Original Article

Deep learning-based detection of primary bone tumors around the knee joint on radiographs: a multicenter study

Danyang Xu[1#], Bing Li[2#], Weixiang Liu[3], Dan Wei[4], Xiaowu Long[5], Tanyu Huang[6], Hongxin Lin[2], Kangyang Cao[2], Shaonan Zhong[2], Jingjing Shao[1], Bingsheng Huang[2], Xian-Fen Diao[3], Zhenhua Gao[1,4]

[1]Department of Radiology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China; [2]Medical AI Lab, School of Biomedical Engineering, Health Science Centre, Shenzhen University, Shenzhen, China; [3]National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Medical School, Shenzhen University, Shenzhen, China; [4]Department of Radiology, Huiya Hospital of The First Affiliated Hospital, Sun Yat-sen University, Huizhou, China; [5]Department of Radiology, Yunfu People's Hospital, Yunfu, China; [6]Department of Radiology, The Second People's Hospital of Huizhou, Huizhou, China

Contributions: (I) Conception and design: Z Gao, XF Diao, B Huang, W Liu; (II) Administrative support: Z Gao, B Huang; (III) Provision of study materials or patients: D Xu, D Wei, X Long, T Huang, J Shao, Z Gao; (IV) Collection and assembly of data: D Xu, D Wei, X Long, T Huang, J Shao, Z Gao; (V) Data analysis and interpretation: B Li, H Lin, K Cao, S Zhong, XF Diao, B Huang, Z Gao, D Xu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Bingsheng Huang, PhD. Medical AI Lab, School of Biomedical Engineering, Health Science Centre, Shenzhen University, No. 1066 Xueyuan Avenue, Fuguang Community, Taoyuan Street, Nanshan District, Shenzhen 518055, China. Email: huangb@szu.edu.cn; Xian-Fen Diao, Doctor of Engineering. National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Medical School, Shenzhen University, No. 1066 Xueyuan Avenue, Fuguang Community, Taoyuan Street, Nanshan District, Shenzhen 518055, China. Email: laodiao@szu.edu.cn; Zhenhua Gao, MD. Department of Radiology, The First Affiliated Hospital, Sun Yat-sen University, No. 58 Zhongshan Er Road, Guangzhou 510080, China; Department of Radiology, Huiya Hospital of The First Affiliated Hospital, Sun Yat-sen University, No. 186 Zhongxing North Road, Dayawan District, Huizhou 516081, China. Email: gaozhh@mail.sysu.edu.cn.

Background: Most primary bone tumors are often found in the bone around the knee joint. However, the detection of primary bone tumors on radiographs can be challenging for the inexperienced or junior radiologist. This study aimed to develop a deep learning (DL) model for the detection of primary bone tumors around the knee joint on radiographs.

Methods: From four tertiary referral centers, we recruited 687 patients diagnosed with bone tumors (including osteosarcoma, chondrosarcoma, giant cell tumor of bone, bone cyst, enchondroma, fibrous dysplasia, etc.; 417 males, 270 females; mean age 22.8±13.2 years) by postoperative pathology or clinical imaging/follow-up, and 1,988 participants with normal bone radiographs (1,152 males, 836 females; mean age 27.9±12.2 years). The dataset was split into a training set for model development, an internal independent and an external test set for model validation. The trained model located bone tumor lesions and then detected tumor patients. Receiver operating characteristic curves and Cohen's kappa coefficient were used for evaluating detection performance. We compared the model's detection performance with that of two junior radiologists in the internal test set using permutation tests.

Results: The DL model correctly localized 94.5% and 92.9% bone tumors on radiographs in the internal and external test set, respectively. An accuracy of 0.964/0.920, and an area under the receiver operating characteristic curve (AUC) of 0.981/0.990 in DL detection of bone tumor patients were for the internal and external test set, respectively. Cohen's kappa coefficient of the model in the internal test set was significantly higher than that of the two junior radiologists with 4 and 3 years of experience in musculoskeletal radiology (Model *vs.* Reader A, 0.927 *vs.* 0.777, P<0.001; Model *vs.* Reader B, 0.927 *vs.* 0.841, P=0.033).

Conclusions: The DL model achieved good performance in detecting primary bone tumors around the knee joint. This model had better performance than those of junior radiologists, indicating the potential for the detection of bone tumors on radiographs.

Keywords: Bone neoplasms; knee joint; deep learning (DL); radiography

Submitted Dec 08, 2023. Accepted for publication May 30, 2024. Published online Jul 12, 2024.
doi: 10.21037/qims-23-1743
View this article at: https://dx.doi.org/10.21037/qims-23-1743

Introduction

Bone tumors are a group of primary or secondary bone neoplastic lesions with various tumor types and biological behaviors (1). Primary bone tumors, such as osteosarcomas and giant cell tumors of bone, are common around the knee joint, including the distal femur, proximal tibia, and proximal fibula (2-4). Malignant bone tumors are relatively rare, comprising 0.2% of human cancers overall and 5–6% in 15- to 24-year-old (5,6). Digital radiography (DR) is recognized as a first-line imaging modality for the evaluation of bone lesions owing to its ability to evaluate the location, internal matrix, and borders of bone lesions, and that it has fast acquisition and is cost-efficient compared with computed tomography (CT) and magnetic resonance imaging (MRI) (7,8). However, plain radiography is a two-dimensional imaging modality, and radiographs have relatively low contrast resolution compared with CT images (9). Thirty to 50% of trabecular bones may have been destroyed before a lesion is visible to the naked eye on plain radiographs (10). Therefore, some bone lesions are occluded on radiographs and can be incorrectly diagnosed by visual observation in daily clinical practice (11). Many junior radiologists and general practitioners may not have developed sufficient training experience to identify and assess bone tumors on radiographs.

Artificial intelligence has been applied in analyzing radiographs for computer-aided diagnosis of bone tumors (11-21). A recent meta-analysis showed that clinicians' sensitivity in diagnosing bone tumors increased with AI assistance, and in certain cases, the machines outperformed human experts (22). von Schacky *et al.* (17) developed a multitask deep learning (DL) model for the simultaneous detection, segmentation, and classification of primary bone tumors on radiographs that correctly detected 59.5% (66 of 111) of bone tumors with an intersection over union (IoU) >0.5 or 82.0% (91 of 111) if the threshold was set to IoU >0. Li *et al.* (18) developed a YOLO model for detection and classification of bone lesions on radiographs in several bones that correctly placed 86.36% (114 of 132) and 85.27% (191 of 224) of the bounding boxes in the internal and external validation sets (IoU >0.5). Breden *et al.* (20) used X-rays of 176 pediatric patients with bone tumor in a single center to develop a DL model for detection of bone tumors around the knee, with an accuracy of 89.1% in the internal test groups. Hinterwimmer *et al.* (21) developed an algorithm to link undiagnosed patients to previous patient histories based on radiographs and simultaneous classification of multiple bone tumors, which achieved the highest mean accuracy, precision and recall (92.86%, 92.86% and 34.08%). These studies have demonstrated the potential role of DL in bone tumor detection. The most common location of primary bone tumors is around the knee joint (23). To the best of our knowledge, no multicenter study of DL for the detection of knee primary bone tumors on radiographs in large series has been reported.

The purpose of this study was to develop a DL model for the localization of primary bone tumors on knee radiographs and differentiating bone tumor patients and from non-tumor patients. Moreover, we aimed to compare its detection performance with that of junior radiologists. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1743/rc).

