Original Article

The potential for different computed tomography-based machine learning networks to automatically segment and differentiate pelvic and sacral osteosarcoma from Ewing's sarcoma

Ping Yin[1], Wenjia Wang[2], Sicong Wang[2], Tao Liu[1], Chao Sun[1], Xia Liu[1], Lei Chen[1], Nan Hong[1]^

[1]Department of Radiology, Peking University People's Hospital, Beijing, China; [2]GE Healthcare, Shanghai, China

Contributions: (I) Conception and design: P Yin, N Hong; (II) Administrative support: L Chen, N Hong; (III) Provision of study materials or patients: P Yin, X Liu; (IV) Collection and assembly of data: P Yin, C Sun, T Liu; (V) Data analysis and interpretation: P Yin, W Wang, S Wang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Nan Hong. Department of Radiology, Peking University People's Hospital, 11 Xizhimen South Street, Xicheng District, Beijing 100044, China. Email: hongnan1968@163.com.

Background: This study aimed to explore optimal computed tomography (CT)-based machine learning and deep learning methods for the identification of pelvic and sacral osteosarcomas (OS) and Ewing's sarcomas (ES).

Methods: A total of 185 patients with pathologically confirmed pelvic and sacral OS and ES were analyzed. We first compared the performance of 9 radiomics-based machine learning models, 1 radiomics-based convolutional neural networks (CNNs) model, and 1 3-dimensional (3D) CNN model, respectively. We then proposed a 2-step no-new-Net (nnU-Net) model for the automatic segmentation and identification of OS and ES. The diagnoses by 3 radiologists were also obtained. The area under the receiver operating characteristic curve (AUC) and accuracy (ACC) were used to evaluate the different models.

Results: Age, tumor size, and tumor location showed significant differences between OS and ES (P<0.01). For the radiomics-based machine learning models, logistic regression (LR; AUC =0.716, ACC =0.660) performed best in the validation set. However, the radiomics-based CNN model had an AUC of 0.812 and ACC of 0.774 in the validation set, which were higher than those of the 3D CNN model (AUC =0.709, ACC =0.717). Among all the models, the nnU-Net model performed best, with an AUC of 0.835 and an ACC of 0.830 in the validation set, which was significantly higher than the primary physician's diagnosis (ACCs ranged from 0.757 to 0.811) (P<0.01).

Conclusions: The proposed nnU-Net model could be an end-to-end, non-invasive, and accurate auxiliary diagnostic tool for the differentiation of pelvic and sacral OS and ES.

Keywords: Radiomics; deep learning; classification; osteosarcomas; Ewing's sarcomas (ES)

Submitted Oct 01, 2022. Accepted for publication Mar 27, 2023. Published online Apr 13, 2023.
doi: 10.21037/qims-22-1042
View this article at: https://dx.doi.org/10.21037/qims-22-1042

^ ORCID: 0000-0003-0301-0514.

Introduction

Osteosarcoma (OS) is the most common primary bone malignancy affecting children and young adults, followed by Ewing's sarcoma (ES) (1-3). OS is most common in the metaphyseal region of long bones in the extremities, whereas ES is more common in the tubular bone, mostly in the diaphysis, and less in the metaphysis or epiphysis. The pelvis and sacrum are unusual locations for OS, and only about 8% of OS cases occur in the pelvis (4). ES in the pelvis and spine account for approximately 25% and 8% of all primary sites, respectively (5). Therefore, the diagnosis of OS and ES in the pelvis and sacrum is sometimes challenging for radiologists due to them being less common sites of occurrence (4).

Pelvic and sacral OS and ES also share numerous similar clinical and imaging features. They are most common in adolescents, often with pain as the earliest clinical symptom. Computed tomography (CT) findings of OS include a mixed-density mass, matrix mineralization, cortical destruction, and unclear margins (4). The common manifestations of ES are "insect-eaten" erosion or osmotic osteolytic lesions and fusiform periosteum reaction; however, rare manifestations of soft tissue components, such as sclerosis, calcification, and pathological fractures, may also occur (6). Tumor bone formation is the most characteristic manifestation of OS, but ES could also produce reactive new bone, which increases the difficulty of identification.

In clinical practice, the preoperative treatments of OS and ES differ. The latest Clinical Practice Guidelines for Bone Tumors published by the National Comprehensive Cancer Network in 2020 pointed out that ES treatment should include primary therapy, local treatment, and adjuvant chemotherapy. Primary treatment includes multidrug chemotherapy and appropriate growth factor support. Local control needs to be selected in accordance with tumor location, size, and age, including extensive resection, targeted radiotherapy combined with chemotherapy, and amputation. The standard treatment for OS is still surgery combined with chemotherapy and neoadjuvant chemotherapy (7). Furthermore, the clinical outcomes of OS and ES in the pelvis and sacrum have been shown to be inferior to those in other sites due to the larger sizes of the lesions when found and the difficulty in obtaining a wide surgical resection margin (5,8,9). Clinically, an auxiliary diagnostic tool is needed to improve the accuracy of distinguishing OS from ES to promptly formulate individualized treatment plans (10).

