npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-00855-3

Deep learning models in classifying primary bone tumors and bone infections based on radiographs

Hua Wang[1,2,7], Yu He[3,7], Lu Wan[1,2], Chenbei Li[1,2], Zhaoqi Li[1,2], Zhihong Li[1,2,4], Haodong Xu[1,2,5] ✉ & Chao Tu[1,2,4,6] ✉

Primary bone tumors (PBTs) present significant diagnostic challenges due to their heterogeneous nature and similarities with bone infections. This study aimed to develop an ensemble deep learning framework that integrates multicenter radiographs and extensive clinical features to accurately differentiate between PBTs and bone infections. We compared the performance of the ensemble model with four imaging models based solely on radiographs utilizing EfficientNet B3, EfficientNet B4, Vision Transformer, and Swin Transformers. The patients were split into external dataset ($N = 423$) and internal dataset [including training ($N = 1044$), test ($N = 354$), and validation set ($N = 171$)]. The ensemble model outperformed imaging models, achieving areas under the curve (AUCs) of 0.948 and 0.963 on internal and external sets, respectively, with accuracies of 0.881 and 0.895. Its performance surpassed junior and mid-level radiologists and was comparable to senior radiologists (accuracy: 83.6%). These findings underscore the potential of deep learning in enhancing diagnostic precision for PBTs and bone infections (Research Registration Unique Identifying Number (UIN): researchregistry10483 and with details are available at https://www.researchregistry.com/register-now#home/registrationdetails/6693845995ba110026aeb754/).

Primary bone tumors (PBTs) are a diverse group of heterogeneous tumors that primarily develop in the skeletal system[1]. Despite their relatively low incidence, these malignancies present significant morbidity and mortality rates[2,3]. Remarkably, bone tumors rank as the third leading cause of cancer-related deaths among individuals under the age of 20 in the United States[4]. Currently, the treatment options for bone tumors remain formidable, traditional treatment options such as chemotherapy and surgical interventions, face significant challenges[1,5]. For instance, chemotherapy often leads to severe side effects and has a limited success rate due to chemoresistance in specific type of bone tumors like osteosarcoma[6,7], while surgical options may result in functional impairments, residual metastasis, and even deformities or disabilities[8–10]. These challenges underscore the need for improved treatment strategies. Radiography is the suggested primary auxiliary examination choice and commonly employed in orthopedic diagnosis as they generally provide a clear evaluation of the lesion's location, internal matrix, margins, and associated periosteal reactions[11]. These destruction signs reflect the biological activity of the lesion, thus allowing for evaluation of the malignancy assessment[12]. However, PBTs exhibit diverse compositions and may present with overlapping radiological and histological features[13,14]. Consequently, the same PBTs may appear differently on radiographs, and different PBTs may exhibit similar radiographic images[15]. Due to the rarity of PBTs, cultivating a professional radiologist often encounters the problem of a long training cycle and insufficient expertise[16]. Bone infections primarily encompass osteomyelitis and joint infections. Notably, clinically distinguishing PBTs from bone infections is challenging for the similarities in clinical practice (e.g., fever, soft tissue swelling, periosteal reaction), leading to potential confusion and challenges in accurate diagnosis[17,18]. Therefore, the pre-operative differential diagnosis of PBTs and bone infections is crucial for precise diagnosis and timely treatment.

[1]Department of Orthopaedics, The Second Xiangya Hospital of Central South University, Changsha, Hunan, China. [2]Hunan Key Laboratory of Tumor Models and Individualized Medicine, The Second Xiangya Hospital of Central South University, Changsha, Hunan, China. [3]Department of Radiology, The Second Xiangya Hospital of Central South University, Changsha, Hunan, China. [4]Shenzhen Research Institute of Central South University, Guangdong, China. [5]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA. [6]Changsha Medical University, Changsha, Hunan, China. [7]These authors contributed equally: Hua Wang, Yu He. ✉e-mail: xuhaodong@csu.edu.cn; tuchao@csu.edu.cn

Traditional diagnostic methods heavily rely on the expertise and subjective judgment of radiologists and pathologists, which can lead to potential errors and delays in treatment options[19–21]. Furthermore, if imaging studies are not interpreted by musculoskeletal radiologists who specialize in this field, discrepancies in readings can occur, reaching up to 28%[22]. In recent years, the emergence of deep learning algorithms especially convolutional neural networks (CNNs) has significantly impacted clinical practices such as assisted diagnosis and drug discovery[23,24]. These advancements have also demonstrated improvements in cancer prognosis[25]. The application of deep learning in cancer diagnosis has considerably enriched the field, showcasing astounding efficiency in solving complex problems with a lower error rate than humans[26,27]. For bone tumors, the development of multitask deep learning models has enabled accurate and simultaneous bounding box placement and segmentation of PBTs in radiographs, and can effectively differentiate benign and malignant PBTs with performance comparable to senior radiologists[28]. Due to the rarity of PBTs, deep learning models in this domain are constrained by limited access to large-scale cohort datasets, resulting in scant efforts to differentiate between bone tumors and other bone pathologies. Furthermore, prevailing models emphasize algorithmic versatility and data diversity, yet they fall short in sufficiently incorporating crucial clinical patient data and prioritizing the interpretability of model outcomes. This trend runs counter to the fundamental ethos of algorithmic design, sometimes it is necessary to pause and delve into a profound comprehension of our meticulously crafted models with professional radiologist interpretation, thereby aligning our efforts with the original essence of algorithmic innovation.

Therefore, the main objective of this study was to create an ensemble deep learning framework using multicenter radiographs and extensive clinical features to accurately differentiate between PBTs and bone infections. While comparing the performance of the ensemble model with four imaging models merely utilizing radiographs, which were built upon four distinct neural networks: EfficientNet B3 (E3), EfficientNet B4 (E4), Vision Transformer (ViT), and Swin Transformers (SWIN). Subsequently, these models' effectiveness was assessed and compared with the diagnostic accuracy of radiologists. In addition, six professional radiologists, categorized into three seniority groups, provided insights and discussions on the clinical implications of the developed models. The research methodology and study flowchart are illustrated in Fig. 1.

Results

Characteristics of study participants

This retrospective study included 1992 patients (median age, 29 years; range, 1–88 years; 796 female) from three hospitals diagnosed of PBTs or bone infections with histopathology reports available as reference (Table 1). The distribution of 1208 patients with PBTs were described in Supplementary Table 1, with 767 benign subtypes, 251 malignant subtypes and 190 intermediate subtypes according to the 2020 World Health Organization (WHO) system for the classification for tumors of bone. While for 784 patients with bone infection, bone tuberculosis counted the highest proportion (Supplementary Table 2). 1569 patients from Hospital 1 were utilized as internal dataset and divided into a training set ($N = 1044$), a test set ($N = 354$) and a validation set ($N = 171$) (Fig. 2a) (screening criteria in Fig. 2b); 423 patients from Hospital 2 and Hospital 3 were used for external validation (Supplementary Fig. 1). Clinical characteristics like age, lesion location, pain, swelling, trauma, C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), alkaline phosphatase (ALP) among all of the bone infection and PBT patients had significantly different distributions (Table 1). The clinical characteristics of patients with PBTs and bone infection were summarized specifically in Supplementary Tables 3 and 4. We further found that clinical characteristics like age, lesion location, pain, swelling, trauma, C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), alkaline phosphatase (ALP) also had statistical differences in the internal dataset (Supplementary Table 5).

Classification performance of models

In the internal test set, the ensemble model outperformed four imaging models (E3, E4, ViT and SWIN) on the binary classification to distinguish PBTs from bone infections respectively ($P < 0.001$ for E3, E4, and ViT; $P = 0.835$ for SWIN; DeLong test) (Table 2 and Supplementary Fig. 2). Specifically, the ensemble model reached an AUC of 0.948 (95% CI, 0.931–0.963) and an accuracy of 88.1% for binary classification, whereas the E3, E4, ViT and SWIN-based models achieved AUCs of 0.903 (95% CI, 0.878–0.927), 0.912 (95% CI, 0.890–0.934), 0.903 (95% CI, 0.880–0.927), and 0.946 (95% CI, 0.929–0.963) as well as accuracies of 84.3%, 84.6%, 84.3%, and 87.2%, respectively (Table 2). The ROC curves and the confusion matrices also demonstrated the best categorizing ability of the ensemble model (Fig. 3 and Supplementary Fig. 3).

In the external test set for validation, the ensemble model also outperformed the four imaging models, which proved the consistency and applicability of the ensemble model ($P < 0.001$ for E3 and E4; $P = 0.002$ for ViT and SWIN; DeLong test) (Table 2 and Supplementary Fig. 2). Specifically, the ensemble model reached an AUC of 0.963 (95% CI, 0.951–0.973) and an accuracy of 89.5% for the classification, while the four imaging models reached AUCs of 0.930 (95% CI, 0.914–0.946), 0.946 (95% CI, 0.932–0.960), 0.951 (95% CI, 0.939–0.964), and 0.957 (95% CI, 0.944–0.969) as well as accuracies of 86.6%, 87.4%, 87.1%, and 88.5%, respectively (Table 2). The confusion matrices and ROC curves in Fig. 3 further visually demonstrated the superior discrimination capability of the ensemble framework. In addition, the result in internal validation set further confirmed the stability and consistency of the ensemble model (Supplementary Fig. 4).

Comparison of performance between the ensemble framework and radiologists

In this study, six professional radiologists were divided into junior expert group (EG1), medium seniority group (EG2), and senior expert group (EG3). The comparative analysis was conducted using the internal test set. As shown in Fig. 3, the ensemble framework significantly outperformed all three radiologist groups ($P < 0.001$ for EG1, EG2, and EG3; Cochran's Q test) (Table 2). The SWIN-based imaging model demonstrated comparable performance to the ensemble model ($P = 0.835$; DeLong test) (Table 2) and also outperformed the three radiologist groups. The other three imaging models (E3, E4, and ViT) achieved superior performance compared to EG1 and EG2, and were comparable to EG3. In addition, we calculated and provided other metrics, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 Score, to facilitate a comprehensive comparison of the performance between the ensemble framework and the radiologists (Table 3).