Methods

Dataset

This study was performed in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by Ethics Committee of the First Affiliated Hospital, Sun Yat-sen University (No. [2022]541) with a waiver for written informed consent due to the retrospective nature of the study. All participating hospitals/institutions were informed and agreed with the study. In

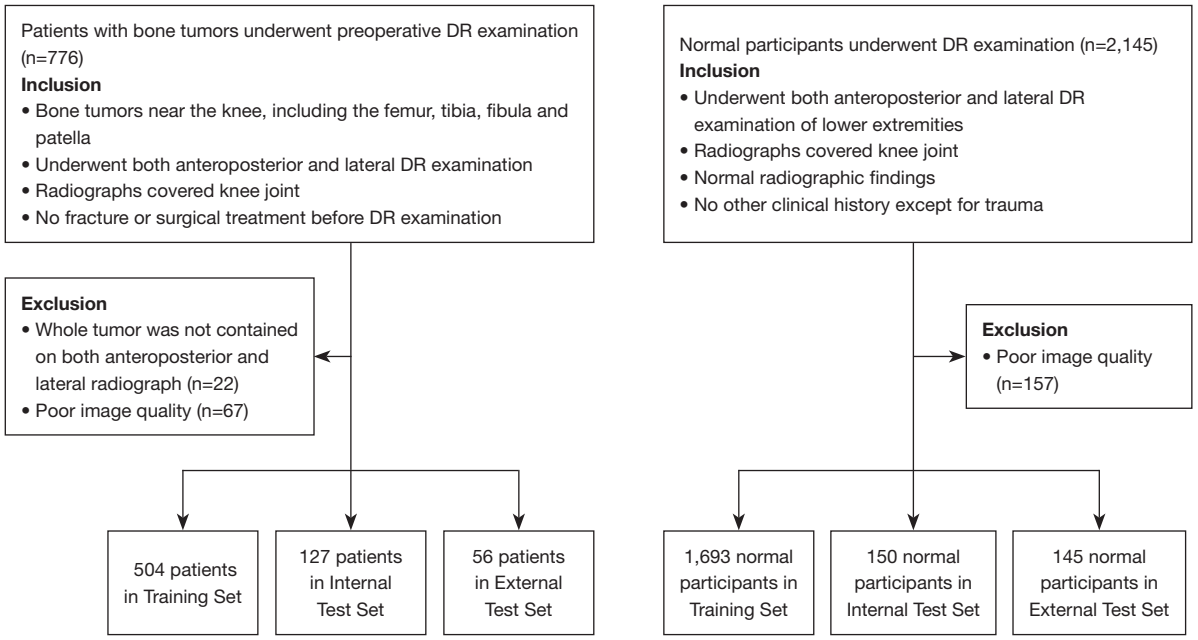

Figure 1 Inclusion and exclusion criteria. DR, digital radiography.

this multicenter study from four centers between January 2003 and June 2022, we included 776 patients with bone tumors who were diagnosed by radiography combined with CT or MRI/clinical follow-up or proven by histopathology according to the 5th World Health Organization classification of bone tumors (24). We also included 2,145 normal participants who were enrolled jointly by two senior radiologists (Z.G. and D.W.) with more than 10 years of experience in reading musculoskeletal radiographs. These four centers in this study are tertiary referral centers in our country.

Inclusion and exclusion criteria are shown in *Figure 1*. Two radiologists (Lixian Chen and Bingkun Guo, with five years and three years of experience in reading musculoskeletal radiographs, respectively) reviewed all images. Images were excluded if both radiologists considered there to be the presence of low image contrast, position errors, motion artifacts, or overlapping of foreign bodies on lesions. The final dataset in these four centers consisted of 687 tumor patients (417 males, 270 females; mean age 22.8±13.2 years) and 1,988 normal participants (1,152 males, 836 females; mean age 27.9±12.2 years). Each patient or participant contained an anteroposterior view and a lateral view of radiographs. These anteroposterior and lateral knee radiographs obtained from the Picture Archiving and Communication System (PACS, Shanghai Atlas Tiger Medical Information Systems Co., Ltd., China; Guangzhou Lijin Digital Medical Systems Co., Ltd., China; Shenzhen Annet Information System Co., Ltd., China) in Digital Imaging and Communication in Medicine (DICOM) format were analyzed.

The characteristics of all subjects are shown in *Table 1*. In this study, the maximum longitudinal diameter of a lesion in an affected bone on a radiograph was used to represent the tumor size.

For model development and internal testing in centers 1, 2, 3, the dataset was split into a training dataset (including 504 tumor patients and 1,693 normal participants) and an internal independent test set (including 127 tumor patients and 150 normal participants). The 127 tumor patients in the internal independent test set were composed of 20% of tumor patients in centers 1, 2, and 3, respectively, i.e., 109 patients from center 1, 13 patients from center 2, 5 patients from center 3. The 150 non-tumor participants in the independent test set consisted of 50 normal participants per center randomly selected from centers 1, 2, and 3. The remaining data from centers 1, 2, and 3 was used as the training set for model development. The internal independent test set was used to evaluate and compare the performance of the junior radiologists and DL model. The external independent test set including 56 tumor patients and 145 normal participants from center 4 was used to

Table 1 Characteristics of the subjects

Characteristics	Center 1	Center 2	Center 3	Center 4
Number of subjects	1,060	977	437	201
Age (years)				
<18	495 (46.7)	156 (16.0)	163 (37.3)	35 (17.4)
≥18	565 (53.3)	821 (84.0)	274 (62.7)	166 (82.6)
Average age (mean ± SD) in years[a]	22.0±12.4	30.8±11.1	26.1±14.4	30.2±1.6
Sex (male/female)[b]	625 (59.0)/435 (41.0)	586 (60.0)/391 (40.0)	265 (60.6)/172 (39.4)	93 (46.3)/108 (53.7)
Number of tumor patients[c]	544	61	26	56
Age (years)				
<18	294 (54.0)	12 (19.7)	7 (26.9)	9 (16.1)
≥18	250 (46.0)	49 (80.3)	19 (73.1)	47 (83.9)
Tumor size (mm)				
<80	230 (42.3)	57 (93.4)	24 (92.3)	46 (82.1)
≥80	314 (57.7)	4 (6.6)	2 (7.7)	10 (17.9)
Malignant (osteosarcoma, chondrosarcoma)	428 (78.7)	3 (4.9)	2 (7.7)	14 (25.0)
Intermediate (giant cell tumor of bone)	116 (21.3)	1 (1.6)	0 (0.0)	42 (75.0)
Benign (bone cyst, enchondroma, fibrous dysplasia, etc.)	0 (0.0)	57 (93.4)	24 (92.3)	0 (0.0)
Average tumor size (mean ± SD) in mm[d]	97.3±47.2	32.6±23.4	44.4±27.9	66.1±25.7

The number in parentheses represents the percentage of cases in a particular class. [a], the age of the subjects in different centers were statistically significant (P<0.001) by using the Kruskal-Wallis test; [b], there were significant differences in sex among subjects in different centers by using the Chi-squared test; P=0.003; [c], the ratio among tumor types in different centers was statistically significant by using the Chi-squared test; P<0.001; [d], the tumor size of the tumor patients in different centers were statistically significant (P<0.001) by using the Kruskal-Wallis test. SD, standard deviation; mm, millimeters.

evaluate the performance of the DL model.