In the past decade, radiomics has been successfully used in the field of medicine, especially for the identification, prognosis, and efficacy evaluation of tumors (11). Deep learning has provided an opportunity to automatically extract imaging features to maximize model performance for the task at hand and has been widely used in medical imaging in recent years (12-19). Its application in the musculoskeletal system is also extensive, mainly focusing on prognostic prediction (20), treatment outcomes (21-23), and relapse (24,25). However, previous studies have not distinguished OS and ES in the pelvis and sacrum from those in other sites, or do not even include the pelvic and sacral regions. Dai *et al.* (2) developed and validated magnetic resonance (MR) radiomics models to identify OS and ES in the pelvis; however, they did not incorporate deep learning methods. Yin *et al.* (26) assessed the performance of the deep neural network and machine learning based on radiomics features to predict benign or malignant sacral tumors and found that both approaches performed well. Deep learning could take radiomics features or raw images as inputs but often requires a large sample size. To our knowledge, no studies have compared the performance of different machine learning and deep learning methods that are based on radiomics features and raw images simultaneously in differentiating OS and ES.

Therefore, the current study aimed to explore optimal CT-based machine learning and deep learning methods for the identification of pelvic and sacral OS and ES, possibly to provide an efficient and accurate auxiliary diagnostic tool for clinical practice. We present the following article in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-1042/rc).

Methods

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the Ethics Committee of Peking University People's Hospital. The requirement for written informed consent for this retrospective analysis was waived. A total of 201 patients with pathologically confirmed pelvic and sacral OS or ES in Peking University People's Hospital from June 2007 to June 2020 were retrospectively analyzed. All patients underwent CT examination within 1 month before

the first operation. A total of 16 patients were excluded due to obvious artifacts on the CT images. Age, sex, maximal tumor size, tumor location (zones I–IV) (27), and the history of malignancy of the patients were also analyzed.

All CT images were obtained using multi-detector row CT systems (Philips iCT 256; Philips Medical Systems, Best, Netherlands; GE Lightspeed VCT 64; GE Medical Systems, Chicago, IL, USA). The scanning parameters were as follows: 120 kV, 100–370 mAs, slice thickness = 5 mm, field of view = 350 mm × 350 mm, and matrix = 512 mm × 512 mm. The reconstruction methods were soft tissue and bone kernel algorithms.

Extraction and selection of radiomics features

MITK software version 2018.04.2 (www.mitk.org) was used for the semi-automatic segmentation of all tumors (28). Multilayer annotation was performed by 2 radiologists along the lesion contour on the axial, sagittal, and coronal planes, respectively, and then the 3-dimensional (3D) lesions were automatically generated using the 3D interpolation annotation tool. Subsequently, all lesions were manually corrected by another 2 musculoskeletal radiologists with 5 and 20 years of experience, respectively, who were blinded to the evaluation.

A total of 1,316 radiomics features of each patient were extracted from the CT images by using the Artificial Intelligence Kit software version 3.3.0 (AK; GE Healthcare, Shanghai, China), including 16 gray-level size-zone matrix features, 24 gray-level co-occurrence matrix features, 18 first-order histogram features, 14 gray-level dependence matrix features, 16 gray-level run-length matrix features, 14 shape features, 5 neighboring gray-tone difference matrix features, 186 Laplacian of Gaussian (LoG_{sigma} =2.0/3.0) features, 744 wavelet features, and 279 local binary pattern features.

Features selection was applied in the training database. Before conducting the analyses, variables with zero variance were excluded. Then, the median was used to fill in the missing value of the features, and Z-score standardization was conducted at the same time. Analysis of variance (ANOVA), correlation analysis with a threshold of 0.7, and least absolute shrinkage and selection operator (LASSO) regression were used to determine the optimal combination characteristics of the model. Finally, 4 radiomics features, including sigma_2_0_mm_3D_glcm_Inverse Variance, HLH_firstorder_Median, LHL_glcm_Cluster Shade, and LLL_glszm_Size ZoneNonUniformity, were left for constructing the radiomics-based machine learning and deep learning models.

Radiomics-based machine learning models

A total of 9 machine learning models were built based on radiomics features by using logistic regression (LR), random forest (RF), support vector machine (SVM), Bayes, decision tree (DT), k-nearest neighbor (KNN), Adaboost, Xgboost, and gradient boosting decision tree (GBDT).

All patients were randomly divided into training (n=130) and validation (n=55) datasets at a ratio of 7:3. The models were trained with the training set using the repeated 5-fold cross-validation method, and estimation performance was evaluated in the validation dataset. *Figure 1* shows the workflow of this study.

Radiomics-based deep learning model

A 3-layer deep-learning convolutional neural network (CNN) model, namely the multilayer perceptron (MLP) model, was also constructed with the selected 4 radiomics features as inputs. MLP is an artificial neural network that has also performed well in previous studies (29,30). The overall architecture is represented in *Figure 2*. Grid Search was used to select the best configuration of hidden layer number nodes. The node numbers for 2 hidden layers were set to 20 and 10, respectively. The learning rate was 0.0001. This radiomics-based deep learning model was derived using Pytorch 1.7.1 (https://pytorch.org), and the training was stopped when the model converged. The model was trained with 100 epochs. The optimizer was the stochastic gradient descent, and the loss function was the weighted cross entropy.

3D deep learning classification networks

The 3D CNN model was directly used to classify the lesions by the original image and its annotation information to avoid the influence of artificial feature selection on the model. CNNs are widely used deep learning architecture in medical image analysis tasks, such as image classification and segmentation. However, the performance of CNNs is significantly influenced by the volume of training data. Building a sufficiently large dataset is extremely challenging due to the difficulty of data acquisition and annotation in 3D medical imaging. For classification, the experiments showed that converging training from scratch was difficult.