Inter-reader reliability

Considering the subjectivity of individual sample predictions and large workload of the monotonous radiographs ($n = 687$), inter-reader reliability among radiologists was much lower than that of the models. We compared the best performing model—the ensemble model with experts of diverse seniority, Cohen κ between expert 6 (radiologist with the highest seniority) and the ensemble had the best consistency: 0.596 (95% CI, 0.560–0.633) (Table 4). The Fleiss κ value among radiologists achieved 0.401 (95% CI, 0.364–0.438) on the internal test set, while the Fleiss κ value among models achieved 0.800 (95% CI, 0.770–0.830) (Table 4). Furthermore, we used Cohen κ value to evaluate consistency between pairs of expert groups (EG1, EG2, and EG3) and consistency between the ensemble model and the other four imaging models. We found as seniority increased, the consistency of judgment rose in radiologists, but the overall consistency of judgment was still lower than that of the models. The Fleiss κ value among EG1, EG2, and EG3 reached 0.267 (95% CI, 0.234–0.300), 0.295 (95% CI, 0.261–0.329), and 0.581 (95% CI, 0.544–0.618), respectively (Table 4). In contrast, the Fleiss κ value among the ensemble model and the imaging models reached 0.805 (95% CI, 0.775–0.835), 0.793 (95% CI, 0.763–0.823), 0.783 (95% CI, 0.752–0.814), and 0.908 (95% CI, 0.886–0.930), respectively (Table 4). This

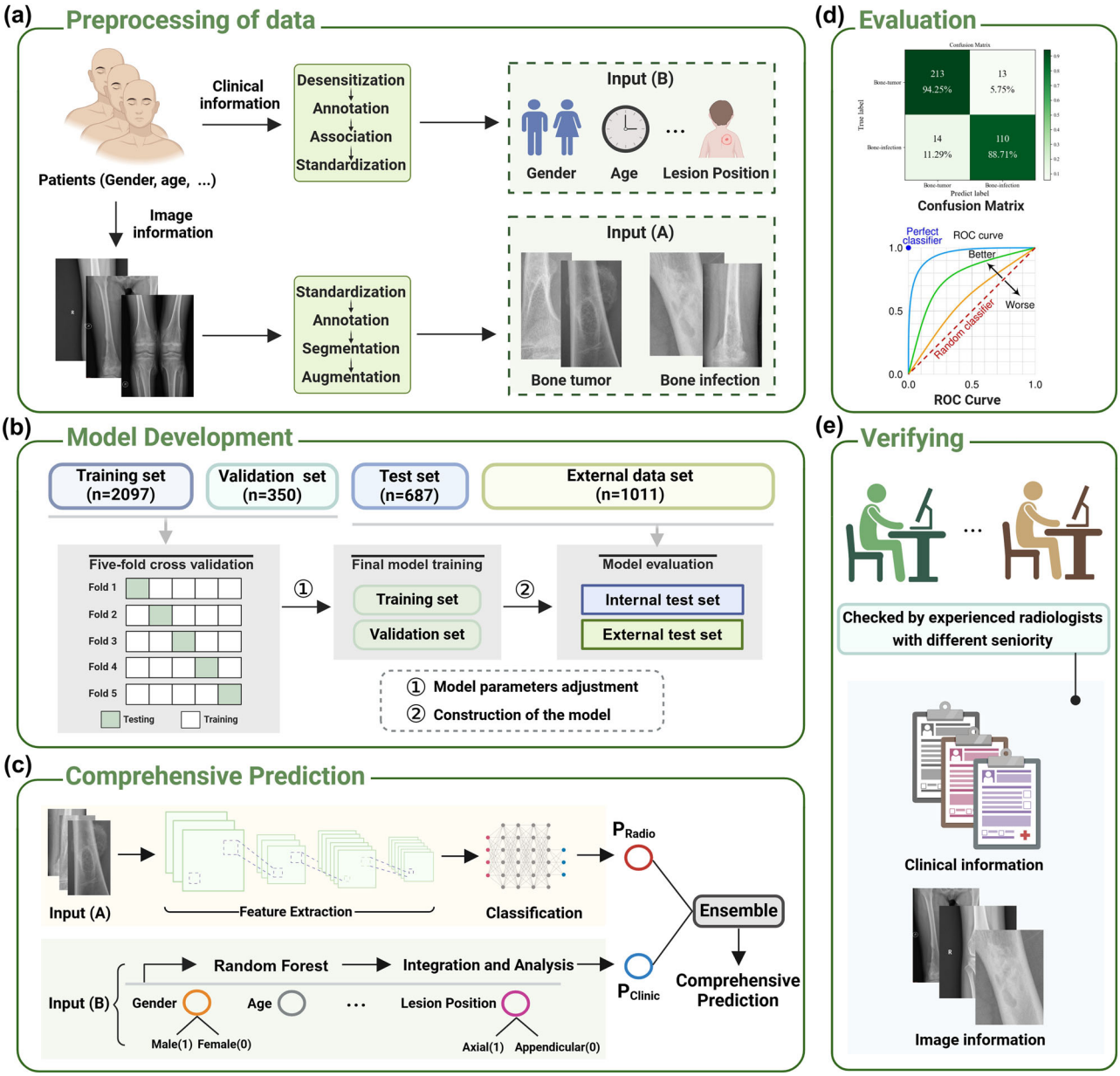

Fig. 1 | Design and flowchart of the deep learning framework. a Preprocessing of data. The input of the models mainly includes image information based on radiographs defined as input (A) and clinical information defined as input (B). **b** Model development. **c** Comprehensive prediction. P_{Radio} and P_{Clinic} refers to the results of the four imaging models (E3, E4, ViT, and SWIN) and the clinic model, respectively. **d** Evaluation. This part is mainly composed of ROC curve and confusion matrix. **e** Verifying. The results of models are compared with radiologists with different seniority. *n* number of the radiographs, E3 EfficientNet B3, E4 EfficientNet B4, ViT vision transformer, SWIN swin transformers. Note: Fig. 1 was Created with BioRender.com.

indicates that a strong disagreement exists among junior radiologists when facing classification of PBTs and bone infection solely on radiograph data.

Visual interpretation of models

In order to accurately interpret the predictions made by the models, we employed techniques such as GradCAM and ScoreCAM to visualize the specific regions within the input data that the model utilizes for its decision-making process (Fig. 4). By identifying and highlighting these key areas, we are able to gain a deeper understanding of how the model arrives at its predictions and make informed assessments about its performance and reliability. In general, the analysis of the highlighted regions on the heat maps reveals that the model primarily focused on identifying PBT or bone infection lesions, such as hemorrhage, necrosis, calcification, cystic lesions, and inflammatory exudation. These findings are in line with the segmentation results, indicating that the model was able to achieve a high level of accuracy in classifying these specific types of lesions. This demonstrates the effectiveness of the model in accurately identifying and categorizing pathological features, ultimately leading to satisfactory classification performance. The distinctions between GradCAM and ScoreCAM are clearly evident in the generated heat maps. GradCAM primarily emphasizes the areas of bone hyperplasia and sclerosis, neglecting those of bone destruction. Conversely, ScoreCAM directs its attention toward both osteogenic and osteoclastogenic regions, resulting in a more precise delineation of lesion boundaries.

Radiologist interpretation

Diagnosis of the ensemble model and radiologists across different types of PBTs and bone infections were explicated in Supplementary Tables 6 and 7, specifically. Some bone tumors were classified incorrectly by experts but correctly by the model (Fig. 5). Giant cell tumors of bone (Fig. 5a) may exhibit obvious aggressiveness, resulting in the blurring of the boundary

Table 1 | Clinical characteristics of included patients with primary bone tumors or bone infections

Characteristics	Patients with PBTs (*N* = 1208)	Patients with bone infection (*N* = 784)	All patients (*N* = 1992)	*P* value
Age (year)	24.85 ± 18.18	45.65 ± 18.95	33.04 ± 21.09	<0.001*
Gender				0.1663
Female	498 (41.23%)	298 (38.01%)	796 (39.96%)	
Male	710 (58.77%)	486 (61.99%)	1196 (60.04%)	
Position				<0.001*
Appendicular	1044 (86.42%)	314 (40.05%)	1358 (68.17%)	
Axial	164 (13.58%)	470 (59.95%)	634 (31.83%)	
Pain				<0.001*
Yes	857 (70.94%)	487 (62.12%)	1344 (67.47%)	
No	351 (29.06%)	51 (6.51%)	402 (20.18%)	
NA	0	246 (31.37%)	246 (18.30%)	
Swelling				<0.001*
Yes	538 (44.54%)	320 (40.82%)	858 (43.07%)	
No	662 (54.80%)	218 (27.81%)	880 (44.18%)	
NA	8 (0.66%)	246 (31.37%)	254 (12.75%)	
Trauma				<0.001*
Yes	173 (14.32%)	47 (5.99%)	220 (11.04%)	
No	858 (71.03%)	491 (62.64%)	1349 (67.72%)	
NA	177 (14.65%)	246 (31.37%)	423 (21.24%)	
CRP				<0.001*
Normal	655 (54.22%)	194 (24.74%)	849 (50.10%)	
Abnormal	440 (36.42%)	389 (49.62%)	829 (34.64%)	
NA	113 (9.36%)	201 (25.64%)	314 (15.76%)	
ESR				0.0089*
Normal	286 (23.68%)	53 (6.76%)	339 (26.27%)	
Abnormal	803 (66.47%)	536 (68.37%)	1339 (22.68%)	
NA	119 (9.85%)	195 (24.87%)	314 (51.05%)	
ALP				0.0086*
Normal	0 (0.00%)	0 (0.00%)	0 (0.00%)	
Abnormal	468 (38.74%)	391 (49.87%)	859 (43.12%)	
NA	740 (61.26%)	393 (50.13%)	1133 (56.88%)	