Data annotation for tumors on radiographs

All original digital radiographs were downloaded in DICOM format from the PACS. All identifiable personal and sensitive information of the subjects on the radiographs was sufficiently anonymized. DICOM images were converted to 8-bit gray Joint Photographic Experts Group (JPEG) format at their original resolution and window width/level using MicroDicom software (Version 3.8.1.422, MicroDicom Ltd., Bulgarian). No further adjustment of the image window length/width was made during the image format conversion. The pixel values of the two-dimensional array in each DICOM file were normalized by scaling values into the range (0–255). Normalized two-dimensional arrays were converted to 3-channel JPEG images by repeating the two-dimensional array three times. Then the JPEG images were loaded into LabelImg software (Version 1.8.1, an open-source image labelling program based on Python, https://github.com/heartexlabs/labelImg). A senior radiologist (Z.G., with 16 years of experience in reading musculoskeletal radiographs) annotated all bone tumor lesions with bounding boxes, which constituted the reference standard for automatic bone tumor detection. Examples of the reference standard for bone tumor detection are shown in *Figure 2*. The rectangular bounding box encompassed the entirety of each lesion with reference to the corresponding CT or MR images, including some adjacent soft tissue and normal bone.

Development of the DL model

All images in JPEG format were padded with zeros and

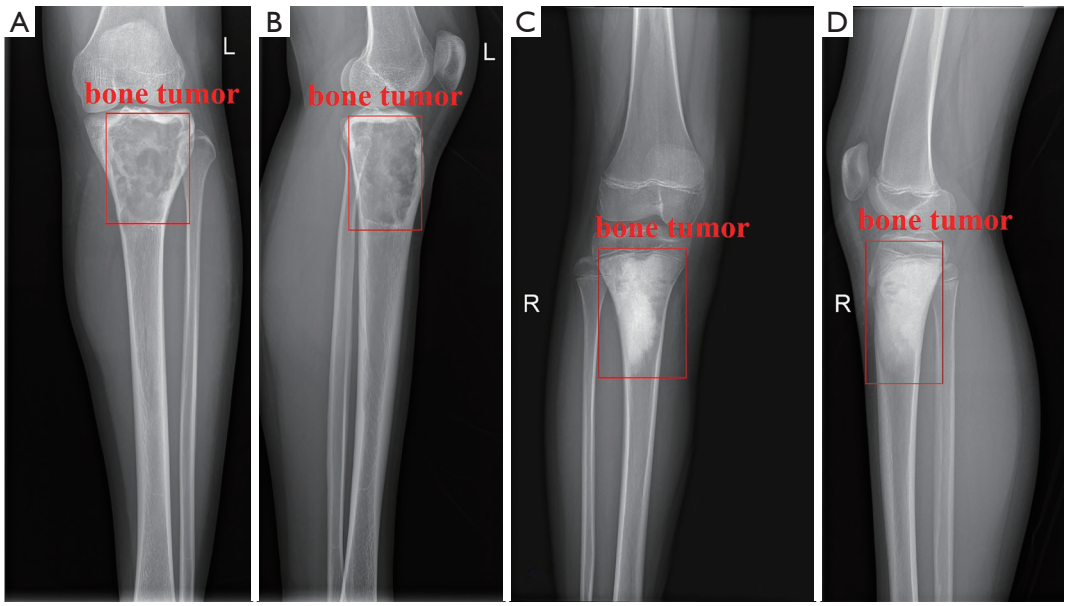

Figure 2 Examples of the diagnostic reference standard for automatic bone tumor detection. (A,B) Anteroposterior (A) and lateral (B) radiographs of the knee of a patient with a giant cell tumor of bone. (C,D) Anteroposterior (C) and lateral (D) radiographs of the knee of a patient with osteosarcoma. Note that the reference standard was placed to fit the margin of the tumor. L, left; R, right.

resized to 3 by 1,280 pixels. Then, the pixel values were normalized by scaling the values into the range (0–1) for input into the DL model. The output of the DL model is the tumor location and the predicted score of the tumor (confidence score).

DL models based on the YOLOv5 convolutional neural network (CNN) architecture (25) were constructed for bone tumor detection. Model weights were initialized with those pretrained with data from the common objects in context (COCO) database (26). The detailed architecture of YOLOv5 used was provided in Appendix 1. The data augmentation methods used during the training process included mosaic, affine transformation (image scaling, translation and flipping), image enhancement (hue, saturation and value transformation). The detailed training parameters are provided in Appendix 1. The training set for model development was further divided into five disjoint partitions of patients, and fivefold cross-validation was used to analyze the model performance.

Independent testing

The internal independent test set (554 radiographs from 277 subjects) and external independent test set (402 radiographs from 201 subjects) were used to evaluate the performance of the proposed method. The test-time augmentation method (27) was used for inference in the test set (see Appendix 1) to increase sample diversity and improve model generalization during independent testing. The results of the test set were obtained by non-maximum suppression across all boxes predicted by the five models selected in cross-validation.

DL model evaluation

In the tumor localization task, the IoU was used to assess the model's performance. In this study, true-positive (TP) tumor localization required the IoU between the bounding box localized by the DL model and the reference bounding box to be greater than 0.2 (28).

For the patient detection task, the receiver operating characteristic (ROC) analysis was used to evaluate the binary discriminatory capacity with different prediction scores, and accuracy, sensitivity, and specificity were also calculated. Each patient contained an anteroposterior view and a lateral view of radiographs. If the model outputs at least one positive diagnostic result on both anteroposterior and lateral radiographs, the patient was considered to have tumor. The standard setting for determining the model diagnostic results was that tumoral bone radiograph was positive and normal bone radiograph was negative. The

patient was considered a normal participant only when the model outputs were negative diagnostic results on both the anteroposterior and lateral radiographs.

Observer evaluation

Two junior radiologists (Readers A and B, with four years and three years of experience, respectively, in reading musculoskeletal radiographs) evaluated the radiographs of bone lesions in DICOM format and were blinded to the histopathologic and clinical data. The radiologists used DICOM Viewer 2.2.9 software (Medixant Company, Poland) to read the radiographs. Then, for each bone radiography, they used LabelImg software (Version 1.8.1) to place a bounding box in the area they considered as a bone tumor. No box was provided for a normal radiograph. At the same time, a stopwatch was used to record the time taken by each radiologist to evaluate all radiographs in the internal test set. They just have been timed for making the diagnosis when reading radiographs, not recording the time taken by radiologists to place a bounding box around the region of the tumor. Note that these two radiologists were unaware of this study.

The bone lesions were considered correctly localized (TP lesion) when the IoU between the bounding box localized by the radiologist and the reference bounding box was greater than or equal to 0.2. If at least one TP lesion was found on a patient's anteroposterior or/and lateral radiographs, the patient was classified as a tumor patient (29). If no bounding boxes are found on a patient's anteroposterior and lateral radiographs, the patient was classified as a normal participant.

Statistical analysis

The Kruskal-Wallis test was used to compare the age and tumor size of patients in different centers. The Chi-squared test was used to compare the ratios of tumor types and sex among different centers and the accuracy of different subgroups. Cohen's kappa coefficient was used to compare the detection performance of the DL model and two radiologists in the internal independent test set. Permutation tests were used to calculate P values. Among centers 1–3, subgroup analyses based on center, age (<18 and ≥18 years old), tumor size (less than median tumor size and greater than or equal to median tumor size), and imaging device were performed in the internal independent test set. All analyses were conducted using MedCalc statistical software (Version 20.0.9.0, MedCalc Software Ltd., Belgium) and Python 3.8.5. All tests were two-sided, and $P<0.05$ was considered statistically significant.

