



A study on whether deep learning models based on CT images for bone density classification and prediction can be used for opportunistic osteoporosis screening

Tao Peng¹ · Xiaohui Zeng¹ · Yang Li² · Man Li² · Bingjie Pu¹ · Biao Zhi¹ · Yongqin Wang¹ · Haibo Qu³

Received: 18 March 2023 / Accepted: 25 August 2023 / Published online: 5 September 2023
© The Author(s) 2023

Abstract

Summary This study utilized deep learning to classify osteoporosis and predict bone density using opportunistic CT scans and independently tested the models on data from different hospitals and equipment. Results showed high accuracy and strong correlation with QCT results, showing promise for expanding osteoporosis screening and reducing unnecessary radiation and costs.

Purpose To explore the feasibility of using deep learning to establish a model for osteoporosis classification and bone density value prediction based on opportunistic CT scans and to verify its generalization and diagnostic ability using an independent test set.

Methods A total of 1219 cases of opportunistic CT scans were included in this study, with QCT results as the reference standard. The training set: test set: independent test set ratio was 703: 176: 340, and the independent test set data of 340 cases were from 3 different hospitals and 4 different CT scanners. The VB-Net structure automatic segmentation model was used to segment the trabecular bone, and DenseNet was used to establish a three-classification model and bone density value prediction regression model. The performance parameters of the models were calculated and evaluated.

Results The ROC curves showed that the mean AUCs of the three-category classification model for categorizing cases into “normal,” “osteopenia,” and “osteoporosis” for the training set, test set, and independent test set were 0.999, 0.970, and 0.933, respectively. The F1 score, accuracy, precision, recall, precision, and specificity of the test set were 0.903, 0.909, 0.899, 0.908, and 0.956, respectively, and those of the independent test set were 0.798, 0.815, 0.792, 0.81, and 0.899, respectively. The MAEs of the bone density prediction regression model in the training set, test set, and independent test set were 3.15, 6.303, and 10.257, respectively, and the RMSEs were 4.127, 8.561, and 13.507, respectively. The *R*-squared values were 0.991, 0.962, and 0.878, respectively. The Pearson correlation coefficients were 0.996, 0.981, and 0.94, respectively, and the *p* values were all < 0.001. The predicted values and bone density values were highly positively correlated, and there was a significant linear relationship.

Conclusion Using deep learning neural networks to process opportunistic CT scan images of the body can accurately predict bone density values and perform bone density three-classification diagnosis, which can reduce the radiation risk, economic consumption, and time consumption brought by specialized bone density measurement, expand the scope of osteoporosis screening, and have broad application prospects.

Keywords Artificial intelligence · Bone density classification · Bone density prediction · Convolutional neural network · Opportunistic CT scan · Osteoporosis screening

Abbreviations

DXA Dual-energy X-ray absorptiometry
QCT Quantitative computed tomography
BMD Bone mineral density
BMD_{QCT} Bone mineral density measured by QCT

ISCD International Society for Clinical Densitometry
ACR American College of Radiology
ROI Region of interest
MSE Mean square error
MAE Mean absolute error
RMSE Root mean square error
CNN Convolutional neural network

Extended author information available on the last page of the article

AUC	Area under the curve
LoA	Limits of agreement
LDCT _{chest}	Low-dose computer tomography of chest
CDCT _{chest}	Conventional dose computer tomography of chest
ABD	Abdomen
M	Male
F	Female
OP	Osteoporosis
OPE	Osteopenia
Nml	Normal
trn	Train set
tst	Test set
ITS	Independent test set

Introduction

Osteoporosis is closely related to population aging. Since 2010, population aging has accelerated in China. According to the data released by the Office of the Leading Group of the State Council for the National Population Census [1, 2], compared to 2010, the total population of China increased by 5.77% in 2020, with an increase of 5.73% and 5.82% for males and females, respectively. The population aged 50 and above increased by 44.10%, with an increase of 43.41% for males and 44.79% for females. The female population is 7.29 million more than the male population. Meanwhile, the population aged below 50 decreased by 7.23%. Based on this data, a clear trend is shown: the population aged 50 and above is rapidly increasing. According to recent studies, the prevalence of osteoporosis in the population aged 50 and above is 29.1%, which is equivalent to 49.3 million females and 10.9 million males who are at risk of osteoporosis and osteoporotic fractures [3]. By 2050, it is projected that there will be 5.99 million cases of osteoporotic fractures annually, costing \$25.43 billion. This represents a 2.7-fold increase since 2010 [4]. Over 20% of hip fracture patients will not survive beyond one year [5], while 20–60% of hip fracture patients will still require assistance to perform various household activities of daily living 1 year after the fracture [6].

Several studies have demonstrated that early intervention in patients with a high risk of fracture can lead to a significant reduction in fracture occurrence [7–9]. Therefore, it is especially critical to conduct osteoporosis screening and fracture risk assessment in patients. Bone density measurement plays a crucial role in the evaluation of osteoporosis and fracture risk. Bone density measurement is an internationally acknowledged diagnostic criterion for osteoporosis, with dual-energy X-ray absorptiometry (DXA) and quantitative computed tomography (QCT) being the two most extensively utilized methods in clinical settings. Compared to QCT, DXA is a two-dimensional method for

measuring bone mineral density (BMD), and its planar nature precludes direct assessment of trabecular bone in the spine. On the other hand, QCT offers a true 3D BMD measurement, and compared to DXA, it is less affected by severe degenerative changes, vascular calcifications, oral contrast agents, and body positions in the spine. As a result, QCT is being increasingly acknowledged as a more precise approach for quantitatively assessing osteoporosis [10]. However, there are also certain limitations of QCT in practical applications. The use of standardized software and strict calibration is required for QCT, and it cannot be applied to different CT scanners simultaneously, which means that complex post-processing and higher application costs are involved. Screening for osteoporosis is still not widely utilized despite its importance [11]. In the USA, less than 23% of individuals have undergone BMD evaluation using DXA as recommended [12]. Access to DXA is scarce in China, with only 0.46 DXA systems available per million inhabitants [13], and the availability of QCT systems is even lower. However, the number of CT scanners is much higher than that of DXA and QCT systems, with 18.2 scanners per million people in 2019 and 34 scanners per million people by the end of 2021 [14]. Numerous patients undergo CT scans every year in China for various medical reasons, and low-dose chest CT scans are utilized as a part of Healthy China 2030, a new long-term health strategy, for early screening of lung cancer. However, the majority of CT images obtained through these scans cannot be utilized for QCT bone density testing.

There has been a growing interest and utilization of artificial intelligence, specifically deep learning and machine learning, in the field of medical imaging in recent years [15–17]. This technology has the potential to provide automated and accurate tools for utilizing a large volume of CT images for osteoporosis screening, enabling early intervention to prevent fractures [18, 19]. This study aims to create a deep learning model that can be used for opportunistic osteoporosis screening. The model classifies the severity of osteoporosis and predicts bone density values, and its performance will be validated.