PBTs primary bone tumors, *N* number of patients, *NA* not appliable, *CRP* C-reactive protein, *ESR* erythrocyte sedimentation rate, *ALP* alkaline phosphatase.
Data in parentheses are percentages. Continuous variables are expressed as mean ± standard deviation. **P* values less than 0.05 are considered statistically significant.

between the lesion and normal bone, wormlike and ethmoidal bone destruction, and soft tissue masses beyond the bone envelope. There is partial image overlap with malignant bone tumors and infections (such as Brodie abscess) on plain film[29]. Synovial osteo-chondromatosis (Fig. 5b) is characterized by multiple cartilage nodules in the joint lumen. When the cartilage nodules are not significantly calcified, especially when bone erosion is present at the same time, it is difficult to distinguish osteoarthritis with free bodies in the joint[30]. There are also cases where both experts and models misclassify. Chondrosarcoma (Fig. 5e) involving the pelvis is more likely to occur in the iliac wing than in the acetabulum. Intramedullary osteolytic lesions with poorly defined acetabular boundaries may be consistent with chondrosarcoma, as well as tuberculosis and osteoarthritis of the hip. The overlap of the structure in plain film makes the calcification of the circular or arc-shaped chondroid stroma, a typical manifestation of chondrosarcoma at the acetabulum, not obvious, and appears to be suspected involvement of the adjacent femoral head. Multiple myeloma (Fig. 5f) tends to occur in the thoracic vertebrae and has a positive pedicle sign (destruction of the vertebral body but retention of the pedicle). When both the vertebral body and pedicle are destroyed at the same time, it is necessary to distinguish them from spinal metastasis and spinal tuberculosis with insignificant paravertebral abscess[31]. There are also cases where the experts got the classification right and the model got it wrong. Sclerosing osteosarcoma has no obvious bone destruction, which is different from the common mixed osteosarcoma with both osteolytic and sclerosing (Fig. 5c). Giant cell tumors of bone occur mostly in the long bone, but can also occur in the vertebral body (Fig. 5d). These relatively uncommon conditions can be recognized by radiologists with extensive clinical experience. However, due to limited training on rare cases, the model tends to focus more on interpreting the more frequently encountered chronic osteomyelitis and spinal tuberculosis.

Some bone infections were classified incorrectly by experts but correctly by the model (Fig. 6). Chronic sclerosing osteomyelitis (Garre osteomyelitis, Fig. 6a) mainly presents with osteosclerosis and lack of dead bone formation, and needs to be distinguished from sclerosing osteosarcoma[32]. When lumbar tuberculosis (Fig. 6b) involves only a single vertebral body and lacks paravertebral space narrowing, formation of paravertebral cold abscess, and soft tissue calcification, it should be differentiated from plasma-cell tumor and giant cell tumor of bone. There are also cases where the experts got the classification right, and the model got it wrong. There is partial overlap between acute suppurative osteomyelitis (Fig. 6d) and Ewing sarcoma. Although the image manifestations of joint tuberculosis (Fig. 6c) occurring in the elbow joint are relatively typical, the number of training cases of joint tuberculosis in the extremities is limited for

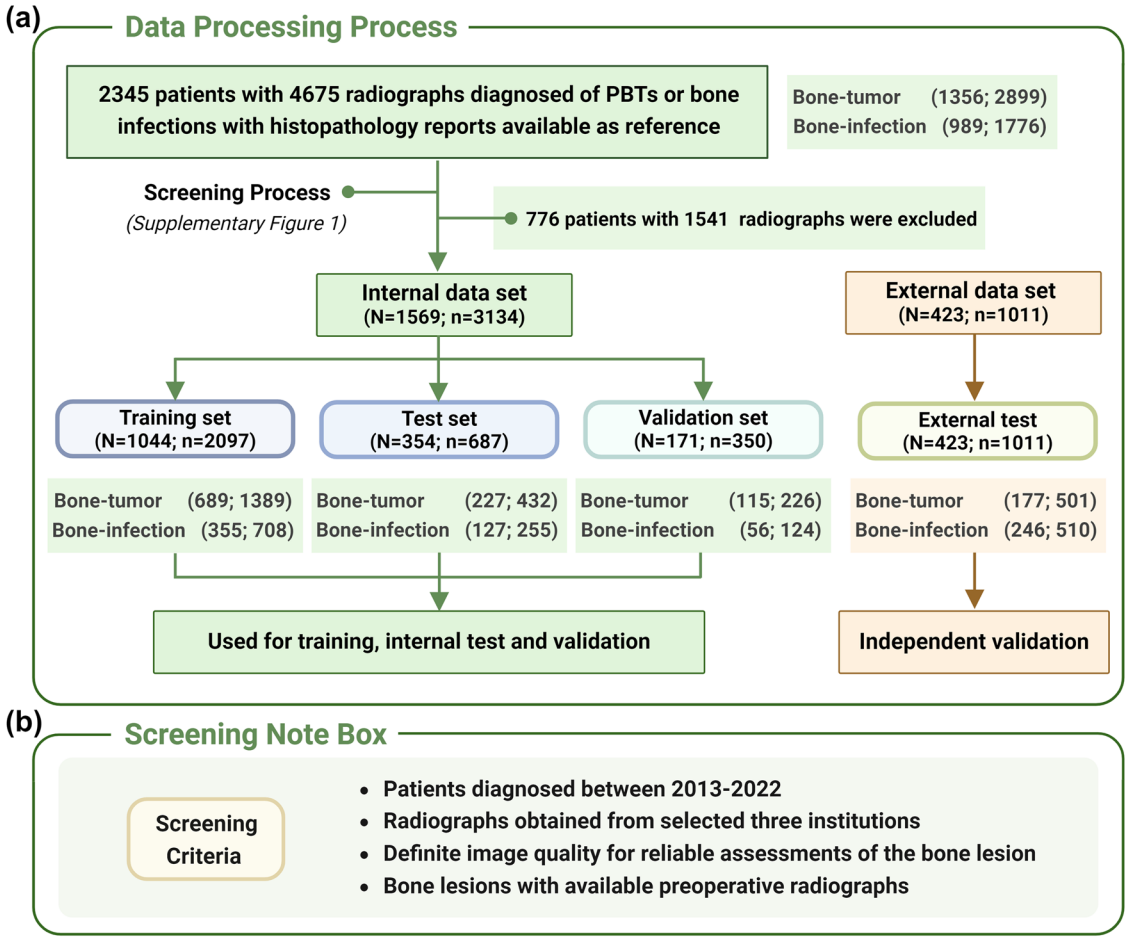

Fig. 2 | Data distribution and the screening criteria of the study. a Data processing process and data distribution across different datasets. **b** Screening criteria of the research. *n* number of the radiographs, *N* number of the patients. Note: Fig. 2 was Created with BioRender.com.

Table 2 | Performance of the models and radiologists of different seniority in internal and external test set

Modality	F1 score	ROC AUC (95% CI)	Accuracy	Sensitivity	Specificity	PPV	NPV	*P* value
Internal test set								
Ensemble model	0.834	0.948 (0.931–0.963)	0.881	0.808	0.924	0.862	0.891	NA
E3	0.879	0.903 (0.878–0.927)	0.843	0.849	0.830	0.912	0.725	<0.001*
E4	0.881	0.912 (0.890–0.934)	0.846	0.856	0.825	0.907	0.741	<0.001*
ViT	0.881	0.903 (0.880–0.927)	0.843	0.840	0.848	0.926	0.702	<0.001*
SWIN	0.903	0.946 (0.929–0.963)	0.872	0.863	0.892	0.947	0.745	0.835
EG1	0.822	NA	0.758	0.764	0.742	0.890	0.535	<0.001*
EG2	0.849	NA	0.802	0.815	0.774	0.887	0.659	<0.001*
EG3	0.870	NA	0.836	0.866	0.783	0.874	0.771	<0.001*
External test set								
Ensemble model	0.887	0.963 (0.951–0.973)	0.895	0.820	0.972	0.968	0.841	NA
E3	0.875	0.930 (0.914–0.946)	0.866	0.818	0.931	0.940	0.794	<0.001*
E4	0.883	0.946 (0.932–0.960)	0.874	0.818	0.953	0.960	0.790	<0.001*
ViT	0.883	0.951 (0.939–0.964)	0.871	0.803	0.977	0.982	0.763	0.002
SWIN	0.894	0.957 (0.944–0.969)	0.885	0.825	0.971	0.976	0.796	0.002

E3 EfficientNet B3, *E4* EfficientNet B4, *ViT* vision transformer, *SWIN* swin transformers, *CI* confidence interval, *PPV* positive predictive value, *NPV* negative predictive value. **P* values less than 0.05 are considered statistically significant.

the model, and more common training cases of tuberculosis come from spinal tuberculosis, resulting in a decrease in the accuracy of model interpretation. There are also cases in which both experts and models misclassify. Brodie abscess appears as a single osteolytic lesion on X-ray, accompanied by peripheral sclerosis with decreasing degree of peripheral sclerosis, which is difficult to distinguish from osteosarcoma and osteoid osteoma (Fig. 6f). When not accompanied by obvious sclerosis, it is difficult to distinguish Langerhans histiocytosis and Ewing sarcoma (Fig. 6e)[33].