Results

Subject characteristics

Table 1 summarizes the characteristics of the dataset used in this study. The study cohort consisted of 2,675 subjects. The age range of patients was 3–70 years. A total of 25.7% of the cohort were bone tumor patients (447 malignant, 159 intermediate, and 81 benign tumors), while 74.3% were normal participants. There was only one tumor lesion in 681 patients and two tumor lesions in 6 patients. The median size of the bone tumors was 80 mm in centers 1, 2 and 3.

DL model performance

For the tumor localization, the DL model correctly placed 94.5% (242 of 256 tumor lesions) of the bounding boxes in the internal test set and 92.9% (104 of 112 tumor lesions) of the bounding boxes in the external test set (IoU >0.2). Only 3.5% (9 of 256) bone tumor lesions in the internal validation set and 5.4% (6 of 112) bone tumor lesions in the external validation set were not detected (IoU =0). Fifty and thirty-two false-positive lesions were detected in the internal and external test set, respectively.

The detection performance of the DL model is shown in *Table 2* and *Figure 3*. The area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity of the DL model in the internal test set were 0.981 [95% confidence interval (CI): 0.957–0.993], 0.964, 0.976, and 0.953, respectively. The stratification analysis for different centers, age groups, and tumor sizes showed very close performance among subgroups with different centers, ages or tumor sizes. The AUC, accuracy, sensitivity, and specificity of the DL model in the external test set were 0.990 (95% CI: 0.964–0.999), 0.920, 0.982 and 0.897, respectively. *Figure 3* depicts the ROC curves for classifying subjects into normal participants or tumor patients. No significant differences in the accuracy of the DL model for detecting bone tumors in the internal independent test set were shown among different X-ray imaging devices (P=0.183) in *Table 3*.

Table 2 The detection performance of the DL model in the independent internal and external test set

Characteristics	AUC (95% CI)	Accuracy	Sensitivity	Specificity
Internal test set	0.981 (0.957–0.993)	0.964 (267/277)	0.976 (124/127)	0.953 (143/150)
Centers[a]				
Center 1	1.000 (0.977–1.000)	0.981 (156/159)	1.000 (109/109)	0.940 (47/50)
Center 2	0.865 (0.755–0.938)	0.952 (60/63)	0.846 (11/13)	0.980 (49/50)
Center 3	0.884 (0.769–0.955)	0.927 (51/55)	0.800 (4/5)	0.940 (47/50)
Age (years)				
<18	1.000 (0.955–1.000)	0.988 (79/80)	1.000 (63/63)	0.941 (16/17)
≥18	0.961 (0.924–0.984)	0.954 (188/197)	0.953 (61/64)	0.955 (127/133)
Tumor size (mm)				
<80	–	0.967 (58/60)	0.967 (58/60)	–
≥80	–	0.985 (66/67)	0.985 (66/67)	–
External test set	0.990 (0.964–0.999)	0.920 (185/201)	0.982 (55/56)	0.897 (130/145)

[a], there were no significant differences in the accuracy of the DL model (P=0.156) in different centers by using the Chi-squared test. AUC, area under the receiver operating characteristic curve; CI, confidence interval; mm, millimeters; DL, deep learning.

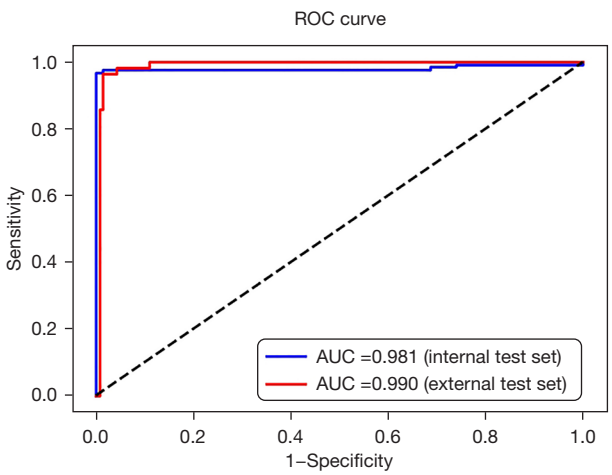

Figure 3 ROC curve of the DL model for detecting bone tumors. ROC, receiver operating characteristic; AUC, area under the receiver operating characteristic curve; DL, deep learning.

Performance comparison between the DL model and observer assessments

Cohen's kappa coefficient between the two radiologists' scores was 0.748 (95% CI: 0.670–0.826), indicating substantial agreement between the two radiologists. The results for the DL model and two radiologists in the independent test set are shown in *Table 4*. The accuracies of Reader A, Reader B, and the DL model were 0.888, 0.921, and 0.964, respectively. The accuracy of the DL model was higher than that of junior radiologists. Cohen's kappa coefficients between the DL model and the diagnostic reference standard in classifying a subject as a normal participant or a tumor patient was significantly higher than those of radiologists (DL model *vs.* Reader A: 0.927 *vs.* 0.777, P<0.001; DL model *vs.* Reader B: 0.927 *vs.* 0.841, P=0.033). There were no significant differences in the accuracy of the DL model and two radiologists for detecting bone tumors between different subgroups based on age and tumor size (all P>0.100). The time taken by the two radiologists and the DL model to evaluate 554 radiographs from 277 subjects was 192, 204, and 2.7 min, respectively.

The confusion matrices for the DL model and two junior radiologists are shown in *Figure 4*. Incorrect diagnoses include missed diagnoses and misinterpreted diagnoses. Of the 277 subjects in the test set, 31 and 22 subjects were diagnosed incorrectly by the two junior radiologists compared to 10 subjects diagnosed incorrectly by our DL model. Furthermore, 11 of 12 subjects with a missed diagnosis of bone tumor by both junior radiologists were found by the DL model (*Figure 5*). On the other hand, 34 of 35 subjects with a misinterpreted diagnosis of bone tumor by both junior radiologists were interpreted correctly by the DL model (*Figure 6*). Examples with missed diagnosis

Table 3 The classification performance of the DL model in the internal independent test set with different digital X-ray imaging devices

Type of equipment	Accuracy	Sensitivity	Specificity
Overall	0.964 (267/277)	0.976 (124/127)	0.953 (143/150)
Devices[a] (manufacturer information)			
Digital Diagnost, Philips Medical Systems	0.963 (156/162)	0.971 (66/68)	0.957 (90/94)
YSIO, Siemens Healthineers	1.000 (40/40)	1.000 (36/36)	1.000 (4/4)
Definium 6000, GE Healthcare	0.867 (26/30)	0.667 (2/3)	0.889 (24/27)
AeroDR C50, Konica Minolta Holdings Inc.	1.000 (21/21)	1.000 (1/1)	1.000 (20/20)
DX-D600, AFGA Medical Systems	1.000 (11/11)	1.000 (11/11)	–
Rad Speed Plus, Shimadzu Medical Systems	1.000 (7/7)	1.000 (7/7)	–
ESSENTA DR Compact, Philips Medical Systems	1.000 (3/3)	–	1.000 (3/3)
XGEO GC80, Samsung Electronics	1.000 (2/2)	–	1.000 (2/2)
Rad Speed M,Shimadzu Medical Systems	1.000 (1/1)	1.000 (1/1)	–

[a], there was no significant difference in the accuracy of the DL model in different digital X-ray imaging devices by using the Chi-squared test (P=0.183). DL, deep learning.