Materials and methods

Ethical approval

All procedures involving human participants were conducted as per the ethical standards of the institutional research committee and the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the Institutional Review Board and the ethics committee (IRB No.PJ2022-047-02) of the Affiliated Hospital of Chengdu University, and the need for individual patient

consent was waived because this was a retrospective study utilizing anonymized data.

Participants

In this study, we collected 1542 cases of images in 6 batches according to the registration order of QCT examination from four CT scanners. The examination methods of these case images included low-dose CT for lung cancer screening, routine-dose CT of the chest, CT of the lumbar vertebrae, and CT of the abdomen. Batches 1, 2, and 3 were from Chengdu University Affiliated Hospital (uCT 550, United Imaging Healthcare, Shanghai, China); batch 4 was from Chengdu University Affiliated Hospital (LightSpeed VCT, GE Healthcare, Waukesha, WI, USA); batch 5 was from West China Second Hospital of Sichuan University (Revolution CT, GE Healthcare, Milwaukee, WI, USA); and batch 6 was from Chengdu University School Hospital (uCT 780, United Imaging Healthcare, Shanghai, China). The images from batches 1, 2, 3, and 5 were used not only for other medical diagnosis but also directly for QCT bone density measurement and obtaining quantitative BMD (BMD_{QCT}) data at the same time. The CT images of batches 4 and 6 were obtained for other diagnostic purposes, but QCT bone density measurements were performed on the patients within 1 month before or after these CT examinations. To eliminate interference from other factors and obtain QCT measurement results as a reference standard with BMD_{QCT} , we excluded cases with a history of thoracic or lumbar spine surgery, vertebral compression fractures, and spinal tumors and cases that did not include the complete inferior margin of the second lumbar vertebra. Figure S1 illustrates the process of data selection and experimental workflow. The tube voltage used for all different CT examination procedures is set to 120 kV. For low-dose chest CT scans, the tube current is set to 50–70 mA, while for other body parts or conventional dose chest CT, the tube current is automatically adjusted.

QCT post-processing and BMD measurement

In this study, QCT measurements were used as the basis for diagnosis and prediction of bone density. According to the recommendations for using QCT to diagnose osteoporosis by the International Society for Clinical Densitometry (ISCD) and the American College of Radiology (ACR), the diagnostic criteria were set as follows: $BMD \geq 120 \text{ mg/cm}^3$ was considered normal, $BMD \leq 80 \text{ mg/cm}^3$ was considered osteoporosis, and BMD between 80 and 120 mg/cm^3 was considered osteopenia [20–22]. Prior to deep learning research, all CT images of the cases were subjected to bone density measurement using a specialized QCT post-processing workstation (Mindways QCT Pro, version 6.1, Austin, Texas, USA), and quality control analysis was performed

using the MINDWAYS Model 4 CT Calibration Phantom on the same day. Cancellous bone in the 12th thoracic vertebra, as well as the 1st and 2nd lumbar vertebrae, was measured using chest CT images, including low-dose chest CT [23]. Lumbar and abdominal CT images were used to measure cancellous bone in the first to third lumbar vertebrae. The oval region of interest (ROI) was delineated in the central region of the vertebral body, avoiding the perivertebral cortical bone and vein areas. This ROI should be as large as possible within the cancellous bone of the vertebral body. After measurement, the mean value of the measured cancellous bone density of the vertebral body was taken as the participant's bone density value.

Data preparation for deep learning

All data were segmented using a pre-trained automatic segmentation model based on the VB-Net architecture on the uAI Research portal (Ver.20220230, United Imaging Intelligence, Shanghai, China). uAI Research Portal (uRP) is a comprehensive medical image analysis software designed for scientific research. Its main objective is to provide a wide range of functionalities and tools that support various visualizations and advanced analysis techniques. These capabilities include automatic segmentation, registration, and classification, making it suitable for a variety of medical imaging modalities, diseases, and application [24]. The segmentation module in uRP can automatically delineate ROIs from single-modal and multimodal 2D/3D data. The VB-Net architecture is one of the partitioning architectures used for this purpose. The V-Net architecture proposed by Milletari et al. [25] was originally developed for prostate segmentation by training an end-to-end fully convolutional network on MRI. It has two distinct paths: the left contraction path extracts high-level context via convolutions and downsampling, while the right expanding path merges high-level context and detailed local information through skip connections, enabling accurate boundary localization. By using residual functions and skip connections, V-Net demonstrates superior segmentation accuracy compared to many classical CNNs. The VB-Net replaces conventional convolutional layers in V-Net with bottleneck structures. This architecture variation led us to name it VB-Net ("B" for bottleneck). The bottleneck structure has three convolutional layers. Performing spatial convolutions on feature maps with reduced channels provides two main benefits: (1) significantly smaller model size and (2) faster inference time. This study utilized VB-Net for precise spinal segmentation to enable subsequent bone density classification and prediction. In this study, we developed a three-category classification model for categorizing bone density into normal, osteopenia, and osteoporosis, as well as a regression model for bone density prediction, based on the DenseNet architecture. The DenseNet architecture,

proposed by Huang et al.²⁶, is a deep learning model widely used for computer vision tasks like image classification and object detection. It introduces unique inter-layer connectivity. Unlike CNNs where information flows sequentially, DenseNet establishes direct connections between layers in dense blocks. This connectivity enables direct gradient access for all layers, allowing efficient network-wide information flow. The DenseNet architecture comprises multiple dense blocks, transition layers, and a classification layer. In each dense block, a layer's inputs are concatenated with previous layers' outputs, forming dense connections that facilitate information and gradient flow, alleviating vanishing gradients, and enabling training of very deep models.

After automatic segmentation of vertebral bodies, the researchers performed erosion operations along the x , y , and z axes in six directions with a depth of 3 mm to obtain trabecular bone as the ROI. They calculated the ROI-related information, including the coordinates of the ROI center point, as well as the width, height, and depth of the ROI, and generated a list of original image paths, class labels, and ROI information pairs to prepare for subsequent deep learning network inference. The network used the original image as a single-channel input, sampled around the ROI region in the original image based on the ROI information, and performed various data augmentations such as rotation and translation to increase the diversity of training samples. The data was preprocessed by resampling, cropping, and data normalization based on the `crop_size`, `spacing`, and `crop_normalizers` parameters.

A bone density three-classification model and a bone density measurement regression model were constructed based on a merged dataset from batch 1, batch 2, and batch 3. The training set and test set were randomly divided in an 8:2 ratio. Batch 4, batch 5, batch 6, and combined batch (i.e., the merged batch 4, batch 5, and batch 6) were used as independent test sets.