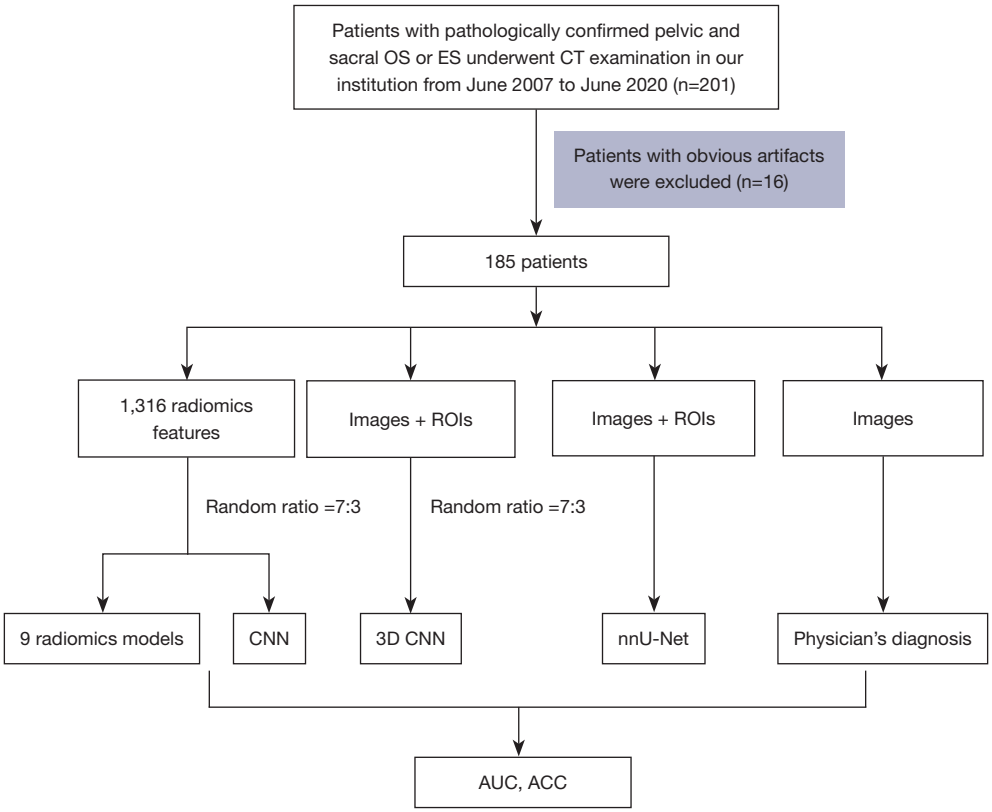

Figure 1 The workflow of this study. OS, osteosarcoma; ES, Ewing's sarcoma; CNN, convolutional neural networks; ROI, region of interest; AUC, area under the curve; ACC, accuracy.

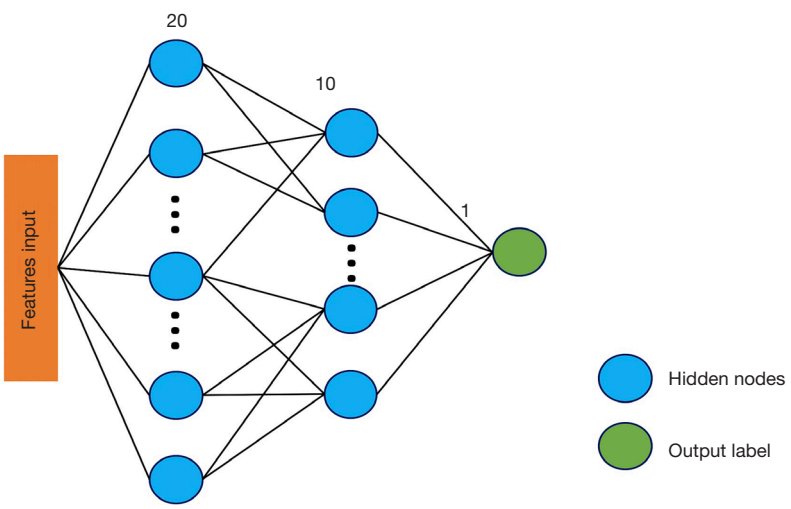

Figure 2 Overall architecture of the radiomics-based CNN model. The model consists of 1 input layer, 2 hidden layers, and 1 output layer. The selected radiomics features were taken as inputs, and the node numbers for 2 hidden layers were set to 20 and 10, respectively. CNN, convolutional neural network.

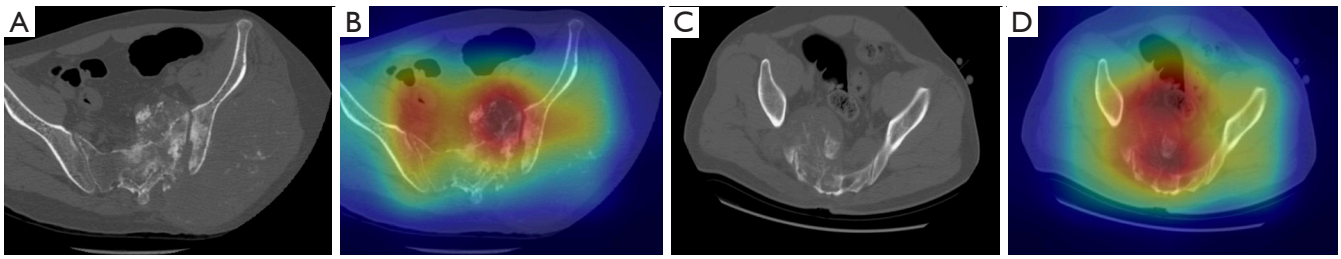

Figure 3 CT raw images and localization heatmaps of patients. (A,B) A 26-year-old male patient with OS; (C,D) A 12-year-old male patient with Ewing's sarcoma. (A,C) CT raw image; (B,D) localization heatmaps, which highlight the important regions for the CNN model to make diagnostic decisions. The model paid more attention to the red regions, which were important for the diagnostic decision, and the deep red regions overlapped with the tumor area. OS, osteosarcoma; CT, computed tomography; CNN, convolutional neural network.