Table 4 Detection performance comparison of the DL model with the two junior radiologists in the internal independent test set

Characteristics	Readers	Accuracy	Sensitivity	Specificity	Kappa coefficient	P[a]
Overall	Model	0.964 (267/277)	0.976 (124/127)	0.953 (143/150)	0.927 (0.883–0.972)	
	Reader A	0.888 (246/277)	0.961 (122/127)	0.827 (124/150)	0.777 (0.704–0.850)	<0.001
	Reader B	0.921 (255/277)	0.945 (120/127)	0.900 (135/150)	0.841 (0.777–0.901)	0.033
Age[b] (years)						
<18	Model	0.988 (79/80)	1.000 (63/63)	0.941 (16/17)	0.962 (0.888–1.000)	
	Reader A	0.900 (72/80)	0.968 (61/63)	0.647 (11/17)	0.673 (0.465–0.881)	0.008
	Reader B	0.938 (75/80)	0.984 (62/63)	0.765 (13/17)	0.800 (0.633–0.968)	0.077
≥18	Model	0.954 (188/197)	0.953 (61/64)	0.955 (127/133)	0.897 (0.832–0.963)	
	Reader A	0.883 (174/197)	0.953 (61/64)	0.850 (113/133)	0.751 (0.657–0.845)	0.015
	Reader B	0.914 (180/197)	0.906 (58/64)	0.917 (122/133)	0.807 (0.720–0.894)	0.110
Tumor size[c] (mm)						
<80	Model	0.967 (58/60)	0.967 (58/60)	–	–	–
	Reader A	0.933 (56/60)	0.933 (56/60)	–	–	–
	Reader B	0.917 (55/60)	0.700 (7/10)	–	–	–
≥80	Model	0.985 (66/67)	0.985 (66/67)	–	–	–
	Reader A	0.983 (66/67)	0.983 (115/117)	–	–	–
	Reader B	0.966 (65/67)	0.966 (113/117)	–	–	–

The linearly weighted Cohen's kappa coefficient (K) measures the agreement between the DL model/readers and the reference standard in detecting a subject as a tumor patient or a normal participant. [a], the difference in Cohen's kappa coefficients between the DL model and the junior radiologists with 3–4 years of experience was statistically significant when P<0.05 by using permutation tests with 10,000 iterations; [b], there were no significant differences in the accuracy of the DL model (P=0.180), Reader A (P=0.550), and B (P=0.508) in different age groups based on 10 years old by using the Chi-squared test. There were no significant differences in the accuracy of the DL model (P=0.345), Reader A (P=0.705), and B (P=0.836) in different age groups based on 18 years old by using the Chi-squared test; [c], there were no significant differences in the accuracy of the DL model (P=0.497), Reader A (P=0.136), and B (P=0.189) in different tumor size groups by using the Chi-squared test. DL, deep learning; mm, millimeters.

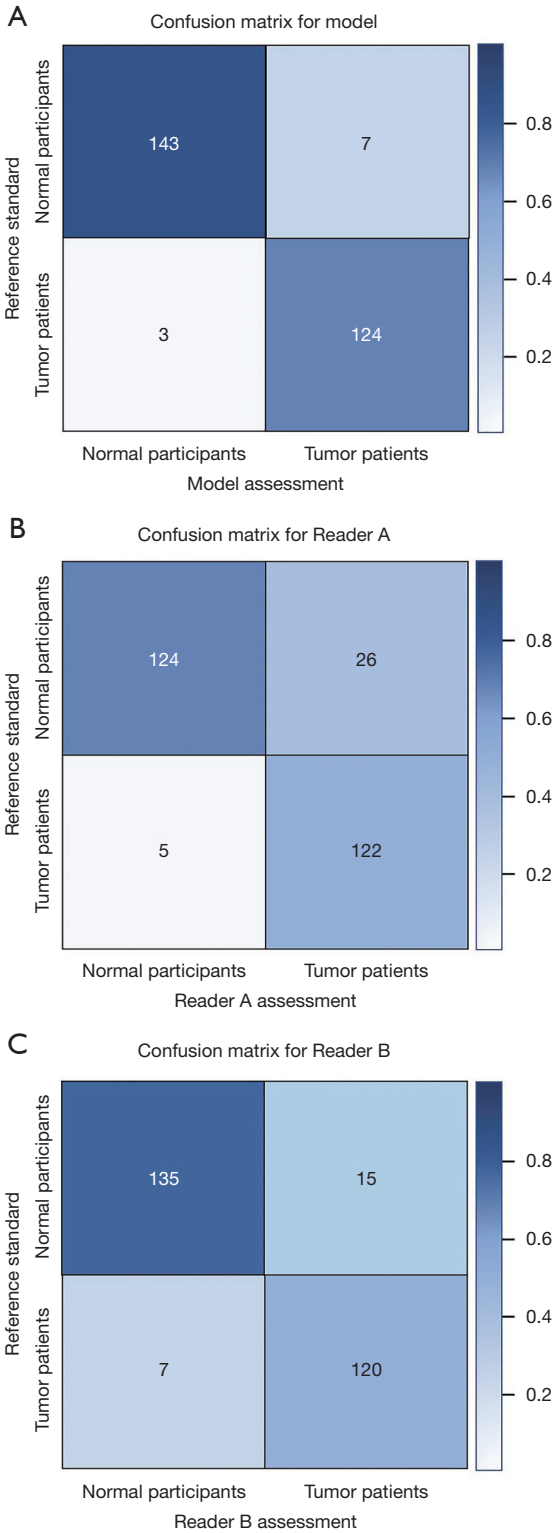

Figure 4 Confusion matrices for the DL model and observer performance assessments in the internal independent test set. DL, deep learning.

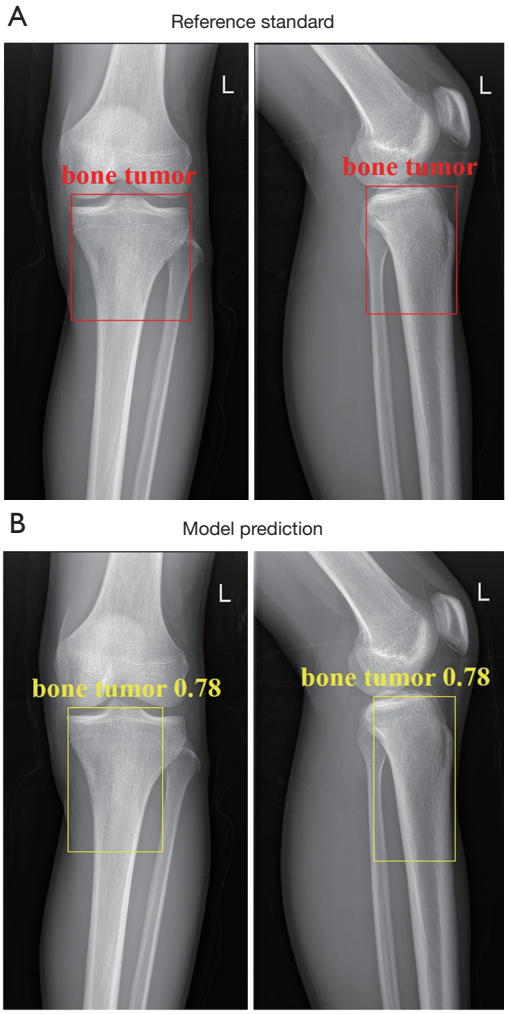

Figure 5 Examples of anteroposterior and lateral radiographs of tibial osteosarcoma. (A) A bone tumor in the diagnostic reference standard box on the radiographs. (B) A tumor detected in the predicted box by the DL model. The values printed on top of the predicted boxes are the confidence scores for each predicted box. This bone tumor was not found with the naked eye by the two junior radiologists. L, left; R, right; DL, deep learning.

and misinterpreted diagnosis by the DL model are shown in *Figures 7,8*. However, a misinterpreted diagnosis of patient by both the radiologist and the DL model was not seen in the independent test set.