Bone density diagnostic three-classification model

Models training, saving, and validation

Focal loss is a loss function commonly employed in machine learning and deep learning models, specifically in object detection tasks [27]. It addresses the issue of class imbalance where there is a significant disparity in sample numbers across different classes. Unlike standard loss functions like cross-entropy which disproportionately penalize misclassifications of minority classes and hinder learning under class imbalance, focal loss assigns greater emphasis to misclassified samples from the minority class. By introducing a modulating factor to cross-entropy loss, focal loss reduces the impact of easy examples during training, gives higher weights to misclassified samples, and focuses on those that

are harder to classify. Considering the imbalanced sample distribution in our study, we ultimately opted for focal loss as the loss function for our network. In this study, the pre-processed samples were input into the DenseNet network for training (Fig. 1). Focal loss was used to evaluate the loss during the training process. In iterative training, the network loss decreased gradually. The model was automatically saved every 20 iterations of training and tested on the test set samples to obtain prediction information for each test sample. Since the decrease of loss value during training only reflects the convergence of the model, it is unnecessary to reduce the loss to a specific value. The training was stopped when the network loss decreased to be adequately low and stable.

Classification performance evaluation

When evaluating the performance of a classification model, various performance metrics are calculated, including AUC, F1 score, recall, precision, specificity, and accuracy. Then, the optimal model is selected based on the comprehensive performance metrics and evaluated using an independent test set, and the ROC curve is plotted.

Bone density prediction regression model

Model training, saving, and validation

After completing image segmentation and ROI establishment according to the previous description, the preprocessed samples were fed into the DenseNet network for training (Fig. 1). The mean squared error (MSE) was used to evaluate the loss during the training process. During the iterative training process, the network's loss gradually decreased, and the model was automatically saved every 20 iterations of training. The saved model was then tested using the test set samples to obtain the predicted information for each test sample. The network training ended when the loss decreased to a sufficiently low level. Parameter table of neural network is shown in TableS4.

Regression model performance evaluation

Calculate the root mean squared error (RMSE) and mean absolute error (MAE) between the predicted bone density values and BMD_{QCT} for each model, select the optimal model, and evaluate it using the independent test sets. Calculate the Pearson correlation coefficient, and plot the Bland–Altman plot and correlation coefficient plot. MAE and root mean square error (RMSE) were calculated using Python V3.7.4, and Pearson correlation coefficient and p value were calculated using R language V4.1.2. The correlation plots and Bland–Altman plots were created using R language V4.1.2 and package `blandr` V0.5.1.

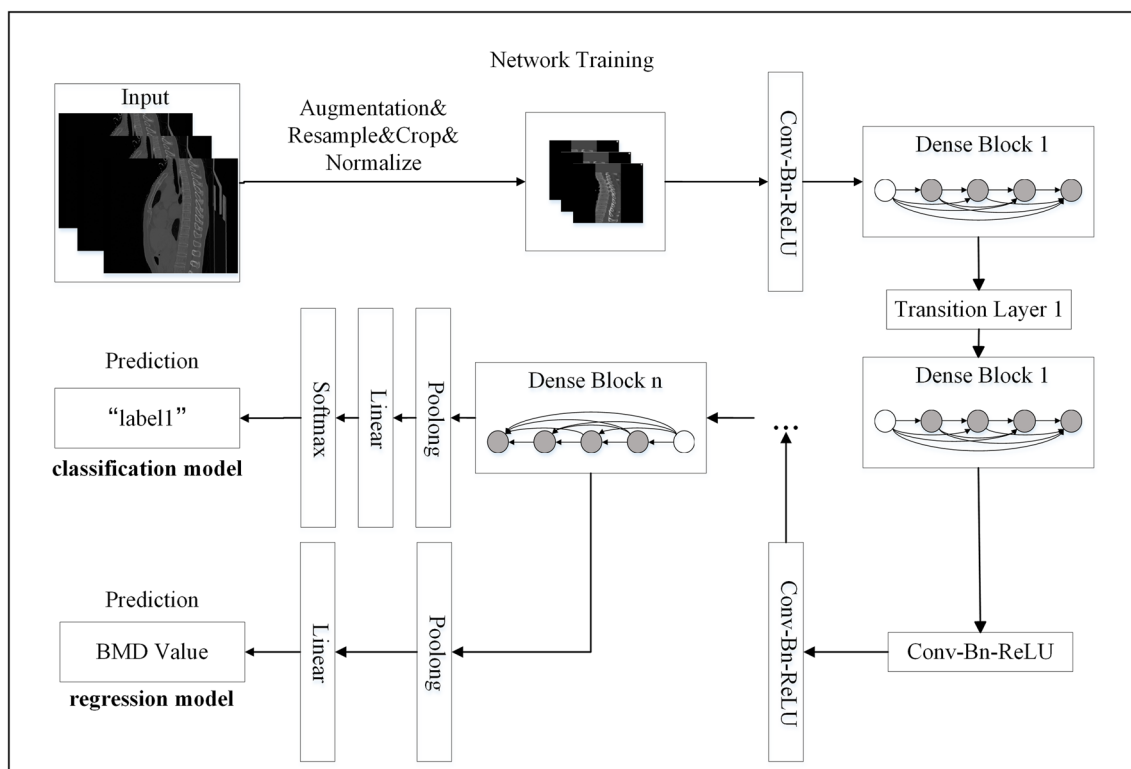


Fig. 1 Bone density classification and bone density value prediction network training flowchart

Results

Patients

A total of 1219 cases were included, with 199, 595, 85, 42, 20, 278, and 340 cases being enrolled in batch 1 to batch 6, as presented in Table S1 and Figure S1. The ratio of the training set, test set, and independent test set was 703:176:340, as illustrated in Table S2. The SPSSPRO online data analysis platform (<https://www.spsspro.com/>) was used to perform one-way ANOVA to test for significant differences in age among different batches. The result showed that $F = 4.225$, $p = 0.001$, indicating that there was a statistically significant difference in age among different batches. Using Pearson’s chi-square test, we analyzed whether there were significant differences in gender, scanning site, and QCT bone density classification among different batches. The results showed that the χ^2 values of these three variables were 68.446, 603.59, and 146.79, respectively, with p values of 0.000, indicating that the differences in these three variables among different batches were statistically significant. Further effect size analysis showed that the Crammer’s V values of gender, scanning site, and QCT bone density classification were 0.237, 0.406, and 0.245, respectively, indicating a moderate degree of difference among these three variables across different batches.

Evaluation of the three-class model for bone density diagnosis

The three-class classification model achieved an area under the curve (AUC) of 0.97 on the test set, with an F1 score of 0.903, accuracy of 0.909, precision of 0.899, recall of 0.908, and specificity of 0.956. Table 1 shows the performance metrics of the three-class model on the training and test sets, while Table 2 displays the confusion matrix for the model. Figure 2A displays the ROC curve for the three-class model on the test set, with the area under the curve (AUC) for the three classifications being 0.99 for normal bone density, 0.94 for osteopenia, and 0.98 for osteoporosis. The three-classification model achieved an AUC of 0.933 on the independent test set (combined batch), with an F1 score of 0.798, accuracy of 0.815, precision of 0.792, recall of 0.81, and specificity of 0.899. Table 1 shows the performance metrics of the three-class model on the independent test set, while Table 2 displays the confusion matrix for the model. Figure 2B displays the ROC curve for the three-class model on the independent test set, with the area under the curve (AUC) for the three classifications being 0.974 for normal bone density, 0.880 for osteopenia, and 0.945 for osteoporosis. Table S3 shows the details of QCT classification mismatched cases in three-class model predictions.