Transfer learning is a powerful method to speed up training convergence and improve accuracy. Med3D (31), a co-train multidomain series of 3D medical models, was used as pre-trained models to initialize the weight. Direct classification was conducted using the original image as input. Data augmentation, such as random rotations, random scaling, random elastic deformations, gamma correction, and mirroring, was performed to avoid overfitting. Resnet18 (32) was used as the base architecture. The loss function was the combination of focal loss (33) and cross entropy in Eq. [1]. Data were shuffled to reduce variance and ensure that the models remained general and overfitted less. The Adam optimizer was used to train the network with a batch size of 8 and adjust the parameters in the supervised learning progress. The Grad-cam++ technique was applied to draw coarse localization heatmaps, which highlighted the important regions for the CNN model to make diagnostic decisions (34) (*Figure 3*). The area under the receiver operating characteristic (ROC) curve (AUC) and accuracy (ACC) were used as evaluation metrics. A total of 100 epochs were trained on the dataset.

$$L=L_{cross\ entropy}+L_{focal\ loss} \quad [1]$$

No-new-Net (nnU-Net) model

The performance of direct classification was relatively limited because of the millions of parameters in the classification network. However, this work only had hundreds of labels. A 2-stage deep learning architecture similar to that in *Figure 4* was proposed to better use the region of interest (ROI) annotations, which contained a segmentation network and a category label calculation module. First, nnU-Net (35) was used to condense and automate key decisions to design a successful segmentation pipeline for the dataset.

Preprocessing was conducted via the following steps: CT scans were cropped to the region with non-zero values to reduce the matrix size and computational burden. The datasets were then resampled to the median voxel spacing by using third-order spline interpolation for the image and neighbor interpolation for the mask to enable the networks to better learn spatial semantics. Finally, the entire dataset was normalized by clipping to the [0.5, 99.5] percentile of these intensity values, and the z-score was then normalized according to the mean and standard deviation of all collected intensity values.

In the training procedure, the 3d_fullres model was used to train the proposed model with a combination of dice and cross-entropy loss in Eq. [2]. The same data augmentation methods were applied to avoid overfitting. The number of training epochs was 100, at which epoch the model converged. We employed 5-fold cross-validation in the training progress. Connected component analysis was used as the postprocessing technique. After inference, each voxel point obtained a predictive probability for the classification. The category label of the whole image was determined according to the maximum number of voxels in 1 category.

$$L=L_{cross\ entropy}+L_{dice\ loss} \quad [2]$$

Physician's diagnosis

All images were read by 3 radiologists with less than 5 years of experience in musculoskeletal CT who were blinded to the radiologists' initial CT report and pathological and

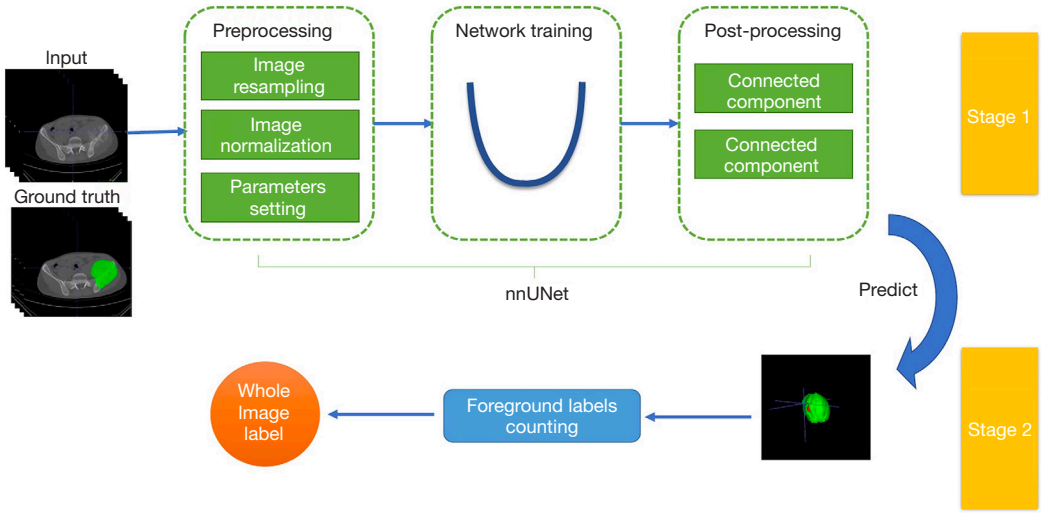

Figure 4 The 2-stage nnU-Net deep learning architecture. This contained a segmentation network and a category label calculation module. Preprocessing was performed before model training. In the training procedure, the model was trained with a combination of dice and cross-entropy loss. Data augmentation was performed to avoid overfitting. Connected component analysis was used as the postprocessing technique. After inference, each voxel point obtained a predictive probability for the classification. The category label of the whole image was determined according to the maximum number of voxels in one category.

clinical information.

Statistical analysis

The AUC and ACC were used to evaluate the performance of the models. R (version 3.5.1; R Foundation for Statistical Computing, Vienna, Austria) and Python (version 3.5.6; Python Software Foundation, Wilmington, DE, USA) were used for all statistical analyses. The Mann–Whitney U test was performed to compare the continuous variables, and the chi-squared test or Fisher's exact was used to compare categorical variables between groups. A comparison of the diagnostic ACC between the radiologists and the nnU-Net model was performed using McNemar's chi-squared test. The dice score was used to evaluate segmentation. A 2-tailed P<0.05 indicated statistical significance.

