6%)	< 0.001

Table 2

Diagnostic performance of three radiologists and ResNet50 deep learning model

	TP	TN	FN	FP	Sensitivity	Specificity	Accuracy
MSK Radiologist	47	139	3	1	47/50 (94%)	139/140 (99%)	186/190 (98%)
Resident R4	47	135	3	5	47/50 (94%)	135/140 (96%)	182/190 (96%)
Resident R1	39	86	11	54	39/50 (78%)	86/140 (61%)	125/190 (66%)
ResNet50 Model	47	128	3	12	47/50 (94%)	128/140 (91%)	175/190 (92%)

TP True Positive; *TN* True Negative; *FN* False Negative; *FP* False Positive