Table 1 Evaluation metrics for the three-classification model for bone density classification

	AUC (95% CI)	F1 score	Accuracy	Precision	Recall	Specificity
Training set	0.999 (0.999–1.000)	0.983	0.983	0.982	0.985	0.992
Test set	0.970 (0.949–0.990)	0.903	0.909	0.899	0.908	0.956
Batch 4	0.937 (0.842–1.000)	0.853	0.857	0.854	0.856	0.928
Batch 5	0.981 (0.918–1.000)	0.888	0.9	0.915	0.889	0.948
Batch 6	0.930 (0.899–0.959)	0.766	0.802	0.755	0.785	0.886
Independent test set (combined batch)	0.933 (0.906–0.960)	0.798	0.815	0.792	0.81	0.899

Combined batch: batches 4, 5, and 6

Table 2 Confusion matrix of the three-classification model

	OP _{trn}	OPE _{trn}	Nml _{trn}	Tot _{trn}	OP _{tst}	OPE _{tst}	Nml _{tst}	Tot _{tst}	OP _{ITS}	OPE _{ITS}	Nml _{ITS}	Tot _{ITS}
OP _{CNN}	189	4	0	193	44	6	0	50	159	24	0	183
OPE _{CNN}	0	221	7	228	2	43	6	51	18	70	6	94
Nml _{CNN}	0	1	281	282	0	2	73	75	0	15	48	63
Tot _{CNN}	189	226	288	703	46	51	79	176	177	109	54	340

OP osteoporosis, OPE osteopenia, Nml normal

Subscript: *trn* train set, *tst* test set, *ITS* independent test set, *CNN* convolutional neural network

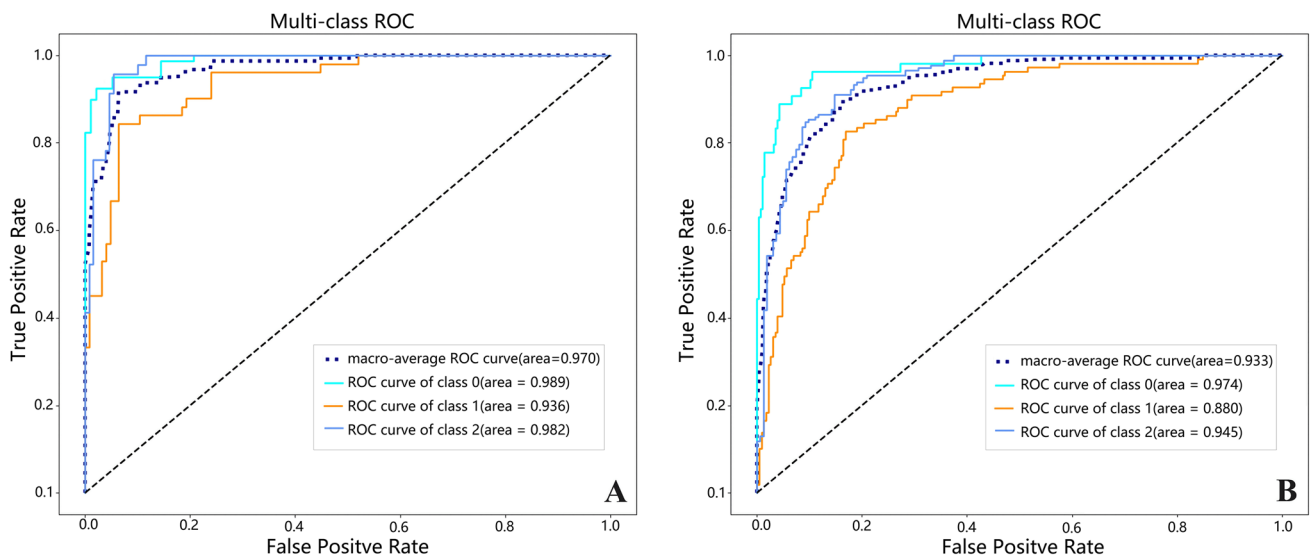


Fig. 2 The ROC curve of the three-classification model for the test set (A) and the independent test set (B), where the light blue line represents the ROC curve for diagnosing class 0 cases, the orange line represents the ROC curve for diagnosing class 1 cases, and the dark blue line represents the ROC curve for diagnosing class 2 cases. The

black dashed line represents the macro-average ROC curve. Class 0 represents normal bone density diagnosed by QCT, class 1 represents osteopenia diagnosed by QCT, and class 2 represents osteoporosis diagnosed by QCT

Performance evaluation of bone density prediction model

The deep learning network was trained for 2200 iterations with mean square error as the loss function. After over 1200 iterations, the loss became stable. The MAEs of the bone density prediction regression model on the training set, test

set, and independent test set are 3.15, 6.303, and 10.257, respectively. The RMSEs for the same sets are 4.127, 8.561, and 13.507, respectively. The *R*-squared values are 0.991, 0.962, and 0.878, respectively. The Pearson correlation coefficients (ρ) are 0.996, 0.981, and 0.94, with *p* values < 0.001, indicating a very strong positive correlation and a significant linear relationship between the bone density values predicted

by the regression model and BMD_{QCT} , as shown in Table 3. Figure 3 shows the Bland–Altman plot and correlation coefficient plot of the bone density values predicted by the deep learning model relative to BMD_{QCT} .

Discussion

In this study, we used opportunistic CT scan data from the chest, abdomen, and spine and applied automatic segmentation to obtain the ROI for trabecular bone. We developed a DenseNet-based three-classification model for bone density and a regression model for predicting bone density values. The experimental results demonstrated that the three-classification model accurately classified the trabecular bone density of the spine, and the bone density value prediction regression model accurately predicted the trabecular bone density values of the spine. The predicted values were found to be highly consistent with the reference values obtained from QCT, indicating a strong positive correlation and significant linear relationship between them.

Analysis of results from the three-classification model

In this study, we developed a three-classification model to classify bone density in opportunistic CT scans, aiming to classify different patients into normal, osteopenia, and osteoporosis categories. In the training, test, and independent test sets, the model achieved macro-average ROC curve AUCs of 0.99, 0.97, and 0.93, respectively. In the independent test set, the AUCs for classifying cases with QCT results of normal bone density, osteopenia, and osteoporosis were 0.974, 0.880, and 0.945, respectively. In the test set, the corresponding AUCs were 0.989, 0.936, and 0.982 (Fig. 2).