Results

Patient characteristics

A total of 185 patients [113 males and 72 females, median age of 22.0 (16.0, 30.0) years, and age range of 4–75 years] were included in this study (*Table 1*). Age, tumor size, and tumor location showed significant differences between the groups (Z_{age}=−4.864, Z_{size}=−5.027, $\chi^2_{location}$=16.836, P<0.01). However, no significant difference was found in terms of sex and history of malignancy between the groups (χ^2_{sex}=2.443, $\chi^2_{history\ of\ malignancy}$=0.013, P>0.05).

Performance of the radiomics classification models

Among the 9 radiomics models, Xgboost performed best (AUC =0.823, ACC =0.836) in the training group, whereas GBDT performed worst (AUC =0.625, ACC =0.680). In the validation group, LR and RF performed better than the other models, and their AUC values were all greater than 0.7. However, LR had a higher ACC value. Therefore, combining the AUC and ACC values showed that LR performed best in the validation group (AUC =0.716, ACC =0.660) among all radiomics models (*Figure 5* and *Table 2*).

Performance of the deep learning classification models

The radiomics-based CNN model achieved an AUC of 0.885 and an ACC of 0.811 in the training set and an AUC of 0.812 and an ACC of 0.774 in the validation set. These values were higher than those in the radiomics-based machine learning models.

Table 1 General characteristics of the included patients

Variable	Ewing's sarcoma	Osteosarcoma	χ^2/Z value	P value
Sex, n (%)			2.443	0.118
Female	26 (32.50)	46 (43.81)		
Male	54 (67.50)	59 (56.19)		
Age (years), median (interquartile range)	17.0 (13.0, 25.6)	26.0 (19.0, 34.3)	−4.864	<0.001
Size (cm), median (interquartile range)	7.90 (5.33, 10.31)	11.90 (8.00, 14.46)	−5.027	<0.001
Location, n (%)			16.836	0.002
Location I	22 (27.50)	28 (26.67)		
Location II	5 (6.25)	8 (7.62)		
Location III	10 (12.50)	8 (7.62)		
Location IV	28 (35.00)	16 (15.24)		
Multi-location	15 (18.75)	45 (42.86)		
History of malignancy, n (%)			0.013	0.909
No	75 (93.75)	98 (93.33)		
Yes	5 (6.25)	7 (6.67)		

Location I includes the iliac crest, Location II includes the acetabulum and its surroundings, Location III includes the pubis and ischium regions, and Location IV refers to the sacrum region. "Multi-location" refers to a tumor that involves more than one area simultaneously.

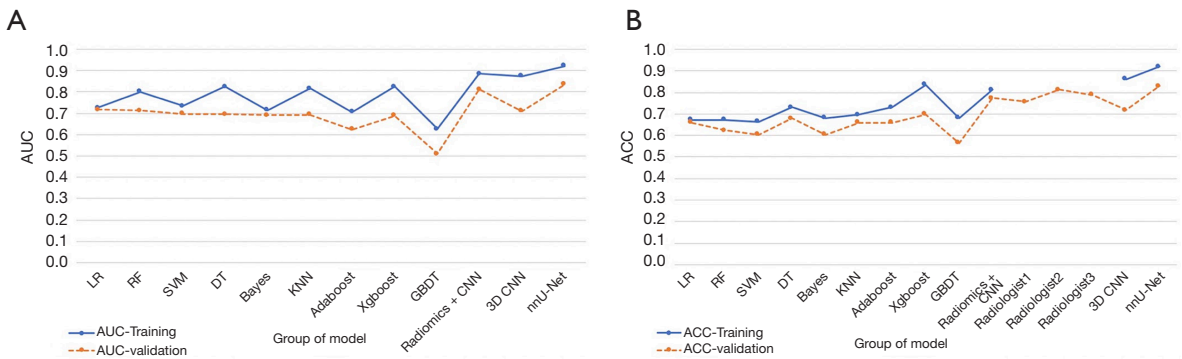

Figure 5 Scatterplots depicting the AUC and ACC of the different models. (A) The AUC scatterplots showed that the nnU-Net model performed best, with an AUC of 0.922 in the training set and an AUC of 0.835 in the validation set. (B) The ACC scatterplots showed that the nnU-Net model performed best, with an ACC of 0.919 in the training set and an ACC of 0.830 in the validation set. AUC, area under the curve; ACC, accuracy.

The performance of the 3D CNN model was relatively poor, only achieving an AUC of 0.709 and an ACC of 0.717 in the validation set, corresponding to a sensitivity of 0.767, a specificity of 0.652, a positive predictive value (PPV) of 0.742, and a negative predictive value (NPV) of 0.682.

The nnU-Net model combines the functions of segmentation and classification. For segmentation, the average dice score of the model was 0.77. For classification, the nnU-Net model performed best among all of the models, with an AUC of 0.922 and an ACC of 0.919 in the training set (sensitivity =0.9, specificity =0.943, PPV =0.955, NPV =0.877), an AUC of 0.835, and an ACC of 0.830 in the validation set (sensitivity =0.8, specificity =0.870, PPV =0.889, NPV =0.769).