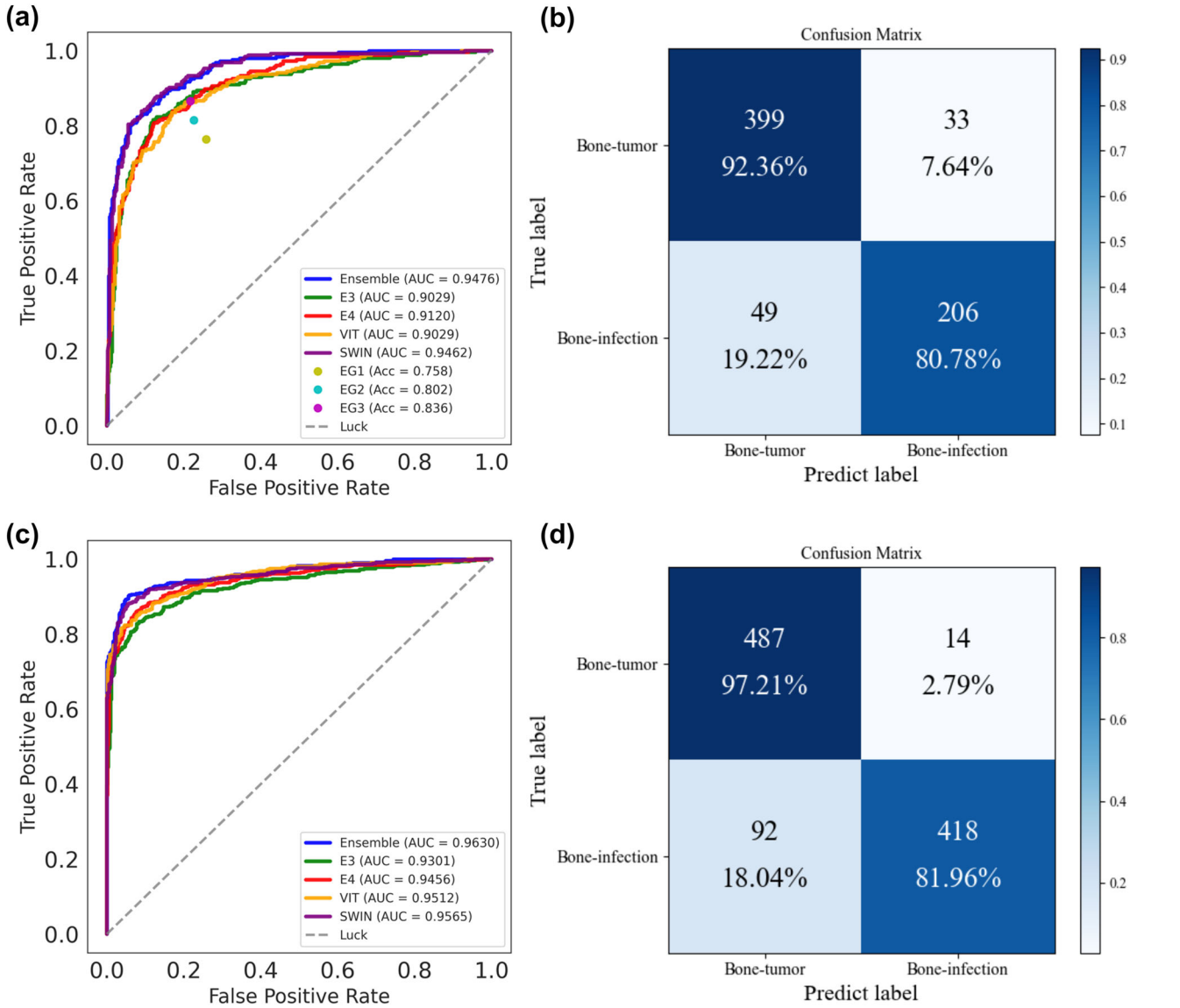

Fig. 3 | Confusion matrix and receiver operating characteristic (ROC) curve of the ensemble model for the binary classification. a, **b** ROC curve and confusion matrices of all models and radiologists' interpretations on the internal test set. **c**, **d** ROC curve and confusion matrices of all models on the external test set. Note: EG1= expert 1+ expert 2 (junior radiologist group); EG2= expert 3+ expert 4 (medium seniority group); EG3= expert 5+ expert 6 (senior radiologist group). EG expert group, E3 EfficientNet B3, E4 EfficientNet B4, ViT vision transformer, SWIN swin transformers, AUC area under the curve, Acc accuracy.

Table 3 | Performance of the experts and models in classifying high-frequency lesions in PBTs and bone infections in the internal test set

Bone tumors	***n*#**	**EG1**	**EG2**	**EG3**	**Expert average**	**E3**	**E4**	**ViT**	**SWIN**	**Ensemble**	**Model average**
Osteochondroma	90	95.6%	93.3%	93.9%	94.3%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Osteosarcoma	49	70.4%	91.8%	77.4%	79.9%	89.8%	95.9%	93.9%	95.9%	95.9%	94.3%
Fibrous dysplasia	49	85.7%	89.8%	92.9%	89.5%	77.6%	81.6%	83.7%	95.9%	87.8%	85.3%
GCT	46	94.6%	95.7%	91.3%	93.8%	91.3%	91.3%	91.3%	89.1%	87.0%	90.0%
Bone infections	***N***	**EG1**	**EG2**	**EG3**	**Expert average**	**E3**	**E4**	**ViT**	**SWIN**	**Ensemble**	**Model average**
Bone TB	170	56.5%	57.4%	86.9%	66.9%	85.3%	86.5%	84.1%	85.3%	90.6%	86.4%
Osteomyelitis	77	50.6%	53.2%	62.9%	54.1%	50.6%	48.7%	46.8%	57.14%	66.2%	54.8%

FDB fibrous dysplasia of bone, *GCT* giant cell of bone, *TB* tuberculosis, *EG* expert group, *E3* EfficientNet B3, *E4* EfficientNet B4, *ViT* vision transformer, *SWIN* swin transformers.
n refers to the number of the radiographs of related high-frequency lesions.

Discussion

Overall, our research introduced an innovative ensemble framework designed to detect and classify PBTs and bone infections concurrently. This framework incorporated two distinct single models: a radiograph-based imaging model and a clinical logistic regression model. By combining these models, we were able to enhance the classification accuracy of radiologists, surpassing the diagnostic capabilities of junior radiologists and aligning closely with those of medium senior radiologists. Our findings suggest that

Table 4 | Inter-reader reliability of the models and radiologists

Inter-reader reliability between the ensemble model and radiologists						
Fleiss κ (95% CI)	0.501 (0.463–0.538)					
Cohen κ (95% CI)	Expert 1	CSTC	Expert 2	CSTC	Expert 3	CSTC
Ensemble model	0.299 (0.265–0.333)	+ +	0.493 (0.456–0.531)	+ + +	0.456 (0.419–0.493)	+ + +
Cohen κ (95% CI)	Expert 3	CSTC	Expert 4	CSTC	Expert 6	CSTC
Ensemble model	0.356 (0.321–0.392)	+ + +	0.570 (0.532–0.607)	+ + +	0.596 (0.560–0.633)	+ + +
Inter-reader reliability among radiologists						
Fleiss κ (95% CI)	0.401 (0.364–0.438)					
Cohen κ (95% CI)	EG1	CSTC	EG2	CSTC	EG3	CSTC
	0.267 (0.234–0.300)	+ +	0.295 (0.261–0.329)	+ +	0.581 (0.544–0.618)	+ + +
Inter-reader reliability among models						
Fleiss κ (95% CI)	0.800 (0.770–0.830)					
Cohen κ (95% CI)	E3	CSTC	E4	CSTC	ViT	CSTC	SWIN	CSTC
Ensemble model	0.805 (0.775–0.835)	+ + + + +	0.793 (0.763–0.823)	+ + + +	0.783 (0.752–0.814)	+ + + +	0.908 (0.886–0.930)	+ + + + +

EG expert group, *CSTC* consistency, *E3* EfficientNet B3, *E4* EfficientNet B4, *ViT* vision transformer, *SWIN* swin transformers, *CI* confidence interval.
Note: EG1= expert 1+ expert 2 (junior radiologist group); EG2= expert 3+ expert 4 (medium seniority group); EG3= expert 5+ expert 6 (senior radiologist group).
CSTC evaluation (consistency evaluation):
0< Fleiss κ, Cohen κ ≤ 0.2, low consistency, "+".
0.2< Fleiss κ, Cohen κ ≤ 0.4, general consistency, "+ +".
0.4< Fleiss κ, Cohen κ ≤ 0.6, moderate consistency, "+ + +".
0.6< Fleiss κ, Cohen κ ≤ 0.8, high consistency, "+ + + +".
0.8< Fleiss κ, Cohen κ ≤ 1.0, extremely high consistency, "+ + + + +".