Discussion

The detection of bone lesions from radiograph is a crucial initial stage before the classification of bone lesions

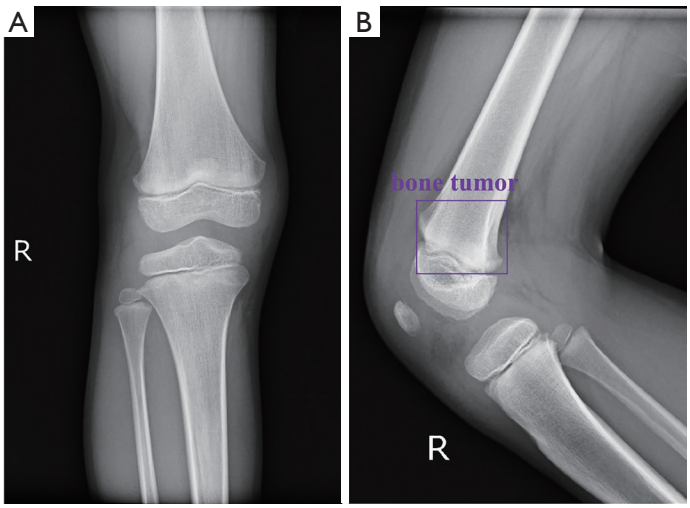

Figure 6 Anteroposterior (A) and lateral (B) radiographs of the right knee of a young normal participant. The area chosen with the bounding box on the lateral radiograph was misinterpreted as a tumor region by the junior radiologists and normal on the anteroposterior radiograph. Both the anteroposterior and lateral radiographs were diagnosed correctly by the DL model. R, right; DL, deep learning.

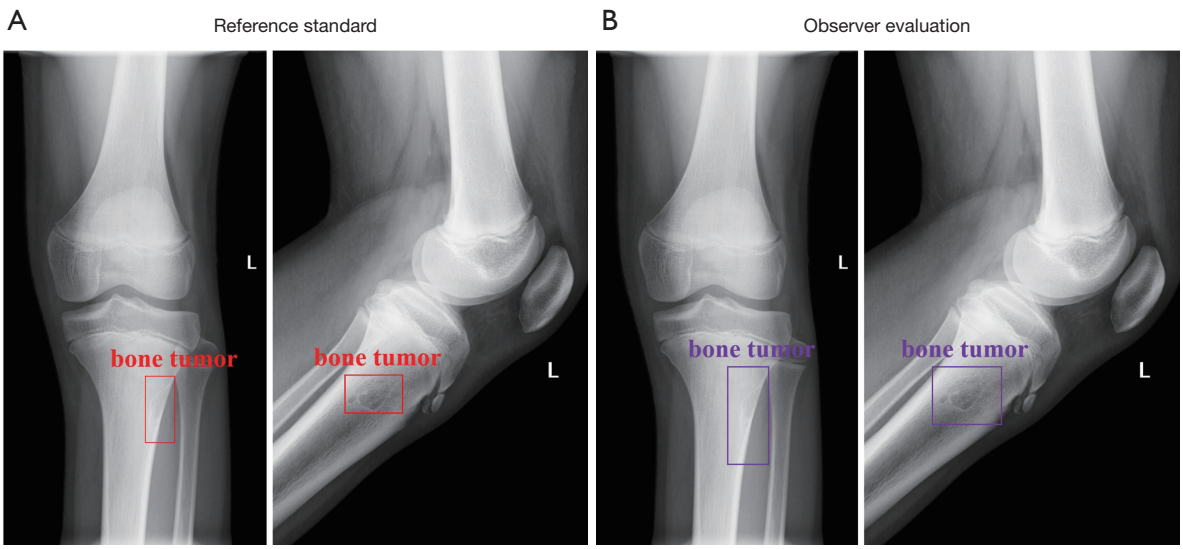

Figure 7 Benign tumors in the left tibia on radiographs were found by one of the two junior radiologists but missed by the other junior radiologist and were not detected by the DL model. (A) A bone tumor in the diagnostic reference standard box on the radiographs. (B) The tumor evaluated in the predicted box by an observer. L, left; DL, deep learning.

(18,30). Doi *et al.* (31) once built a DL-based NLP model for metastasis detection, but their research was based on radiology reports, not radiographs. In our study, we developed a DL model that located and detected bone tumor lesions on radiographs of pediatric and adult patients from four different centers across the country. The detection performance of the DL model (accuracies of 0.964 and 0.920 in the internal and external independent test set, respectively) was better than that of two junior radiologists (accuracies of 0.888 and 0.921 for Reader A and B, respectively) who had four years and three years of experience, respectively, in reading musculoskeletal

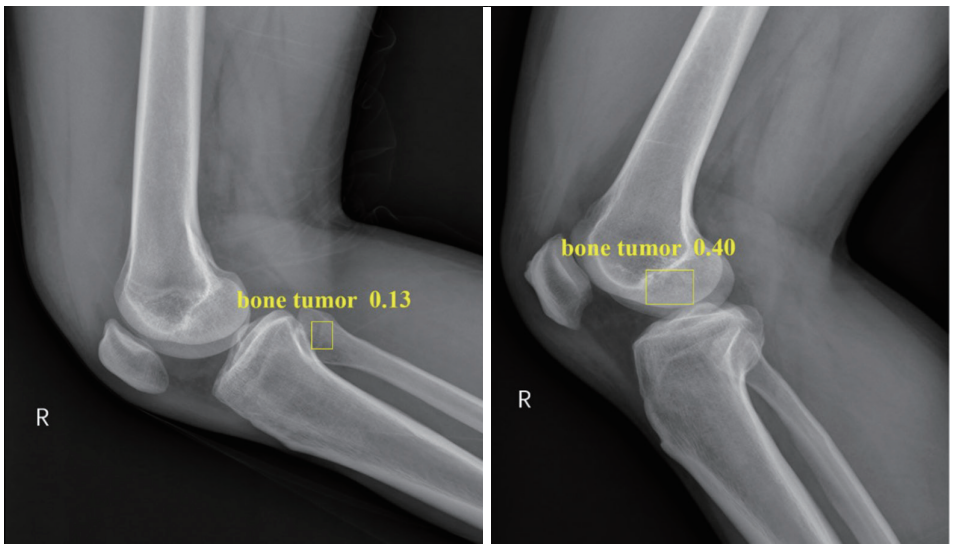

Figure 8 Two radiographs of normal knees diagnosed correctly by the two radiologists but misinterpreted as tumors by the DL model. The values on top of the predicted boxes are the confidence scores for the predicted box. R, right; DL, deep learning.

radiographs) in the internal test set. We compared the time taken by the DL model and radiologists to evaluate 554 radiographs from 277 subjects. Remarkably, our DL model took less time (2.7 min) than that of the junior radiologists (approximately 3 h), indicating that the DL model may help reduce the workload of radiologists in clinical practice. The DL model may help the lesser experienced radiologists and general practitioners in screening potential bone tumors on radiographs, who should be timely transferred for further classification diagnosis.