Table 2 Performance of the different models in the training and validation sets

Models' or radiologists' performance	AUC	ACC	Sensitivity	Cut-off	Specificity	PPV	NPV
LR	0.725 (0.716)	0.672 (0.660)	0.788 (0.652)	0.390	0.586 (0.667)	0.586 (0.600)	0.788 (0.714)
RF	0.800 (0.711)	0.672 (0.623)	0.558 (0.391)	0.378	0.757 (0.800)	0.630 (0.600)	0.697 (0.632)
SVM	0.733 (0.696)	0.664 (0.604)	0.481 (0.391)	0.409	0.800 (0.767)	0.641 (0.562)	0.675 (0.622)
DT	0.824 (0.695)	0.730 (0.679)	0.923 (0.826)	0.541	0.586 (0.567)	0.623 (0.594)	0.911 (0.810)
Bayes	0.715 (0.690)	0.680 (0.604)	0.519 (0.261)	0.267	0.800 (0.867)	0.659 (0.600)	0.691 (0.605)
KNN	0.816 (0.693)	0.697 (0.660)	0.654 (0.609)	0.400	0.729 (0.700)	0.642 (0.609)	0.739 (0.700)
Adaboost	0.705 (0.624)	0.730 (0.660)	0.538 (0.348)	0.474	0.871 (0.900)	0.757 (0.727)	0.718 (0.643)
Xgboost	0.823 (0.688)	0.836 (0.698)	0.731 (0.609)	0.423	0.914 (0.767)	0.864 (0.667)	0.821 (0.719)
GBDT	0.625 (0.510)	0.680 (0.566)	0.25 (0.087)	0.441	1.000 (0.933)	1.000 (0.500)	0.642 (0.571)
Radiomics + CNN	0.885 (0.812)	0.811 (0.774)	0.865 (0.826)	0.425	0.771 (0.733)	0.738 (0.704)	0.885 (0.846)
Radiologist1	–	0.757	0.739	–	0.768	0.675	0.819
Radiologist2	–	0.811	0.8	–	0.818	0.75	0.857
Radiologist3	–	0.789	0.753	–	0.8	0.725	0.822
3D CNN	0.874 (0.709)	0.862 (0.717)	0.786 (0.767)	0.495	0.962 (0.652)	0.965 (0.742)	0.773 (0.682)
nnU-Net	0.922 (0.835)	0.919 (0.830)	0.900 (0.800)	0.506	0.943 (0.870)	0.955 (0.889)	0.877 (0.769)

LR, logistic regression; RF, random forest; SVM, support vector machine; DT, decision tree; KNN, k-nearest neighbor; GBDT, gradient boosting decision tree; CNN, convolutional neural network; AUC, area under the curve; ACC, accuracy; PPV, positive predictive value; NPV, negative predictive value. Training set, in front of the brackets. Validation set, in brackets.

Results of radiologist's diagnosis

The ACCs of the 3 primary radiologists' judgments were 0.757, 0.811, and 0.789, and the mean value was 0.786. The results of the doctors' judgments were higher than those of the 3D CNN and radiomics models. However, the performance of the nnU-Net model was significantly higher than the primary physician's diagnosis ($\chi^2_{\text{radiologist 1 } vs. \text{ nnU-Net}}$ =23.04, $\chi^2_{\text{radiologist 2 } vs. \text{ nnU-Net}}$ =13.07, $\chi^2_{\text{radiologist 3 } vs. \text{ nnU-Net}}$ =19.05, P<0.01).

Discussion

In this study, the performance of different machine learning and deep learning methods was investigated, and a 2-step nnU-Net model was proposed for the automatic segmentation and identification of OS and ES. Significant differences were found in terms of age, tumor size, and tumor location for differentiating pelvic and sacral OS and ES. Among the 9 radiomics models, LR performed best in the validation group. The radiomics-based CNN model performed better than the 3D CNN model. Also, the nnU-Net model performed best among all the models, with an AUC of 0.835 and an ACC of 0.830 in the validation set, which were higher than those of the primary physician's diagnosis. The proposed nnU-Net model could automatically segment and identify OS and ES lesions. Thus, it is not only more efficient than the radiomics and 3D CNN model but also more convenient and practical, and does not involve manual intervention in the whole process.

In this study, significant differences were found in terms of age, tumor size, and tumor location for differentiating pelvic and sacral OS and ES. Consistent with the result of a previous study (5), the median age of ES was significantly lower than that of OS. ES usually occurs in the second decade of life, whereas OS has 2 peak periods, namely, adolescence and older adulthood (2). Furthermore, the maximum tumor size of ES was significantly smaller than that of OS. ES is also more prevalent in the sacral region, followed by the iliac crest and multiple sites, whereas OS was more prevalent in multiple sites, followed by the iliac

crest and sacral regions. Previous studies have reported that OS and ES in the pelvis often involve the sacroiliac joint in multiple locations (4,8). This finding may be related to the fact that OS is generally larger and more likely to involve multiple areas.

Until now, few studies on the differentiation of pelvic and sacral tumors using machine learning methods have been conducted (2,3,26,36-38). Although the data analysis process is relatively cumbersome, the traditional radiomics analysis is suitable for small sample data. In the present study, the traditional radiomics method was used first, and the effects of 9 machine learning models were compared to distinguish OS from ES. Consistent with the result of a previous study (26), LR performed best in the validation set. Due to the poor performance of the radiomics models, a CNN model was then constructed based on previously filtered radiomics features. The results showed that the radiomics-based CNN model performed better than the conventional machine learning methods. However, the performance of the radiomics-based CNN model is still not very satisfactory, and the analysis process is complicated due to the use of the radiomics analysis method.