According to the classification criteria set in the study, there were a total of 34 cases where the reference value (BMD_{QCT}) was osteopenia but predicted as osteoporosis, including four in the training set, six in the test set, and

24 in the independent testing set. Among these cases, 27 (79.41%) had a BMD_{QCT} difference from the 80 mg/cm³ boundary of less than 10, and even some cases had a difference of less than 1. There were 18 cases in total, including one in the training set, two in the test set, and 15 in the independent testing set, where the BMD_{QCT} was osteopenia but predicted as normal. Among these cases, 11 (61.11%) had a BMD_{QCT} difference from the 120 mg/cm³ boundary of less than 10. There were 19 cases in total, including seven in the training set, six in the test set, and six in the independent testing set, where the BMD_{QCT} was normal but predicted as osteopenia. Among these cases, 16 (84.21%) had a BMD_{QCT} difference from the 120 mg/cm³ boundary of less than 10. There were 20 cases in total, including two in the test set and 18 in the independent testing set, where the BMD_{QCT} was osteoporosis but predicted as osteopenia. Among these cases, 14 (77.78%) had a BMD_{QCT} difference from the 80 mg/cm³ boundary of less than 10. No cases were predicted as osteoporosis when the BMD_{QCT} was normal, and no cases were predicted as normal when the BMD_{QCT} was osteoporosis. As BMD values differ between different vertebrae in QCT measurements, the study set the average value as the reference value, so the predicted value is meaningful when it is close to the BMD_{QCT} . BMD value prediction can more accurately reflect bone density and avoid the classification crossover phenomenon caused by the predicted value being close to the boundary in classification.

According to the ROC curves of the three-classification model, the AUCs for diagnosing cases with normal bone density and osteoporosis in the test set were similar, at 0.99 and 0.98, respectively, and higher than the AUC (0.94) for diagnosing cases with osteopenia as the reference value. In the independent testing set, the AUCs for diagnosing cases with normal bone density and osteoporosis were 0.974 and 0.945, respectively, also higher than the AUC (0.88) for diagnosing cases with osteopenia, which has two threshold values of 80 mg/cm³ and 120 mg/cm³. The higher likelihood of cases close to the threshold values in cases with osteopenia makes it easier to make classification errors, compared to normal bone density and osteoporosis, which only have one threshold value. Therefore, the classification model has a stronger diagnostic ability for normal bone density and osteoporosis.

In the study by Rastegar et al., they used radiomics to classify bone mineral densitometry images of the hip joint and lumbar spine, achieving an AUC of 0.50–0.78. However, this method had lower classification ability than the deep learning combined with CT scanning method used in this study [28]. Yasaka et al. also used a convolutional neural network (CNN) model to predict BMD from lumbar spine CT images and compared it with DXA. The AUC for diagnosing osteoporosis was 0.965 and 0.970 in the internal and

Table 3 Performance metrics for the bone density prediction regression model

Set	MAE	RMSE	R-squared	ρ	p value
Training set	3.15	4.127	0.991	0.996	<0.001
Test set	6.303	8.561	0.962	0.981	<0.001
Batch 4	11.44	15.512	0.892	0.948	<0.001
Batch 5	15.169	18.508	0.85	0.943	<0.001
Batch 6	9.725	12.733	0.85	0.931	<0.001
Independent test set (combined batch)	10.257	13.507	0.878	0.94	<0.001

Combined batch: batches 4, 5, and 6

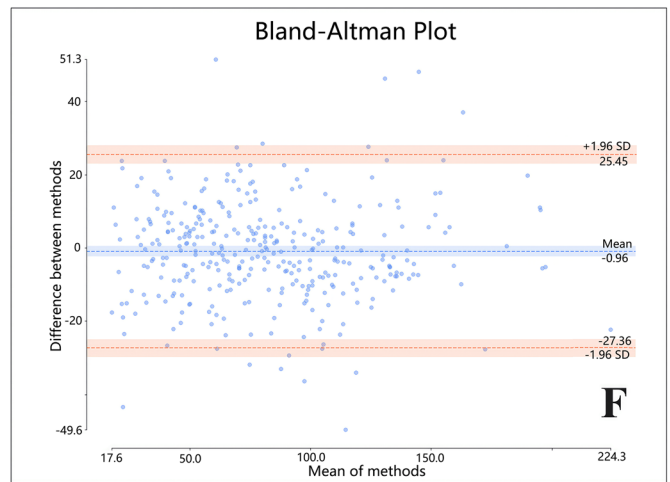
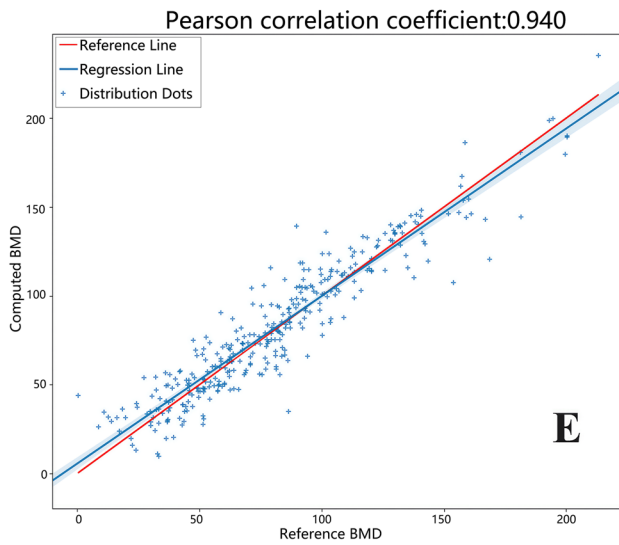
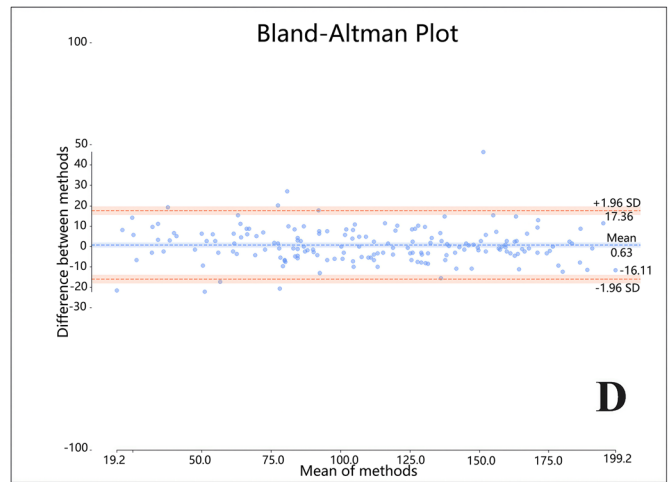
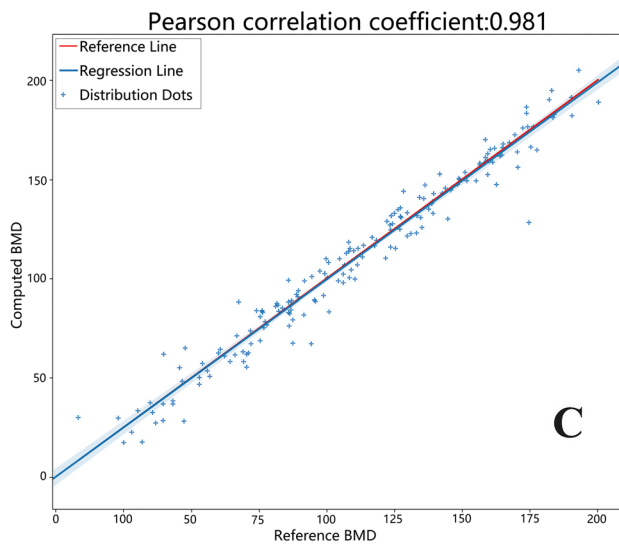
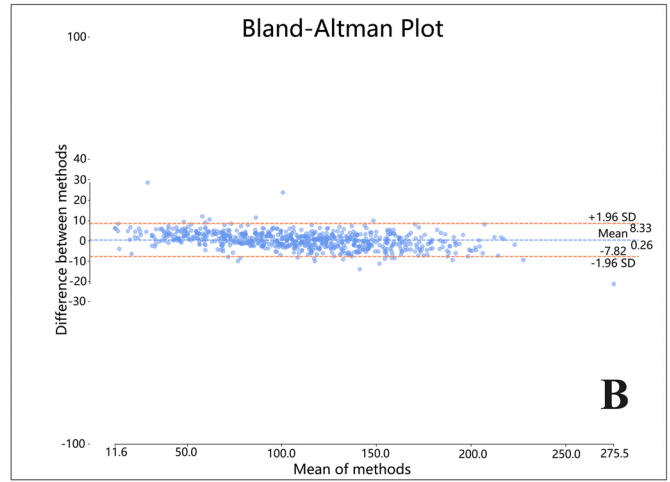
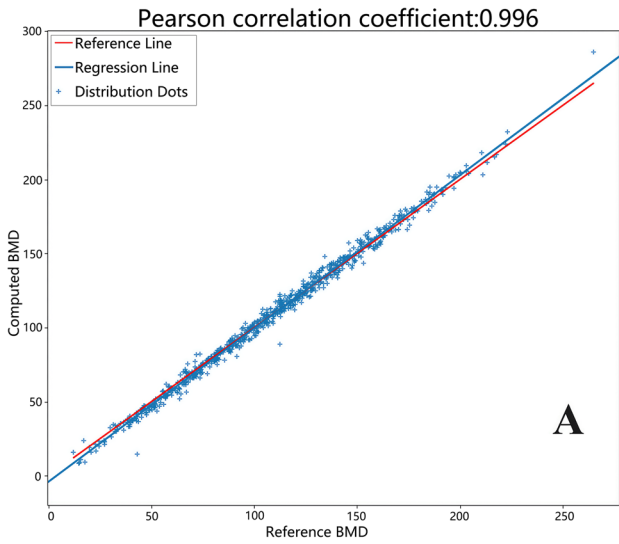