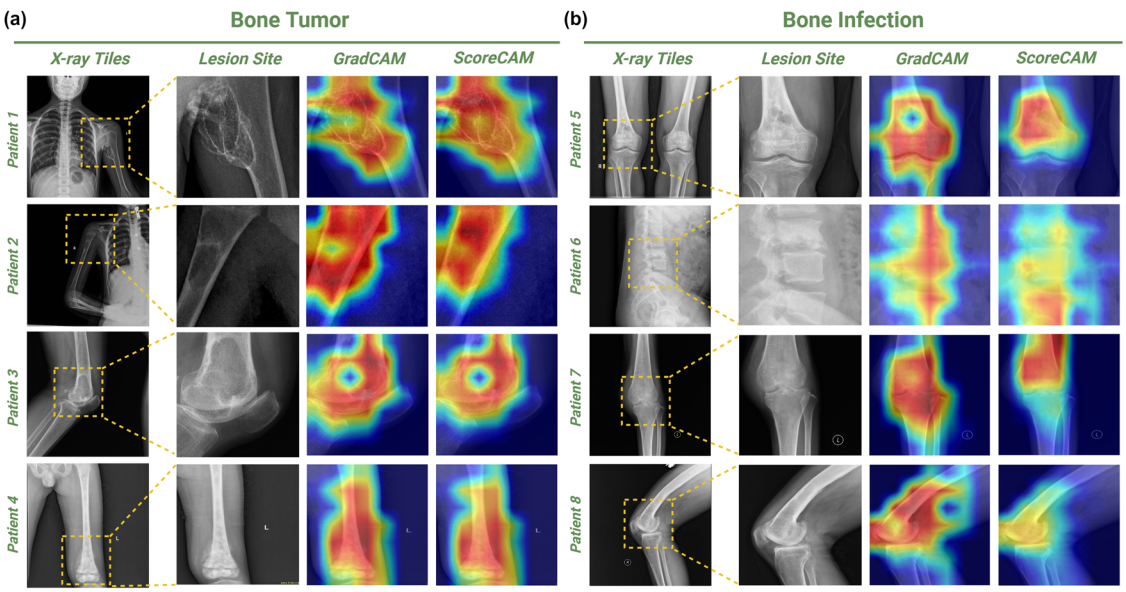

Fig. 4 | Visualization of PBTs and bone infections in four cases respectively. **a** Visualization of PBTs. Patient 1, a 10-year-old girl with chondrosarcoma on the left proximal humerus; Patient 2, a 10-year-old boy with a simple bone cyst on the right humerus; Patient 3, a 65-year-old female with giant cell tumor of bone on the left distal femur; Patient 4, a 9-year-old boy with osteosarcoma on the left distal femur. **b** Visualization of Bone infection. Patient 5, a 72-year-old male with chronic suppurative osteomyelitis of the lower right femur; Patient 6, a 31-year-old male with tuberculosis of lumbar vertebrae 3 and 4 with spinal canal stenosis; Patient 7, a 68-year-old female with tuberculosis of left knee joint; Patient 8, a 65-year-old male with right distal femoral osteomyelitis. Starting from the left, the first column is the original flat film image. The second column is an area cut as small as possible against the edge of the lesion. The third column is the GradCAM-generated heat map. The fourth is the heat map generated by ScoreCAM.

this ensemble approach holds promise for improving the accuracy and efficiency of detecting and classifying PBTs and bone infections in clinical settings.

In the realm of medical imaging, numerous deep learning models have been developed to aid in the diagnosis and classification of skeletal diseases using data from radiographs[28,34,35], CT[36–38], and MRIs[39–41]. However, the majority of these models have primarily concentrated on feature extraction from images and enhancing the accuracy of classification judgments to optimize model performance, neglecting the initial goal of utilizing deep learning as an auxiliary tool to enhance the diagnostic accuracy of clinicians. Consequently, our study aims to shed light on this issue by employing GradCAM and ScoreCAM to visualize the areas of focus within the models. In the course of our research, we have observed that GradCAM tends to prioritize the identification of bone hyperplasia and sclerosis, while overlooking areas of bone destruction. Conversely, ScoreCAM demonstrates a more balanced approach by highlighting both osteogenic and osteoclastogenic regions, resulting in a more precise delineation of lesion boundaries. This distinction underscores the importance of selecting the appropriate

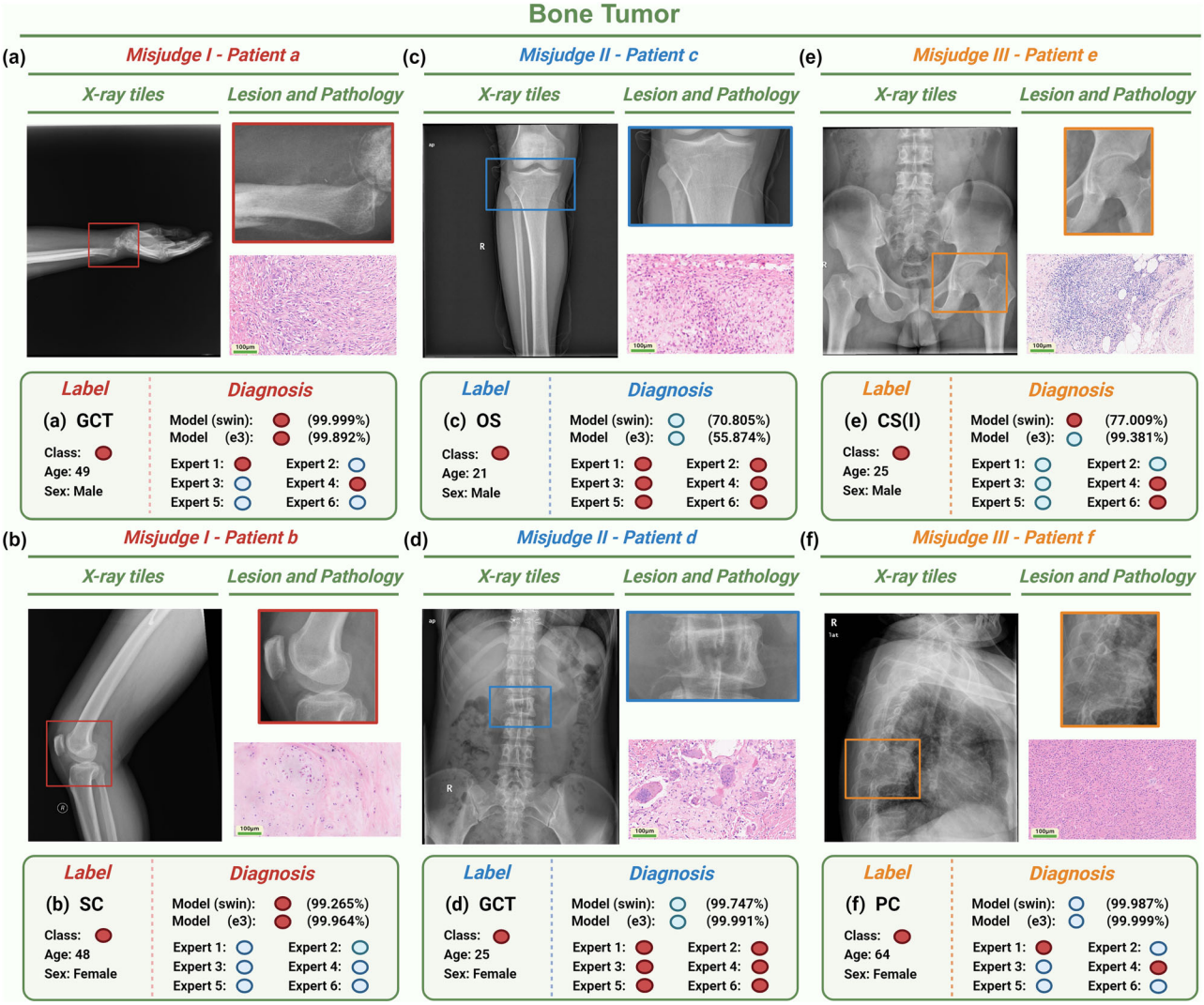

Fig. 5 | Bone tumor cases misclassified by experts and models in the internal test set. a, **b** The models mostly predict correctly but the experts mostly predict incorrectly based on the radiographs from Patient a and Patient b. **c**, **d** The models mostly predict incorrectly but the experts mostly predict correctly based on the radiographs from Patient c and Patient d. **e**, **f** Both of the models and the experts mostly predict incorrectly based on the radiographs from Patient e and Patient f. Model classification shows the probability of SWIN model and E3 model, which respectively correspond to the best and worst predictions in the imaging models. Red circles refer to bone tumors. Blue circles refer to bone infections. Bar = 100 μm. E3 EfficientNet B3, SWIN swin transformers, GCT giant cell of bone, SC Synovial chondromatosis, OS osteosarcoma, CS Chondrosarcoma, PC plasmacytoma. Note: Fig. 5 was Created with BioRender.com.

methodology for image analysis in order to achieve optimal results in the identification and characterization of bone abnormalities. Further investigation into the comparative effectiveness of these techniques may yield valuable insights for enhancing diagnostic accuracy and treatment planning in the field of medical imaging. Additionally, a group of experienced radiologists is enlisted to provide insightful clinical explanations for instances of misjudgment in representative cases, thereby facilitating a deeper comprehension of the models' functionality and ultimately improving its utility in the medical field.

Manual annotations of ROI which served as ground truth for various deep learning models have long been regarded as a relatively challenging and intricate task, especially in CT- or MRI-based deep learning models[37,42]. Despite the continuous emergence of novel segmentation algorithms in recent years like Mask R-CNN, 3D CNN[43,44] and so on, the segmentation performance of models built upon these algorithms often falls short of expectations. Issues such as misidentifying lesion locations or producing inaccurate segmentations frequently result in IoU and Dice scores that do not meet desired standards. Such discrepancies can introduce bias into subsequent classification model assessments and necessitate intricate manual verification and corrections in later stages. Therefore, in terms of research design, compared with multitask deep learning framework, our research prioritizes the accuracy and interpretability of the deep learning model. All of the segmentation and labeling of lesion areas in the radiographs are meticulously carried out by professional radiologists.