Therefore, a deep learning method without the need for manual intervention that only required raw images and ROIs as inputs was further adopted. Contrary to expectations, the performance of the 3D CNN model was not very good and overfitted. This may be related to the fact that CNN usually requires a large sample size, and this study only included 100 cases. More importantly, the 3D CNN model also requires delineated lesions as inputs, which is a time-consuming process.

We then applied the nnU-Net model to further simplify the process and improve efficiency. Isensee *et al.* (35) were the first to propose nnU-Net, which is a deep learning-based segmentation method that automatically configures itself, including preprocessing, network architecture, model training, and post-processing, for any new task. It only requires the original image as input, and can automatically complete the segmentation of lesions. Unlike Isensee *et al.*'s study (35), nnU-Net was applied in the present work to first perform automatic segmentation of the lesions on OS and ES images and then classify them without manual intervention, forming an end-to-end process. Considering that deep learning models usually require large data samples, this voxel-based prediction method, which was based on the principle that the minority is subordinate to the majority, greatly optimized the accuracy of this classification task (39). Therefore, the problem of insufficient input of small sample data analysis could be solved, and the stability and efficiency of the model could be improved. The results demonstrated that the nnU-Net model performed best among all models and better than the radiologists. Using the proposed model, OS or ES lesions could be automatically identified by only inputting the original images, without manual delineation or any intervention.

This study has certain limitations that should be noted. First, this was a retrospective study, and all the images were collected on 2 different machines. Although the data were preprocessed, they may still have affected the performance of the model. Second, this was a single-center study with a small sample size and lacking an external verification group. Although a suitable method for a small sample of data was developed, more external data are still needed for validation. Third, to better compare the performance of various machine-learning and deep-learning models, only CT data were included. In the future, more data (such as semantic features, clinical data, and genes) will be included to improve the performance of the proposed model.

Conclusions

The proposed nnU-Net model could be an end-to-end, non-invasive, and accurate auxiliary diagnostic tool for the differentiation of pelvic and sacral OS and ES, which may relatively improve the diagnostic efficiency of clinicians.

Acknowledgments

Funding: This study was supported by the National Natural Science Foundation of China (No. 82001764), the Peking University People's Hospital Scientific Research Development Funds (Nos. RDY2020-08, RS2021-10), and the Beijing United Imaging Research Institute of Intelligent Imaging Foundation (No. CRIBJQY202105).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-1042/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-1042/coif). WW and SW are employees of GE Healthcare. The other authors have no 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). This study was approved by the Ethics Committee of Peking University People's Hospital, and the requirement for written informed consent for this retrospective analysis was waived.

References

1. Zhao X, Wu Q, Gong X, Liu J, Ma Y. Osteosarcoma: a review of current and future therapeutic approaches. Biomed Eng Online 2021;20:24.
2. Dai Y, Yin P, Mao N, Sun C, Wu J, Cheng G, Hong N. Differentiation of Pelvic Osteosarcoma and Ewing Sarcoma Using Radiomic Analysis Based on T2-Weighted Images and Contrast-Enhanced T1-Weighted Images. Biomed Res Int 2020;2020:9078603.
3. Kubo T, Furuta T, Johan MP, Adachi N, Ochi M. Percent slope analysis of dynamic magnetic resonance imaging for assessment of chemotherapy response of osteosarcoma or Ewing sarcoma: systematic review and meta-analysis. Skeletal Radiol 2016;45:1235-42.
4. Park SK, Lee IS, Cho KH, Lee YH, Yi JH, Choi KU. Osteosarcoma of pelvic bones: imaging features. Clin Imaging 2017;41:59-64.
5. Zhu C, Olson KA, Roth M, Geller DS, Gorlick RG, Gill J, Laack NN, Randall RL. Provider views on the management of Ewing sarcoma of the spine and pelvis. J Surg Oncol 2018;117:417-24.
6. Patnaik S, Yarlagadda J, Susarla R. Imaging features of Ewing's sarcoma: Special reference to uncommon features and rare sites of presentation. J Cancer Res Ther 2018;14:1014-22.
7. Ni M. Update and interpretation of 2021 National Comprehensive Cancer Network (NCCN) "Clinical Practice Guidelines for Bone Tumors". Zhongguo Xiu Fu Chong Jian Wai Ke Za Zhi 2021;35:1186-91.
8. Rajiah P, Ilaslan H, Sundaram M. Imaging of sarcomas of pelvic bones. Semin Ultrasound CT MR 2011;32:433-41.
9. Liang H, Guo W, Yang R, Tang X, Yan T, Ji T, Yang Y, Li D, Xie L, Xu J. Radiological characteristics and predisposing factors of venous tumor thrombus in pelvic osteosarcoma: A mono-institutional retrospective study of 115 cases. Cancer Med 2018;7:4903-13.
10. Sambri A, Fujiwara T, Fiore M, Giannini C, Zucchini R, Cevolani L, Donati DM, De Paolis M. The Role of Imaging in Computer-Assisted Tumor Surgery of the Sacrum and Pelvis. Curr Med Imaging 2022;18:137-41.
11. Zhong J, Hu Y, Si L, Jia G, Xing Y, Zhang H, Yao W. A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation. Eur Radiol 2021;31:1526-35.
12. Ryu SM, Seo SW, Lee SH. Novel prognostication of patients with spinal and pelvic chondrosarcoma using deep survival neural networks. BMC Med Inform Decis Mak 2020;20:3.
13. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J Arthroplasty 2018;33:2358-61.
14. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. Radiology 2019;290:290-7.
15. Shimizu H, Nakayama KI. Artificial intelligence in oncology. Cancer Sci 2020;111:1452-60.
16. 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.
17. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, Gallivanone F, Cozzi A, D'Amico NC, Sardanelli F. AI applications to medical images: From machine learning to deep learning. Phys Med 2021;83:9-24.
18. Qu B, Cao J, Qian C, Wu J, Lin J, Wang L, Ou-Yang L, Chen Y, Yan L, Hong Q, Zheng G, Qu X. Current development and prospects of deep learning in spine image analysis: a literature review. Quant Imaging Med Surg 2022;12:3454-79.
19. Lin H, Xiao H, Dong L, Teo KB, Zou W, Cai J, Li T.