Fig. 3 Correlation plot and Bland–Altman plot of train set (**B**). **A** A scatter plot of BMD_{QCT} and predicted BMD in training set (**A**, **B**), test set (**C**, **D**), and independent test set (**E**, **F**). The horizontal axis shows BMD_{QCT} . The vertical axis shows predicted BMD. Correlation plot with confidence interval set at 95% (shadowed area). Red line represents reference line; blue line is regression line. Blue crosses are distribution dots. The Pearson's correlation coefficient for the training set is 0.996; for the test set, it is 0.981; and for the independent test set, it is 0.94. The R -squared value for the training set is 0.991; for the test set, it is 0.962; and for the independent test set, it is 0.878. The p value is less than 0.001. **B** In the Bland–Altman plot, the horizontal axis shows the average of the results for each sample measured by the two methods, and the vertical axis shows the difference between the two methods. The limits of agreement (LoA) are shown as dashed orange lines with 95% confidence intervals (light orange areas), while the bias is shown as a dashed blue line with a 95% confidence interval (light blue area). The degree of agreement between the two measurements is reflected by the tightness of the distribution of points around the central mean line. Outliers are observations that lie above and below the light red bands, respectively. In the Bland–Altman plot of the training set, the mean difference is 0.26, the higher LoA is 8.33, the lower LoA is -7.82 , and 0.036% (25/703) of the points are outside the 95% LoA. For the test set, the mean difference is 0.629, the higher LoA is 17.41, the lower LoA is -16.15 , and 0.051% (9/176) scatters are outside the 95% LoA. For the independent test set, the mean difference is -0.956 , the higher LoA is 25.49, the lower LoA is -27.40 , and 0.05% (17/340) points are outside the 95% LoA

external test sets, respectively, which is similar to our study. However, unlike our study, the AUC in the independent testing set was lower than in the test set [29].

In this study, we established a regression model to predict BMD values, and the MAEs on the test set and independent test set were 6.303 and 10.257, respectively, while the RMSEs were 8.561 and 13.507, respectively. MAE is the average of the absolute errors, and RMSE is the square root of the mean squared difference between the predicted and reference values. In this study, the comparison of MAE and RMSE between the test set and independent test set showed no significant difference, indicating that the differences between the predicted BMD values and reference values were stable, and there were no large outliers. The R^2 values in the test set and independent test set were 0.962 and 0.878, respectively, indicating a good fit of the regression equation. In addition, the Pearson's correlation coefficients on the test set and independent test set were 0.981 and 0.94, respectively, with $p < 0.001$, indicating a strong linear correlation and high similarity between the predicted BMD values and BMD_{QCT} . Fang et al. used a regression model based on DenseNet to predict BMD values for some cases in three testing cohorts, with R^2 ranging from 0.780 to 0.948, similar to this study [30]. Pan et al. also predicted BMD values using a deep learning model, with R^2 ranging from 0.964 to 0.968. Bland–Altman analysis also showed good consistency between the predicted BMD values and QCT reference values, consistent with the results of this study [31].