It is a challenge for junior radiologists to identify small or unobvious bone tumors from normal bones on radiographs due to the limited detection capacity of plain radiography (overlapping projection of bone and surrounding soft tissue structures and relatively low image contrast resolution compared to that of CT images) (9,10) and less clinical experience (tumoral bone density slightly higher or lower than that of the surrounding normal bone or unfamiliarity with density changes of the epiphyseal-metaphyseal junction during the development of long bone) in reading musculoskeletal radiographs. Our results indicated that the DL model might help radiologists reduce missed diagnoses (*Figure 5*) and misinterpreted diagnoses of primary bone tumors around the knee joint on radiographs (*Figure 6*).

Our model from the multicenter study demonstrated good bone tumor detection capacity in different subgroup analysis based on center, age (<18 and ≥18 years old) and tumor size (*Table 2*). A similar study by Breden *et al.* (20) used X-rays of 176 patients with bone tumor in a single center to develop a DL model for detection of bone tumors around the knee, achieving an accuracy of 89.1% in the internal test groups using classification algorithm, without an external testing or comparison of detection performance between the DL model and radiologist evaluation. Their study only focused on pediatric patients (≤18 years old), and identified suspicious X-ray images as a whole but not clearly detect where the tumor was. In contrast to the study by Breden *et al.*, our study covered a wide range of age stages and applied to a broad population (patients ranged in age from 3 to 70 years). On the other hand, our model has been externally validated with higher efficacy of detecting bone tumors. Li *et al.* (18) developed a YOLO model for detection of bone lesions on radiographs in several bones, achieving accuracies of 86.36% and 85.27% in the internal and external validation sets using IoU >0.5, respectively. In their study, the YOLO model detection capacity for bone tumor in different subgroup analysis based on age and tumor size has not been discussed. Our DL model correctly localized 94.5% and 92.9% bone tumors on radiographs in the internal and external test set using IoU >0.2, respectively. Most outpatients come to the hospital because of site-specific discomfort, and therefore imaging examination is usually limited to a specific site. Primary bone tumors of the extremities are commonly found in the bones around the knee joint. Our model for detecting bone tumors focuses on the knee region, which fits the clinical

scenario.

For the centers 2 and 3, the model's performance was lower than that of the center 1. The probable reasons may be as follows: in these two centers we collected much fewer tumor patients compared with those from center 1; and there was a large difference in the ratio of benign, intermediate and malignant bone tumors in these four centers (*Table 1*). The good model performance in external testing groups supports generalizability of our algorithm.

The DL model made some misdiagnoses on metaphysis epiphysis (*Figure 8*) probably because of the density change of the epiphyseal-metaphyseal junction during epiphyseal closure in the development of long bone. If the training dataset does not include density changes in the whole process of epiphyseal closure, it may be misinterpreted as a localized lesion. On the other hand, several relatively small or inconspicuous bone changes in bone tumors were not detected by the DL model despite the overall high sensitivity of the DL model (*Figure 7*). These false-negative detections were probably because of the limited sample size of such tumors. This illustrated a pitfall of existing DL methods; namely, relatively small or inconspicuous abnormalities may be missed because of insufficient data on such abnormalities in the training set used for developing the model (32).

Next, we analyzed several limitations of this study, some of which can be further investigated in future work. First, primary bone tumors of the knee were analyzed without considering other bone tumors elsewhere. Although all malignant bone tumors and most benign bone tumors were postoperatively proven by histopathologic findings, some benign bone tumors, such as enchondroma and fibrous dysplasia, were diagnosed by using imaging examination and clinical follow-up. All patients with malignant bone tumors were examined by radiography and MRI in the study. Second, we mainly collected some common types of primary bone tumors around the knee joint, such as osteosarcoma and giant cell tumors of bone. This was because these bone tumors are much more prevalent (1). Osteochondroma, which represented a larger proportion of bone tumors in previous studies, was not included in the current study (11,18,20) because it can be easily found on radiographs with the naked eye due to its characteristic extraosseous growth shape. Third, considering the fact that there are more trauma patients without fractures across the country, normal bone radiographs as a normal control group in this study were chosen from some patients who presented with trauma. Fourth, in this study, we aimed to develop a DL model for the detection of primary bone tumors on knee radiographs and compared its performance with that of junior radiologists. We only targeted detecting bone tumor from normal bone on radiographs, without involving benign and malignant classification of bone tumors or differentiation of other bone diseases. Raman spectroscopy provides spectra which are great indicators for the analysis and monitoring of several diseases including bone tumors (33,34). The next study on the classification of benign and malignant bone tumors, prognosis, etc. can be based on some new techniques such as Raman spectroscopy. Finally, our study only focused on bone tumors occurring in the knee region. Thus, our DL model may have a relatively narrow application in a real-world scenario. To better extend the application of the DL model, we will continue to optimize the model in future studies by collecting data from patients with other bone diseases and other sites.

Conclusions

In conclusion, we developed a DL model for the accurate detection of bone tumors on knee radiographs and our model showed better performance than that of junior radiologists. Our DL model may help the lesser experienced radiologists or general practitioners in detecting potential bone tumors on radiographs and has the potential to reduce missed diagnoses and misinterpreted diagnoses with less time taken. In future research, we will select other state-of-the-art versions (e.g., YOLOv7 or v8) and some new techniques such as Raman spectroscopy to develop model on the classification of benign and malignant bone tumors, prognosis, etc.

Acknowledgments

We thank two junior radiologists (Dr. Lixian Chen and Dr. Bingkun Guo, with 3 and 4 years of experience respectively) for their help in evaluating 554 radiographs in 227 subjects. Also, we thank Dr. Xinqu Huang and Dr. Xiaofeng Chen for their help in data collection.

Funding: This work was funded by the Natural Science Foundation of Guangdong Province, China (No. 2022A1515011593) and Medical Research Foundation of Guangdong Province, China (No. A2021010).

Footnote

Reporting Checklist: The authors have completed the

TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-1743/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1743/coif). All authors report the funding from the Natural Science Foundation of Guangdong Province, China (No. 2022A1515011593) and the Medical Research Foundation of Guangdong Province, China (No. A2021010). The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Ethics Committee of The First Affiliated Hospital, Sun Yat-sen University (No. [2022]541). Written informed consent was waived by the Ethics Committee due to the retrospective nature of the study. All participating hospitals/institutions were informed and agreed with the study.