The utilization of deep learning techniques has significantly improved the clinical diagnosis of medical images in computer-assisted imaging settings. Despite these advancements, distinguishing between PBTs and bone infections remains a challenging task. Previous research has successfully developed and validated deep learning models for classifying different types of PBTs using radiographic and demographic data[28,45]. However, these studies primarily concentrate on categorizing benign, intermediate, and malignant PBTs, rather than differentiating bone tumors from other musculoskeletal diseases that may be easily confused with PBTs. It is worth noting that while MRI-based deep learning models have been created to enhance the diagnosis of patients with PBTs and bone infections[42], biases were present in the patient data collection due to variations in diagnosis and treatment protocols across different medical centers. Furthermore, these

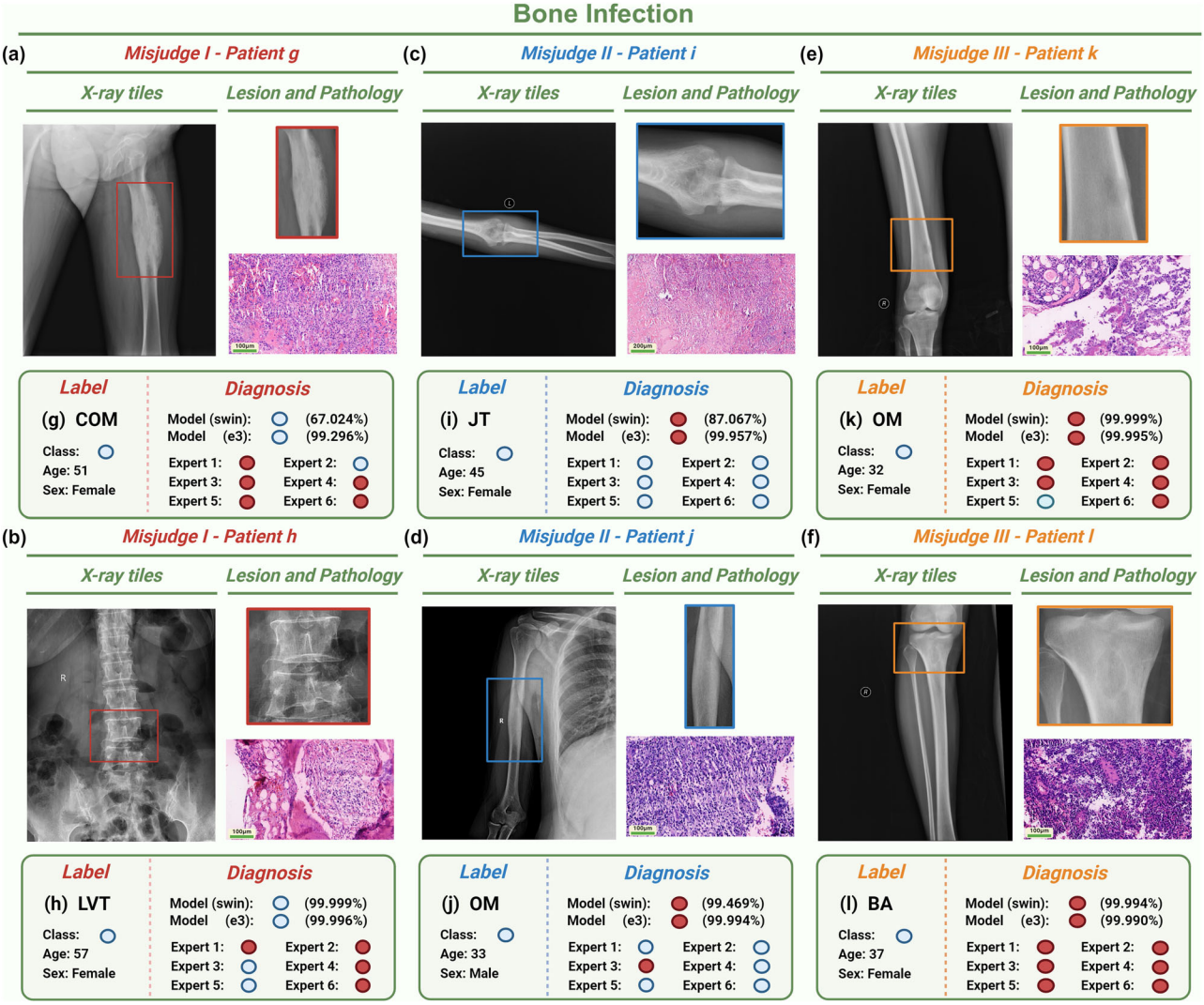

Fig. 6 | Bone infection cases misclassified by experts and models in the internal test set. a, **b** The models mostly predict correctly but the experts mostly predict incorrectly based on the radiographs from Patient g and Patient h. **c**, **d** The models mostly predict incorrectly but the experts mostly predict correctly based on the radiographs from Patient i and Patient j. **e**, **f** Both of the models and the experts mostly predict incorrectly based on the radiographs from Patient k and Patient l. Model classification shows the probability of SWIN model and E3 model, which respectively correspond to the best and worst predictions in the imaging models. Red circles refer to bone tumors. Blue circles refer to bone infections. Bar = 100 μm. E3 EfficientNet B3, SWIN swin transformers, COM chronic osteomyelitis, LVT lumber vertebra tuberculosis, JT joint tuberculosis, OM osteomyelitis, BA brodie's abscess. Note: Fig. 6 was Created with BioRender.com.

studies have overlooked important biomarkers such as CRP, ESR, ALP, lactate dehydrogenase (LDH) and so on. Combining the completeness of clinical information can better restore the original appearance and characteristics of the disease. Our ensemble model which encompasses sufficient clinical information outperformed the other four models merely based on the image data. These cases underscore the necessity for more systematic approaches to data gathering and organization, encompassing a broader spectrum of bone lesions and data points to enhance the accuracy of the models.

This study has limitations. Firstly, bone infections are more common than PBTs and benign subtypes in PBTs are far more common than malignant ones. However, because the hospitals selected were regionally superior medical centers, patients with intractable diseases have high tendency. Secondly, our external validation set includes a children's specialty hospital (Hospital 3), while it does help increase the diversity of our study population to some extent, making our research more representative, it may introduce some bias in terms of population distribution. Thirdly, the segmentation and labeling of lesion areas in the radiographs were entirely carried out by radiologists manually, making the research multifarious, although it may bring better work. In addition, in the collection process of clinical information, we found that for some examination like ALP and LDH, not all patients need this examination. In addition, doctors from different hospitals and departments may also exist examination preference, which lead to large amount of missing information. In the future, more cases with radiograph images from representative hospitals and more standardized collection of clinical information need to be researched to improve the generalizability and completeness of the model.

This groundbreaking study introduces a radiograph-based deep learning framework designed to enhance the classification of PBTs and bone infections, while also elucidating the clinical interpretation of these models. The ensemble deep learning framework, utilizing multicenter radiographs and clinical data, significantly improves the diagnostic accuracy for the binary classification. The results of the model have been meticulously visualized and professionally explained by expert radiologists. The ensemble model is more accurate and reliable in diagnosis compared with radiologists. These findings hold immense potential to guide orthopedic surgeons in making informed treatment decisions, thereby facilitating timely interventions for patients in need.

Methods

In this research, the methodology is mainly composed of data collection, preprocessing, annotation, model design, and development. The subsequent analysis was performed in compliance with all relevant ethical regulations, including the Declaration of Helsinki, as approved by the institutional review board of human studies of the Second Xiangya Hospital of Central South University (protocol number: no.2022-040) (Hospital 1). In addition, this retrospective study was approved by the local institutional review boards of Xiangya Hospital of Central South University (Hospital 2) and Hunan Children's Hospital of Central South University (Hospital 3), and informed consent was waived because of the retrospective nature[28]. The study was performed in accordance with national and international guidelines, and followed the recommended guidelines Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines (Supplementary Table 8)[46].

Research participants and data

This retrospective multicenter study collected patients via consecutive sampling between 2013 and 2022 from two cohorts: training cohort (from Hospital 1) and testing cohort (from Hospital 2 and Hospital 3) (Supplementary Fig. 1). After screening, 1569 patients diagnosed of PBTs or bone infections with histopathology reports available as reference were finally included in the internal dataset. While 423 patients from another two medical centers were collected for validation (Fig. 2a and Supplementary Fig. 1). These lesions were identified to have bone involvement through preoperative radiographs and were histologically diagnosed following biopsy or surgery. The criteria for evaluating the accuracy of both expert classifications and model classifications are grounded in pathological results, serving as the "ground truth". (i) For the inclusion criteria, lesions were confirmed and diagnosed as PBTs according to the 2020 World Health Organization (WHO) system for the classification for tumors of bone[1] while bone infections were confirmed and proven by histology and (or) bacterial culture. The other vital inclusion criteria are evident as well as available clinical information and preoperative radiographs. (ii) The screening criteria were respectively described in Fig. 2b: (a) radiographs were from patients diagnosed between 2013 and 2022 (b) in selected three hospitals; (c) radiographs with robust quality for reliable assessments of the bone lesions and (d) all of these radiographs were preoperative. With reference to previous literature[42,45,47], clinical characteristics of the included patients' contained age, gender, lesion position (appendicular or axial), "whether the lesion painful ?", "whether the lesion swelling ?", "whether a recent history of trauma ?", and we further collected examination data including C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), and alkaline phosphatase (ALP). All of the clinical data of the patients were reviewed and obtained from the patients' electronic medical records after data desensitization and standardization.

Image preprocessing and annotation

During the preprocessing stage, all of the radiographs were screened and selected based on the inclusion and exclusion criteria above. Notably, radiograph images like artifacts or foreign bodies which might significantly hinder the observation of lesions were regarded as poor-quality radiographs. One senior seniority radiologist (Y.H.) with systematic musculoskeletal fellowship training (12 years work experience) and one medium seniority clinical orthopedist (C.T.) (8 years work experience) independently reviewed these radiographs without the patients' information, and the quality of them would decide by consensus. Radiographs were kept and downloaded as Digital Imaging and Communications in Medicine (DICOM) files from the picture archiving and communication system (PACS) at their original sizes and resolutions. All of these radiograph images have undergone desensitization processing of disengaging patient-protected health information from DICOM data to meet the relevant legal criteria and requirements of US (HIPAA) as well as European (GDPR)[42]. Delineating the region of interest (ROI) was performed by two proficient radiologists (Y.Q. with 3-5 years of experience and J.G. with 3-5 years of experience in screening musculoskeletal radiographs images). ROIs were meticulously outlined via Click 2 Crop (version 5.2.2) (https://click-2-crop.en.softonic.com/) to closely segment pertinent entities present in each PBT or bone infection. Instances where disagreements arose between the two radiologists regarding contentious boundaries of these entities were subjected to further scrutiny. In such cases, a distinguished senior radiologist (Y.H.), boasting an impressive 12 years of experience in screening musculoskeletal radiographs, undertook the task of confirming the final delineations of ROIs. The smallest rectangular box that can completely cover the ROI was manually annotated as the boundary box by senior seniority radiologist (Y.H.) to ensure accuracy. Afterward, the annotated ROIs were used as ground truth for the model development process.