As an important medical imaging examination, CT scan is an important diagnostic tool for many diseases

[32], especially after experiencing the COVID-19 pandemic for 3 years, people pay more attention to health issues, and the number of people visiting hospitals or health centers for examination has increased significantly, with many undergoing CT scans, but many of them have not undergone DXA or QCT bone density examinations [33]. Since there is a significant overlap between patients undergoing CT scans for other diseases and those with osteoporosis risk factors and most of the current body CT scans are volumetric scans, there is an opportunity to use these CT images for bone density assessment [11]. If we have a method to directly evaluate bone density based on these CT data, we can significantly expand the scope of osteoporosis screening, enable more people to receive early intervention, and prevent fractures, thus achieving the goal of improving quality of life, reducing mortality, and lowering costs [16, 31]. This study demonstrates that accurate prediction of bone density values and diagnosis of bone density abnormalities can be achieved through deep learning neural networks, reducing the economic and time costs of additional DXA bone density testing and even eliminating the need for additional radiation exposure for patients [34]. Moreover, it can accurately measure trabecular bone density. Compared with QCT, the biggest advantage of the bone density prediction method in this study is that it is not limited to a single device, does not require planning and calibration of QCT scans before the examination, and can use retrospective CT data for BMD measurement without additional manual post-processing, significantly expanding the scope of osteoporosis screening and improving efficiency [35]. It can also achieve diagnostic prediction results comparable to QCT. Studies have shown that there are significant differences in the access and quality of DXA services globally, and the deep learning screening method based on opportunistic CT scans is expected to reduce this variability [36].

Although DXA is still one of the gold standard methods to categorize patients into “normal,” “osteopenia,” or “osteoporosis” according to the WHO classification, QCT can provide a reliable assessment standard as it assesses bone density based on three-dimensional volumetric data and can avoid interference from factors such as aortic calcification and osteoporosis [37]. Therefore, we adopted QCT as the reference standard for BMD measurement in this study. In addition, using QCT as the reference standard can provide more research samples for our study, as the number of subjects undergoing QCT far exceeds those who have undergone both CT and DXA examinations. In our study, we excluded cases that would significantly affect vertebral density beforehand. For the deep learning modeling and test in the regions of interest, we used all scanned vertebrae, while the ROI for chest CT cases measured by QCT was the L1-2 vertebrae, and the ROI for abdominal and lumbar CT cases was the L1-3 vertebrae. The average BMD values of each vertebra

measured by QCT were used as the reference values. According to the results of the final regression model, although the predicted and reference BMD values were highly consistent, there were still some small differences between them. Considering the measurement range, the BMD values predicted by the regression model may better reflect the overall spine bone density situation.

Limitations and prospects

(1) The compressed vertebrae were not automatically excluded during the automatic segmentation process, and the method used was to manually exclude relevant cases. (2) The method used to select vertebral bodies and subchondral bone through automatic segmentation and erosion did not specifically remove small influencing factors such as bone islands. (3) The training set was from a single device. (4) Bone density measurements were concentrated on the spine, and the relationship between hip joint bone density measurements and bone density has not yet been studied. (5) The relationship between the predicted bone density values and the risk of fractures has not yet been studied. (6) In future studies, an automatic segmentation model should be established to remove vertebrae and factors affecting vertebral bone density in images that are not suitable, such as compression fractures, bone islands, vertebral hemangiomas, and internal fixation, and retain vertebrae that meet the criteria to expand the model's applicability range.

Conclusion

In this study, we utilized opportunistic CT scan data from different medical institutions and CT devices for chest, abdomen, and spine to establish and validate a deep learning convolutional neural network-based bone density three-classification model and bone density value prediction regression model. We found that these models can accurately predict bone density values and perform bone density three-classification diagnosis, reducing the radiation risk, economic costs, and time consumption associated with specialized bone density measurement, expanding the population scope of osteoporosis screening, and helping to improve the screening efficiency and accuracy of osteoporosis, with broad application prospects in reducing the incidence of osteoporotic fractures, improving patient quality of life, and reducing mortality rates.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00198-023-06900-w>.

Funding The study was funded by the Health Commission of Chengdu (Sichuan, China) (grants 2021036, 2021045). Tao Peng has received research grants from Health Commission of Chengdu (grant 2021036). Yongqin Wang has received research grants from Health Commission of Chengdu (grant 2021045).

Declarations

Conflict of interest None.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.


References

1. Sixth National Population Census of the People's Republic of China. Beijing, China; 2010. Available from: <https://www.Stats.Gov.Cn/Tjsj/Pcsj/Rkpc/6rp/Indexch.Htm>. Chinese
2. Seventh National Population Census of the People's Republic of China. Beijing, China; 2020. Available from: <https://www.Stats.Gov.Cn/Tjsj/Pcsj/Rkpc/7rp/Indexch.Htm>. Chinese
3. Zeng Q, Li N, Wang Q et al (2019) The prevalence of osteoporosis in China, a nationwide, multicenter DXA survey. *J Bone Miner Res* 34(10):1789–1797. <https://doi.org/10.1002/jbmr.3757>
4. Si L, Winzenberg TM, Jiang Q, Chen M, Palmer AJ (2015) Projection of osteoporosis-related fractures and costs in China: 2010–2050. *Osteoporos Int* 26(7):1929–1937. <https://doi.org/10.1007/s00198-015-3093-2>
5. Brauer CA (2009) Incidence and mortality of hip fractures in the United States. *JAMA* 302(14):1573. <https://doi.org/10.1001/jama.2009.1462>
6. Dyer SM, Crotty M, Fairhall N et al (2016) A critical review of the long-term disability outcomes following hip fracture. *BMC Geriatr* 16:158. <https://doi.org/10.1186/s12877-016-0332-0>
7. Rubin KH, Rothmann MJ, Holmberg T et al (2018) Effectiveness of a two-step population-based osteoporosis screening program using FRAX: the randomized Risk-stratified Osteoporosis Strategy Evaluation (ROSE) study. *Osteoporos Int* 29(3):567–578. <https://doi.org/10.1007/s00198-017-4326-3>
8. McCloskey E, Johansson H, Harvey NC et al (2018) Management of patients with high baseline hip fracture risk by FRAX reduces hip fractures—a post hoc analysis of the SCOOP study: reduced hip fractures in high-risk patients with osteoporosis management. *J Bone Miner Res* 33(6):1020–1026. <https://doi.org/10.1002/jbmr.3411>
9. Turner DA, Khioe RFS, Shepstone L et al (2018) The cost-effectiveness of screening in the community to reduce osteoporotic fractures in older women in the UK: economic evaluation of the SCOOP study: cost-effectiveness of community screening for fracture risk. *J Bone Miner Res* 33(5):845–851. <https://doi.org/10.1002/jbmr.3381>
10. Wu Y, Guo Z, Fu X et al (2019) The study protocol for the China Health Big Data (China Biobank) project. *Quant Imaging Med Surg*. 9(6):1095–1102. <https://doi.org/10.21037/qims.2019.06.16>
11. Brett AD, Brown JK (2015) Quantitative computed tomography and opportunistic bone density screening by dual use of computed tomography scans. *J Orthop Transl* 3(4):178–184. <https://doi.org/10.1016/j.jot.2015.08.006>
12. Cosman F, de Beur SJ, LeBoff MS et al (2014) Clinician's guide to prevention and treatment of osteoporosis. *Osteoporos Int* 25(10):2359–2381. <https://doi.org/10.1007/s00198-014-2794-2>