References

1. Choi JH, Ro JY. The 2020 WHO Classification of Tumors of Bone: An Updated Review. Adv Anat Pathol 2021;28:119-38.
2. Arora RS, Alston RD, Eden TO, Geraci M, Birch JM. The contrasting age-incidence patterns of bone tumours in teenagers and young adults: Implications for aetiology. Int J Cancer 2012;131:1678-85.
3. Beird HC, Bielack SS, Flanagan AM, Gill J, Heymann D, Janeway KA, Livingston JA, Roberts RD, Strauss SJ, Gorlick R. Osteosarcoma. Nat Rev Dis Primers 2022;8:77.
4. Cole S, Gianferante DM, Zhu B, Mirabello L. Osteosarcoma: A Surveillance, Epidemiology, and End Results program-based analysis from 1975 to 2017. Cancer 2022;128:2107-18.
5. He Y, Wang J, Zhang J, Yuan F, Ding X. A prospective study on predicting local recurrence of giant cell tumour of bone by evaluating preoperative imaging features of the tumour around the knee joint. Radiol Med 2017;122:546-55.
6. Zhou L, Zhu H, Lin S, Jin H, Zhang Z, Dong Y, Yang Q, Zhang C, Yuan T. Computerised tomography features of giant cell tumour of the knee are associated with local recurrence after extended curettage. Int Orthop 2022;46:381-90.
7. Gemescu IN, Thierfelder KM, Rehnitz C, Weber MA. Imaging Features of Bone Tumors: Conventional Radiographs and MR Imaging Correlation. Magn Reson Imaging Clin N Am 2019;27:753-67.
8. Expert Panel on Musculoskeletal Imaging; Bestic JM, Wessell DE, Beaman FD, Cassidy RC, Czuczman GJ, Demertzis JL, Lenchik L, Motamedi K, Pierce JL, Sharma A, Sloan AE, Than K, Walker EA, Ying-Kou Yung E, Kransdorf MJ. ACR Appropriateness Criteria® Primary Bone Tumors. J Am Coll Radiol 2020;17:S226-38.
9. Gould CF, Ly JQ, Lattin GE Jr, Beall DP, Sutcliffe JB 3rd. Bone tumor mimics: avoiding misdiagnosis. Curr Probl Diagn Radiol 2007;36:124-41.
10. Krych A, Odland A, Rose P, Dahm D, Levy B, Wenger D, Stuart M, Sim F. Oncologic conditions that simulate common sports injuries. J Am Acad Orthop Surg 2014;22:223-34.
11. Liu R, Pan D, Xu Y, Zeng H, He Z, Lin J, Zeng W, Wu Z, Luo Z, Qin G, Chen W. A deep learning-machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. Eur Radiol 2022;32:1371-83.
12. Bandyopadhyay O, Biswas A, Bhattacharya BB. Bone-Cancer Assessment and Destruction Pattern Analysis in Long-Bone X-ray Image. J Digit Imaging 2019;32:300-13.
13. He Y, Pan I, Bao B, Halsey K, Chang M, Liu H, Peng S, Sebro RA, Guan J, Yi T, Delworth AT, Eweje F, States LJ, Zhang PJ, Zhang Z, Wu J, Peng X, Bai HX. Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. EBioMedicine 2020;62:103121.
14. Ho NH, Yang HJ, Kim SH, Jung ST, Joo SD. Regenerative Semi-Supervised Bidirectional W-Network-Based Knee Bone Tumor Classification on Radiographs Guided by Three-Region Bone Segmentation. IEEE Access 2019;7:154277-89.
15. Park CW, Oh SJ, Kim KS, Jang MC, Kim IS, Lee YK, Chung MJ, Cho BH, Seo SW. Artificial intelligence-based

classification of bone tumors in the proximal femur on plain radiographs: System development and validation. PLoS One 2022;17:e0264140.
16. von Schacky CE, Wilhelm NJ, Schäfer VS, Leonhardt Y, Jung M, Jungmann PM, Russe MF, Foreman SC, Gassert FG, Gassert FT, Schwaiger BJ, Mogler C, Knebel C, von Eisenhart-Rothe R, Makowski MR, Woertler K, Burgkart R, Gersing AS. Development and evaluation of machine learning models based on X-ray radiomics for the classification and differentiation of malignant and benign bone tumors. Eur Radiol 2022;32:6247-57.
17. von Schacky CE, Wilhelm NJ, Schäfer VS, Leonhardt Y, Gassert FG, Foreman SC, Gassert FT, Jung M, Jungmann PM, Russe MF, Mogler C, Knebel C, von Eisenhart-Rothe R, Makowski MR, Woertler K, Burgkart R, Gersing AS. Multitask Deep Learning for Segmentation and Classification of Primary Bone Tumors on Radiographs. Radiology 2021;301:398-406.
18. Li J, Li S, Li X, Miao S, Dong C, Gao C, Liu X, Hao D, Xu W, Huang M, Cui J. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. Eur Radiol 2023;33:4237-48.
19. Pan C, Lian L, Chen J, Huang R. FemurTumorNet: Bone tumor classification in the proximal femur using DenseNet model based on radiographs. J Bone Oncol 2023;42:100504.
20. Breden S, Hinterwimmer F, Consalvo S, Neumann J, Knebel C, von Eisenhart-Rothe R, Burgkart RH, Lenze U. Deep Learning-Based Detection of Bone Tumors around the Knee in X-rays of Children. J Clin Med 2023;12:5960.
21. Hinterwimmer F, Serena RS, Wilhelm N, Breden S, Consalvo S, Seidl F, Juestel D, Burgkart RHH, Woertler K, von Eisenhart-Rothe R, Neumann J, Rueckert D. Recommender-based bone tumour classification with radiographs-a link to the past. Eur Radiol 2024. [Epub ahead of print]. doi: 10.1007/s00330-024-10672-0.
22. Salehi MA, Mohammadi S, Harandi H, Zakavi SS, Jahanshahi A, Shahrabi Farahani M, Wu JS. Diagnostic Performance of Artificial Intelligence in Detection of Primary Malignant Bone Tumors: a Meta-Analysis. J Imaging Inform Med 2024;37:766-77.
23. Bahamonde L, Catalan J. Bone tumors around the knee: risks and benefits of arthroscopic procedures. Arthroscopy 2006;22:558-64.
24. WHO Classification of Tumours Editorial Board. Soft tissue and bone tumours. WHO Classification of Tumours, 5th Edition. IARC Press, 2020.
25. Jocher G, Stoken A, Borovec J, Chaurasia A, Xie T, Changyu L, et al. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. 2021. doi: 10.5281/ZENODO.4679653.
26. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL, editors. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T. editors. Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, 2014:740-55.
27. Kimura M. Understanding Test-Time Augmentation. Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I, 2021:558-69.
28. Zhang L, Li Y, Chen H, Wu W, Chen K, Wang S. Anchor-free YOLOv3 for mass detection in mammogram. Expert Systems with Applications 2022. doi: 10.1016/j.eswa.2021.116273.
29. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. Radiol Artif Intell 2019;1:e180001.
30. Fritz B, Fritz J. Artificial intelligence for MRI diagnosis of joints: a scoping review of the current state-of-the-art of deep learning-based approaches. Skeletal Radiol 2022;51:315-29.
31. Doi K, Takegawa H, Yui M, Anetai Y, Koike Y, Nakamura S, Tanigawa N, Koziumi M, Nishio T. Deep learning-based detection of patients with bone metastasis from Japanese radiology reports. Jpn J Radiol 2023;41:900-8.
32. Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 2006;28:594-611.
33. Manganelli Conforti P, D'Acunto M, Russo P. Deep Learning for Chondrogenic Tumor Classification through Wavelet Transform of Raman Spectra. Sensors (Basel) 2022;22:7492.
34. Lau CPY, Ma W, Law KY, Lacambra MD, Wong KC, Lee CW, Lee OK, Dou Q, Kumta SM. Development of deep learning algorithms to discriminate giant cell tumors of bone from adjacent normal tissues by confocal Raman spectroscopy. Analyst 2022;147:1425-39.

Cite this article as: Xu D, Li B, Liu W, Wei D, Long X, Huang T, Lin H, Cao K, Zhong S, Shao J, Huang B, Diao XF, Gao Z. Deep learning-based detection of primary bone tumors around the knee joint on radiographs: a multicenter study. Quant Imaging Med Surg 2024;14(8):5420-5433. doi: 10.21037/qims-23-1743