Design of the imaging models

For the classification of the radiographs, imaging models were built upon four distinct neural networks: EfficientNet B3 (E3), EfficientNet B4 (E4), Vision Transformer (ViT), and Swin Transformers (SWIN)[48–50]. These models were selected based on their state-of-the-art performance in image classification tasks and their ability to capture diverse features from medical images. Specifically, EfficientNet represents a lineage of Convolutional Neural Networks (CNNs) that utilize compound scaling to harmonize the depth, width, and resolution of the network, achieving optimal performance with fewer parameters compared to traditional CNNs[50]. Thanks to this innovative methodology, EfficientNet consistently attains state-of-the-art accuracy, yet with markedly fewer parameters. This makes it a prime choice for an array of computer vision applications[50,51]. The Vision Transformer (ViT) introduces a novel architecture that processes images as sequences of patches using Transformer blocks, originally designed for natural language processing tasks. This architecture has demonstrated significant potential in handling visual data. The Swin Transformer further refines this approach by incorporating a hierarchical structure and local self-attention mechanisms, enabling it to manage diverse resolutions and scales effectively. Collectively, these models represent some of the most advanced frameworks in computer vision.

Addressing the constraints of our limited label data, we adopted a transfer learning strategy. All four imaging models were initialized with weights pre-trained on the extensive ImageNet dataset, followed by fine-tuning on our proprietary bone dataset[52]. The original classification heads of these models, designed for 1000-class classification, were replaced with a single output node equipped with a sigmoid activation function to facilitate binary predictions (PBTs vs. bone infection).

Model training and evaluation

The internal dataset from Hospital 1 was partitioned into training, validation, and test set at a ratio of 7:1:2, respectively. The dataset from Hospital 2 and Hospital 3 was set aside as an external test set to evaluate the generalizability of our models across different data sources. Each of the four imaging models was trained independently using a batch size of 128 over 100 epochs. We employed Binary Cross-Entropy loss as our loss function. Optimization of the model was achieved through Stochastic Gradient Descent with an initial learning rate of 0.1. This rate was decayed by a factor of 10 every 30 epochs. For testing, we utilized the weights from the epoch exhibiting the best performance on the validation dataset.

Our algorithms were developed in Python 3.7 and executed on a machine equipped with an NVIDIA RTX 3090 GPU. The deep learning framework used in this study is PyTorch. In terms of data preprocessing, all images underwent resizing and normalization. Specifically, images were resized to a resolution of 224×224 pixels and normalized using the mean and standard deviation of the training dataset. To further enhance performance, we incorporated standard data augmentation techniques during training, including random horizontal and vertical flips with a probability of 0.5 for each.

Model ensemble

To further optimize performance, we integrated the predictions from the four imaging models (E3, E4, ViT, and SWIN) with traditional machine-

learning models based on patients' clinical characteristics. The hyperparameters utilized in the four imaging models and the ensemble model are depicted in Supplementary Table 9. Specifically, we designed and evaluated several machine-learning models, including Random Forest (RF), Adaptive Boosting (AdaBoost), Gradient Boosted Decision Trees (GBDT), Light Gradient Boosting Machine (LightGBM), Decision Tree (DT), Logistics Regression (LR), Extreme Gradient Boosting (XGBoost) and K-Nearest Neighbor (KNN). Given the missing clinical data and the significant differences in clinical features between PBTs and bone infections, the clinical characteristics included in the ensemble model were age, gender, and lesion location.

The construction of the ensemble model involved a two-step 5-fold cross-validation approach to avoid self-validation. In the first step, the four trained imaging models were used to score each patient (Supplementary Fig. 5). In the second step, these scores were integrated with clinical features using traditional machine-learning methods, with fivefold cross-validation utilized for hyperparameter tuning (Supplementary Fig. 6). Through systematic comparison, we determined that the ensemble model utilizing Random Forest achieved the highest AUC (Supplementary Fig. 7). The final ensemble framework integrates both clinical characteristics and imaging information, providing a comprehensive diagnostic tool for PBTs and bone infection classification.

Visualization and examples

To interpret the models' predictions, we use GradCAM and ScoreCAM to visualize the regions that our model relies on for decision-making. GradCAM calculates the gradient of the target class score with respect to feature maps. It then applies global-average-pooling to these gradients to determine the importance weights for each feature map. This weighted combination, when subjected to a ReLU activation, produces a coarse localization map highlighting the most relevant image regions. As GradCAM is model-agnostic, it can be applied to four different models in our approach. In contrast, ScoreCAM, an extension of GradCAM, does not use gradients. Instead, it activates each feature map in the target layer individually and forwards these to obtain the class score. The final saliency map is derived by linearly combining the activation maps with their respective scores. This results in sharper and more precise visual explanations than GradCAM provides. Together, these two methods offer insights into the regions of an X-ray that our model considers essential for predictions.

Radiologist evaluation

To assess and contrast the precision of clinical doctors and the classification judgments made by various deep learning models, we have enlisted the participation of three distinct groups of radiologists varying in seniority. Within this study, three expert groups (EG) with different seniority were designed. Individuals classified as junior radiologists possessed 2–4 years of experience (Q.L. and J.G.) and were responsible for analyzing 1500 musculoskeletal radiograph reports annually (EG1). While senior radiologists (Prof. P. and Prof. L.) had accumulated over 10 years of experience in the field (EG3)[42,47]. In addition, we engaged another group of refresher radiologists (M.W. and Y.Z.) with 8–10 years of experience referred as medium seniority group (EG2). Each radiologist independently evaluated radiographs and associated clinical data using a conventional PACS system, with the diagnoses being made without prior knowledge of the pathological and/or bacterial culture results. The inter-reader reliability among radiologists were evaluated through Fleiss κ and Cohen κ[53].

Statistics analysis

All statistical analyses were conducted using the opensource R software (version 4.2.3; R Foundation). Evaluation of the classification performance involved the use of the receiver operating characteristic (ROC) curve, along with metrics such as the area under the curve (AUC), accuracy, sensitivity, specificity, and confusion matrices. The mean AUC was specifically employed to assess the average performance of these four distinct imaging models. Statistical differences in clinicopathologic features among groups were analyzed using the Kruskal–Wallis rank-sum test for continuous variables and the chi-square test for categorical variables. Statistical differences between the AUC curves of different models were assessed using the DeLong test[54], while the statistical differences between the models and radiologist experts were evaluated using the Cochran's Q test[55,56], which is appropriate for multiple sets of paired data. Calculation of 95% confidence intervals (CI) was performed using the Wilson method. *P* values below 0.05 were considered as statistically significant.

Data availability

The raw data collected and processed in this study are supervised under the corresponding institutions. All of the imaging data in this study has been desensitized and publicly released with restricted access on Zenodo (https://zenodo.org/) at https://doi.org/10.5281/zenodo.13858807. This DOI represents all versions, and will always resolve to the latest one. The data are available by emailing the corresponding author with all requests for academic use. The requirements will be evaluated concerning institutional policies, and data can only be shared for non-commercial academic usage with a formal material transfer agreement. All requests will be promptly reviewed within a timeframe of 30 working days.

Code availability

The pipeline development and experiments are conducted in Python with PyTorch as a primary tool. All of the codes for reproducing this study (Deep learning pipeline for Tumors and Infections based on Radiographs Predicition) can be found at https://github.com/CSUXY-2YY/DeepTIRP.

Received: 1 August 2024; Accepted: 25 February 2025;
Published online: 13 March 2025

References

1. Choi, J. H. & Ro, J. Y. The 2020 WHO classification of tumors of bone: an updated review. *Adv. Anat. Pathol.* **28**, 119–138 (2021).
2. Ferguson, J. L. & Turner, S. P. Bone cancer: diagnosis and treatment principles. *Am. Fam. Physician* **98**, 205–213 (2018).
3. Molina, E. R., Chim, L. K., Barrios, S., Ludwig, J. A. & Mikos, A. G. Modeling the tumor microenvironment and pathogenic signaling in bone sarcoma. *Tissue Eng. Part B Rev.* **26**, 249–271 (2020).
4. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **73**, 17–48 (2023).
5. Bölling, T., Hardes, J. & Dirksen, U. Management of bone tumours in paediatric oncology. *Clin. Oncol. (R. Coll. Radiol.)* **25**, 19–26 (2013).
6. Zhang, W. et al. PRKDC induces chemoresistance in osteosarcoma by recruiting GDE2 to stabilize GNAS and activate AKT. *Cancer Res* **84**, 2873–2887 (2024).
7. Yu, S. & Yao, X. Advances on immunotherapy for osteosarcoma. *Mol. Cancer* **23**, 192 (2024).
8. Casali, P. G. et al. Bone sarcomas: ESMO-PaedCan-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **29**, iv79–iv95 (2018).
9. Liu, J. et al. Anticancer and bone-enhanced nano-hydroxyapatite/gelatin/polylactic acid fibrous membrane with dual drug delivery and sequential release for osteosarcoma. *Int. J. Biol. Macromol.* **240**, 124406 (2023).
10. Meltzer, P. S. & Helman, L. J. New horizons in the treatment of osteosarcoma. *New Engl. J. Med.* **385**, 2066–2076 (2021).
11. Li, J. et al. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. *Eur. Radiol.* **33**, 4237–4248 (2023).
12. Caracciolo, J. T., Temple, H. T., Letson, G. D. & Kransdorf, M. J. A modified lodwick-madewell grading system for the evaluation of lytic bone lesions. *AJR Am. J. Roentgenol.* **207**, 150–156 (2016).
13. Kovacs, S. K., Manassaporn, A., Nielsen, G. P. & Hung, Y. P. Molecular and immunohistochemical testing of bone tumours: review and update. *Histopathology* **82**, 794–811 (2023).