13. Cheng X, Zhao K, Zha X et al (2021) Opportunistic screening using low-dose CT and the prevalence of osteoporosis in China: a nationwide, multicenter study. *J Bone Miner Res* 36(3):427–435. <https://doi.org/10.1002/jbmr.4187>
14. Tencent. CT industry perspective in China: the high-end wins the world. [Press release]. Shenzhen, China: Tencent; 2022 Jan 27 [cited 2023 Feb 18]. Available from: <https://new.qq.com/rain/a/20220127A01L8Z00.Chinese>
15. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018) Artificial intelligence in radiology. *Nat Rev Cancer* 18(8):500–510. <https://doi.org/10.1038/s41568-018-0016-5>
16. Boutin RD, Lenchik L (2020) Value-added opportunistic CT: insights into osteoporosis and sarcopenia. *Am J Roentgenol* 215(3):582–594. <https://doi.org/10.2214/AJR.20.22874>
17. Navarro F, Shit S, Ezhov I, et al. (2019) Shape-aware complementary-task learning for multi-organ segmentation. In: Suk HI, Liu M, Yan P, Lian C, eds. *Machine Learning in Medical Imaging*. Vol 11861. Lecture Notes in Computer Science. Springer International Publishing; 2019:620–627. https://doi.org/10.1007/978-3-030-32692-0_71
18. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E (2008) FRAX™ and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 19(4):385–397. <https://doi.org/10.1007/s00198-007-0543-5>
19. Huntjens KMB, van Geel TACM, van den Bergh JPW et al (2014) Fracture liaison service: impact on subsequent nonvertebral fracture incidence and mortality. *J Bone Jt Surg*. 96(4):e29. <https://doi.org/10.2106/JBJS.L.00223>
20. Löffler MT, Sollmann N, Mei K et al (2020) X-ray-based quantitative osteoporosis imaging at the spine. *Osteoporos Int* 31(2):233–250. <https://doi.org/10.1007/s00198-019-05212-2>
21. American College of Radiology. ACR–SPR–SSR practice parameter for the performance of musculoskeletal quantitative computed tomography (QCT). Available via <https://www.acr.org/-/media/ACR/Files/Practice-Parameters/qct.pdf?la=en>. Accessed 11 Mar 2023
22. Engelke K, Adams JE, Armbricht G et al (2008) Clinical use of quantitative computed tomography and peripheral quantitative computed tomography in the management of osteoporosis in adults: The 2007 ISCD Official Positions. *J Clin Densitom* 11(1):123–162. <https://doi.org/10.1016/j.jocd.2007.12.010>
23. Salzmann SN, Shirahata T, Yang J et al (2019) Regional bone mineral density differences measured by quantitative computed tomography: does the standard clinically used L1–L2 average correlate with the entire lumbosacral spine? *Spine J* 19(4):695–702. <https://doi.org/10.1016/j.spinee.2018.10.007>
24. Wu J, Xia Y, Wang X et al (2023) uRP: An integrated research platform for one-stop analysis of medical images. *Front Radiol* 3:1153784. <https://doi.org/10.3389/fradi.2023.1153784>
25. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016:565–571. <https://doi.org/10.1109/3DV.2016.79>
26. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2261–2269. <https://doi.org/10.1109/CVPR.2017.243>
27. Lin TY, Goyal P, Girshick R, He K, Dollár P. (2017) Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE; 2017:2999–3007. <https://doi.org/10.1109/ICCV.2017.324>
28. Rastegar S, Vaziri M, Qasempour Y et al (2020) Radiomics for classification of bone mineral loss: a machine learning study. *Diagn Interv Imaging* 101(9):599–610. <https://doi.org/10.1016/j.diii.2020.01.008>
29. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O (2020) Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. *Eur Radiol* 30(6):3549–3557. <https://doi.org/10.1007/s00330-020-06677-0>
30. Fang Y, Li W, Chen X et al (2021) Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. *Eur Radiol* 31(4):1831–1842. <https://doi.org/10.1007/s00330-020-07312-8>
31. Pan Y, Shi D, Wang H et al (2020) Automatic opportunistic osteoporosis screening using low-dose chest computed tomography scans obtained for lung cancer screening. *Eur Radiol* 30(7):4107–4116. <https://doi.org/10.1007/s00330-020-06679-y>
32. Radwan R, Tang A, Beasley W (2018) Computed tomography as a first-line investigation for elderly patients admitted to a surgical assessment unit. *Ann R Coll Surg Engl* 100(4):285–289. <https://doi.org/10.1308/rcsann.2017.0231>
33. Pickhardt PJ, Lee SJ, Liu J et al (2019) Population-based opportunistic osteoporosis screening: validation of a fully automated CT tool for assessing longitudinal BMD changes. *Br J Radiol* 92(1094):20180726. <https://doi.org/10.1259/bjr.20180726>
34. Jang S, Graffy PM, Ziemlewicz TJ, Lee SJ, Summers RM, Pickhardt PJ (2019) Opportunistic osteoporosis screening at routine abdominal and thoracic CT: normative L1 trabecular attenuation values in more than 20 000 adults. *Radiology* 291(2):360–367. <https://doi.org/10.1148/radiol.2019181648>
35. Pickhardt P, Bodeen G, Brett A, Brown JK, Binkley N (2013) Comparison of lunar DXA and QCT at the femoral neck using asynchronous calibration of CT colonography exams. *J Clin Densitom* 16(3):273–274. <https://doi.org/10.1016/j.jocd.2013.05.037>
36. On behalf of the International Society for Clinical Densitometry (ISCD) and the International Osteoporosis Foundation (IOF), Clynes MA, Westbury LD, et al. Bone densitometry worldwide: a global survey by the ISCD and IOF. *Osteoporos Int*. 2020;31(9):1779–1786. <https://doi.org/10.1007/s00198-020-05435-8>
37. Löffler MT, Jacob A, Valentinitz A et al (2019) Improved prediction of incident vertebral fractures using opportunistic QCT compared to DXA. *Eur Radiol* 29(9):4980–4989. <https://doi.org/10.1007/s00330-019-06018-w>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Tao Peng¹  · Xiaohui Zeng¹ · Yang Li² · Man Li² · Bingjie Pu¹ · Biao Zhi¹ · Yongqin Wang¹ · Haibo Qu³

✉ Tao Peng
pengtao919@163.com

Xiaohui Zeng
887132@qq.com

Yang Li
yang.li20@uui-ai.com

Man Li
1209091163@qq.com

Bingjie Pu
1347002983@qq.com

Biao Zhi
280210702@qq.com

Yongqin Wang
50608813@qq.com

Haibo Qu
windowsqhb@126.com

¹ Department of Radiology, Affiliated Hospital of Chengdu University, 82 2Nd N Section of Second Ring Rd, Chengdu 610081, Sichuan Province, China

² Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd, Shanghai 200232, China

³ Department of Radiology, West China Second University Hospital of Sichuan University, Chengdu 610041, Sichuan Province, China