Deep learning for automatic target volume segmentation in radiation therapy: a review. Quant Imaging Med Surg 2021;11:4847-58.
20. Chen W, Zhou C, Yan Z, Chen H, Lin K, Zheng Z, Xu W. Using machine learning techniques predicts prognosis of patients with Ewing sarcoma. J Orthop Res 2021;39:2519-27.
21. Huang B, Wang J, Sun M, Chen X, Xu D, Li ZP, Ma J, Feng ST, Gao Z. Feasibility of multi-parametric magnetic resonance imaging combined with machine learning in the assessment of necrosis of osteosarcoma after neoadjuvant chemotherapy: a preliminary study. BMC Cancer 2020;20:322.
22. Lin P, Yang PF, Chen S, Shao YY, Xu L, Wu Y, Teng W, Zhou XZ, Li BH, Luo C, Xu LM, Huang M, Niu TY, Ye ZM. A Delta-radiomics model for preoperative evaluation of Neoadjuvant chemotherapy response in high-grade osteosarcoma. Cancer Imaging 2020;20:7.
23. Dufau J, Bouhamama A, Leporq B, Malaureille L, Beuf O, Gouin F, Pilleul F, Marec-Berard P. Prediction of chemotherapy response in primary osteosarcoma using the machine learning technique on radiomic data. Bull Cancer 2019;106:983-99.
24. Liu J, Lian T, Chen H, Wang X, Quan X, Deng Y, Yao J, Lu M, Ye Q, Feng Q, Zhao Y. Pretreatment Prediction of Relapse Risk in Patients with Osteosarcoma Using Radiomics Nomogram Based on CT: A Retrospective Multicenter Study. Biomed Res Int 2021;2021:6674471.
25. Chen H, Liu J, Cheng Z, Lu X, Wang X, Lu M, Li S, Xiang Z, Zhou Q, Liu Z, Zhao Y. Development and external validation of an MRI-based radiomics nomogram for pretreatment prediction for early relapse in osteosarcoma: A retrospective multicenter study. Eur J Radiol 2020;129:109066.
26. Yin P, Mao N, Chen H, Sun C, Wang S, Liu X, Hong N. Machine and Deep Learning Based Radiomics Models for Preoperative Prediction of Benign and Malignant Sacral Tumors. Front Oncol 2020;10:564725.
27. Enneking WF, Dunham WK. Resection and reconstruction for primary neoplasms involving the innominate bone. J Bone Joint Surg Am 1978;60:731-46.
28. Wolf I, Vetter M, Wegner I, Böttger T, Nolden M, Schöbinger M, Hastenteufel M, Kunert T, Meinzer HP. The medical imaging interaction toolkit. Med Image Anal 2005;9:594-604.
29. Bou Assi E, Gagliano L, Rihana S, Nguyen DK, Sawan M. Bispectrum Features and Multilayer Perceptron Classifier to Enhance Seizure Prediction. Sci Rep 2018;8:15491.
30. Kwon K, Kim D, Park H. A parallel MR imaging method using multilayer perceptron. Med Phys 2017;44:6209-24.
31. Chen S, Ma K, Zheng Y. "Med3d: Transfer learning for 3d medical image analysis." arXiv preprint arXiv:1904.00625 (2019).
32. Tandel GS, Tiwari A, Kakde OG, Gupta N, Saba L, Suri JS. Role of Ensemble Deep Learning for Brain Tumor Classification in Multiple Magnetic Resonance Imaging Sequence Data. Diagnostics (Basel) 2023.
33. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell 2020;42:318-27.
34. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018:839-847. doi: 10.1109/WACV.2018.00097.
35. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203-11.
36. Yin P, Zhi X, Sun C, Wang S, Liu X, Chen L, Hong N. Radiomics Models for the Preoperative Prediction of Pelvic and Sacral Tumor Types: A Single-Center Retrospective Study of 795 Cases. Front Oncol 2021;11:709659.
37. Yin P, Mao N, Zhao C, Wu J, Sun C, Chen L, Hong N. Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. Eur Radiol 2019;29:1841-7.
38. Yin P, Mao N, Zhao C, Wu J, Chen L, Hong N. A Triple-Classification Radiomics Model for the Differentiation of Primary Chordoma, Giant Cell Tumor, and Metastatic Tumor of Sacrum Based on T2-Weighted and Contrast-Enhanced T1-Weighted MRI. J Magn Reson Imaging 2019;49:752-9.
39. Sohsah GN, Ibrahimzada AR, Ayaz H, Cakmak A. Scalable classification of organisms into a taxonomy using hierarchical supervised learners. J Bioinform Comput Biol 2020;18:2050026.

Cite this article as: Yin P, Wang W, Wang S, Liu T, Sun C, Liu X, Chen L, Hong N. The potential for different computed tomography-based machine learning networks to automatically segment and differentiate pelvic and sacral osteosarcoma from Ewing's sarcoma. Quant Imaging Med Surg 2023;13(5):3174-3184. doi: 10.21037/qims-22-1042