14. Tao, Y. et al. Qualitative histopathological classification of primary bone tumors using deep learning: a pilot study. *Front. Oncol.* **11**, 735739 (2021).
15. Rozeman, L. B., Cleton-Jansen, A. M. & Hogendoorn, P. C. Pathology of primary malignant bone and cartilage tumours. *Int. Orthop.* **30**, 437–444 (2006).
16. Kellish, A. S. et al. Reliability and accuracy in radiographic measurements of musculoskeletal tumors. *J. Orthop. Res.* **40**, 1654–1660 (2022).
17. Ulaner, G., Hwang, S., Landa, J., Lefkowitz, R. A. & Panicek, D. M. Musculoskeletal tumours and tumour-like conditions: common and avoidable pitfalls at imaging in patients with known or suspected cancer: Part B: malignant mimics of benign tumours. *Int. Orthop.* **37**, 877–882 (2013).
18. Ulaner, G., Hwang, S., Lefkowitz, R. A., Landa, J. & Panicek, D. M. Musculoskeletal tumors and tumor-like conditions: common and avoidable pitfalls at imaging in patients with known or suspected cancer: Part A: benign conditions that may mimic malignancy. *Int. Orthop.* **37**, 871–876 (2013).
19. Anderson, M. E., Wu, J. S. & Vargas, S. O. CORR (®) tumor board: do orthopaedic oncologists agree on the diagnosis and treatment of cartilage tumors of the appendicular skeleton? *Clin. Orthop. Relat. Res.* **475**, 2172–2175 (2017).
20. Benz, M. R., Crompton, J. G. & Harder, D. PET/CT variants and pitfalls in bone and soft tissue sarcoma. *Semin. Nucl. Med.* **51**, 584–592 (2021).
21. Rozenberg, A. et al. Clinical impact of second-opinion musculoskeletal subspecialty interpretations during a multidisciplinary orthopedic oncology conference. *J. Am. Coll. Radiol.* **14**, 931–936 (2017).
22. Rozenberg, A. et al. Second opinions in orthopedic oncology imaging: can fellowship training reduce clinically significant discrepancies? *Skelet. Radiol.* **48**, 143–147 (2019).
23. Mullowney, M. W. et al. Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* **22**, 895–916 (2023).
24. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
25. Zhu, W., Xie, L., Han, J. & Guo, X. The application of deep learning in cancer prognosis prediction. *Cancers* **12**, 603 (2020).
26. Chen, X. et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* **79**, 102444 (2022).
27. Painuli, D., Bhardwaj, S. & Köse, U. Recent advancement in cancer diagnosis using machine learning and deep learning techniques: a comprehensive review. *Comput Biol. Med.* **146**, 105580 (2022).
28. von Schacky, C. E. et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology* **301**, 398–406 (2021).
29. Murphey, M. D. et al. From the archives of AFIP. Imaging of giant cell tumor and giant cell reparative granuloma of bone: radiologic-pathologic correlation. *Radiographics* **21**, 1283–1309 (2001).
30. Miller, T. T. Bone tumors and tumorlike conditions: analysis with conventional radiography. *Radiology* **246**, 662–674 (2008).
31. Delorme, S. & Baur-Melnyk, A. Imaging in multiple myeloma. *Eur. J. Radiol.* **70**, 401–408 (2009).
32. van de Meent, M. M., Pichardo, S. E. C., Rodrigues, M. F., Verbist, B. M. & van Merkesteyn, J. P. R. Radiographic characteristics of chronic diffuse sclerosing osteomyelitis/tendoperiostitis of the mandible: a comparison with chronic suppurative osteomyelitis and osteoradionecrosis. *J. Craniomaxillofac. Surg.* **46**, 1631–1636 (2018).
33. Gould, C. F., Ly, J. Q., Lattin, G. E. Jr., Beall, D. P. & Sutcliffe, J. B. 3rd Bone tumor mimics: avoiding misdiagnosis. *Curr. Probl. Diagn. Radiol.* **36**, 124–141 (2007).
34. Hill, B. G., Krogue, J. D., Jevsevar, D. S. & Schilling, P. L. Deep learning and imaging for the orthopaedic surgeon: how machines "read" radiographs. *J. Bone Jt. Surg. Am.* **104**, 1675–1686 (2022).
35. Kijowski, R., Liu, F., Caliva, F. & Pedoia, V. Deep learning for lesion detection, progression, and prediction of musculoskeletal disease. *J. Magn. Reson Imaging* **52**, 1607–1619 (2020).
36. Liu, P. et al. Deep learning to segment pelvic bones: large-scale CT datasets and baseline models. *Int. J. Comput. Assist Radiol. Surg.* **16**, 749–756 (2021).
37. Noguchi, S. et al. Deep learning-based algorithm improved radiologists' performance in bone metastases detection on CT. *Eur. Radiol.* **32**, 7976–7987 (2022).
38. Arthur, A. et al. A CT-based radiomics classification model for the prediction of histological type and tumour grade in retroperitoneal sarcoma (RADSARC-R): a retrospective multicohort analysis. *Lancet Oncol.* **24**, 1277–1286 (2023).
39. Hallinan, J. et al. Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine MRI. *Radiology* **300**, 130–138 (2021).
40. Wennmann, M. et al. Combining deep learning and radiomics for automated, objective, comprehensive bone marrow characterization from whole-body MRI: a multicentric feasibility study. *Invest. Radiol.* **57**, 752–763 (2022).
41. Zheng, H. D. et al. Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from MRI. *Nat. Commun.* **13**, 841 (2022).
42. Ye, Q. et al. Automatic detection, segmentation, and classification of primary bone tumors and bone infections using an ensemble multi-task deep learning framework on multi-parametric MRIs: a multi-center study. *Eur. Radiol*. **34**, 4287–4299 (2023).
43. Bitarafan, A., Nikdan, M. & Baghshah, M. S. 3D image segmentation with sparse annotation by self-training and internal registration. *IEEE J. Biomed. Health Inf.* **25**, 2665–2672 (2021).
44. Pereira, H. M., Marchiori, E. & Severo, A. Magnetic resonance imaging aspects of giant-cell tumours of bone. *J. Med. Imaging Radiat. Oncol.* **58**, 674–678 (2014).
45. He, Y. et al. Deep learning-based classification of primary bone tumors on radiographs: a preliminary study. *EBioMedicine* **62**, 103121 (2020).
46. Mongan, J., Moy, L. & Kahn, C. E. Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).
47. Eweje, F. R. et al. Deep learning for classification of bone lesions on routine MRI. *EBioMedicine* **68**, 103402 (2021).
48. Heidari, M. et al. Enhancing efficiency in vision transformer networks: design techniques and insights. Preprint at https://arxiv.org/abs/2403.19882 (2024).
49. Liu, Z. & et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 10006-10017 (IEEE, 2021).
50. Tan, M. & Le, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 (eds. Kamalika, C. & Ruslan, S.) 6105--6114 (PMLR, Proceedings of Machine Learning Research, 2019).
51. Mozaffari, J., Amirkhani, A. & Shokouhi, S. B. A survey on deep learning models for detection of COVID-19. *Neural Comput. Appl.* 1–29 (2023).
52. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
53. Ragab, H. et al. DeePSC: a deep learning model for automated diagnosis of primary sclerosing cholangitis at two-dimensional MR cholangiopancreatography. *Radiol. Artif. Intell.* **5**, e220160 (2023).
54. Kato, M. et al. A machine learning model for predicting the lymph node metastasis of early gastric cancer not meeting the endoscopic curability criteria. *Gastric Cancer* **27**, 1069–1077 (2024).

55. Hou, P. et al. A paradigm shift in oncology imaging: a prospective cross-sectional study to assess low-dose deep learning image reconstruction versus standard-dose iterative reconstruction for comprehensive lesion detection in dual-energy computed tomography. *Quant. Imaging Med. Surg.* **14**, 6449–6465 (2024).
56. Cochran, W. G. The comparison of percentages in matched samples. *Biometrika* **37**, 256–266 (1950).

Acknowledgements

The authors would like to express our gratitude to BioRender (https://app.biorender.com/) for assistance in creating the figures (Figs. 1, 2, 5, and 6). The authors are very grateful for the active participation of radiologists with diverse seniority: junior radiologist group (Q.L. and J.G.); medium seniority group (M.W. and Y.Z.); senior radiologist group (Prof. P. and Prof. L.). This work was supported by the National Natural Foundation of China (82272664, 82172500 and 32300528), The Science and Technology Innovation Program of Hunan Province (2023RC3085, 2023RC3080), Hunan Provincial Health High-Level Talent Scientific Research Project (R2023054), Hunan Provincial Natural Science Foundation of China (2022JJ30843), the Science and Technology Development Fund Guided by Central Government (2021Szvup169), Hunan Provincial Administration of Traditional Chinese Medicine Project (D2022117), Hunan Provincial Health High-Level Talent Scientific Research Project (R2023054), Key Project of Scientific Research of the Education Department of Hunan Province (24A0008), the Excellent Youth Foundation of Hunan Scientific Committee (2024JJ2084), the Scientific Research Fund of Hunan Provincial Education Department (23B0023) and the Scientific Research Program of Hunan Provincial Health Commission (B202304077077). The study sponsors did not have any role in the study design, the collection, analysis and interpretation of data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author contributions

C.T. and H.D.X. conceived and designed the study, performed the data analysis. H.W. and Y.H. contributed to the data collection, results interpretation, and manuscript preparation. L.W., C.B.L. and Z.Q.L. were participated in data collection. Z.H.L. was responsible for the supervision of the project. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-00855-3.

Correspondence and requests for materials should be addressed to Haodong Xu or Chao Tu.