Bone 140 (2020) 115561

Contents lists available at ScienceDirect

Bone

journal homepage: www.elsevier.com/locate/bone

Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study

Bin Zhang[a,b,1], Keyan Yu[c,1], Zhenyuan Ning[d,1], Ke Wang[d,1], Yuhao Dong[e], Xian Liu[f], Shuxue Liu[g], Jian Wang[c], Cuiling Zhu[c], Qinqin Yu[c], Yuwen Duan[c], Siying Lv[c], Xintao Zhang[c], Yanjun Chen[c], Xiaojia Wang[h], Jie Shen[i], Jia Peng[j], Qiuying Chen[a,b], Yu Zhang[d,*], Xiaodong Zhang[c,**], Shuixing Zhang[a,***]

[a] Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, PR China
[b] Jinan University, Guangzhou, Guangdong, PR China
[c] Department of Medical Imaging, The Third Affiliated Hospital of Southern Medical University (Academy of Orthopedics Guangdong Province), Guangzhou, PR China
[d] School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, PR China
[e] Department of Catheterization Lab, Guangdong Cardiovascular Institute, Guangdong Provincial Key Laboratory of South China Structural Heart Disease, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, PR China
[f] Department of Radiology, Guangdong Provincial Hospital of Chinese Medicine, Guangzhou, Guangdong, PR China
[g] The Affiliated Zhongshan Hospital of Traditional Chinese Medicine University of Guangzhou, Guangdong, PR China
[h] Bone mineral density test room, Health Management Centre, The Third Affiliated Hospital of Southern Medical University (Academy of Orthopedics Guangdong Province), Guangzhou, PR China
[i] Department of endocrinology, The Third Affiliated Hospital of Southern Medical University (Academy of Orthopedics Guangdong Province), Guangzhou, PR China
[j] Department of computed tomography, The Affiliated Zhongshan City Hospital of Sun Yat-sen University, PR China

ARTICLE INFO

Keywords:
Osteoporosis
Postmenopausal women
Bone mineral density
Dual-energy X-ray absorptiometry
Deep learning
Lumbar spine X-rays

ABSTRACT

Osteoporosis is a prevalent but underdiagnosed condition. As compared to dual-energy X-ray absorptiometry (DXA) measures, we aimed to develop a deep convolutional neural network (DCNN) model to classify osteopenia and osteoporosis with the use of lumbar spine X-ray images. Herein, we developed the DCNN models based on the training dataset, which comprising 1616 lumbar spine X-ray images from 808 postmenopausal women (aged 50 to 92 years). DXA-derived bone mineral density (BMD) measures were used as the reference standard. We categorized patients into three groups according to DXA BMD T-score: normal ($T \geq -1.0$), osteopenia ($-2.5 < T < -1.0$), and osteoporosis ($T \leq -2.5$). T-scores were calculated by using the BMD dataset of young Chinese female aged 20–40 years as a reference. A 3-class DCNN model was trained to classify normal BMD, osteoporosis, and osteopenia. Model performance was tested in a validation dataset (204 images from 102 patients) and two test datasets (396 images from 198 patients and 348 images from 147 patients respectively). Model performance was assessed by the receiver operating characteristic (ROC) curve analysis. The results showed that in the test dataset 1, the model diagnosing osteoporosis achieved an AUC of 0.767 (95% confidence interval [CI]: 0.701–0.824) with sensitivity of 73.7% (95% CI: 62.3–83.1), the model diagnosing osteopenia achieved an AUC of 0.787 (95% CI: 0.723–0.842) with sensitivity of 81.8% (95% CI: 67.3–91.8); In the test dataset 2, the model diagnosing osteoporosis yielded an AUC of 0.726 (95% CI: 0.646–0.796) with sensitivity of 68.4% (95% CI: 54.8–80.1), the model diagnosing osteopenia yielded an AUC of 0.810 (95% CI, 0.737–0.870) with sensitivity of 85.3% (95% CI, 68.9–95.0). Accordingly, a deep learning diagnostic network may have the potential in screening osteoporosis and osteopenia based on lumbar spine radiographs. However, further studies are necessary to verify and improve the diagnostic performance of DCNN models.

* Correspondence to: Yu Zhang, No. 1023 Shatai Road, Baiyun District, Guangzhou, Guangdong 510515, PR China.
** Correspondence to: Xiaodong Zhang, No. 183 Zhongshan Road, Tianhe District, Guangzhou 510630, PR China.
*** Correspondence to: Shuixing Zhang, No. 613 Huangpu West Road, Tianhe District, Guangzhou, Guangdong 510627, PR China.
E-mail addresses: yuzhang@smu.edu.cn (Y. Zhang), ddautumn@126.com (X. Zhang), shui7515@126.com (S. Zhang).
[1] Bin Zhang, Keyan Yu, Zhenyuan Ning, and Ke Wang contributed equally to this work.

https://doi.org/10.1016/j.bone.2020.115561
Received 14 November 2019; Received in revised form 16 July 2020; Accepted 23 July 2020
Available online 28 July 2020

1. Introduction

Osteoporosis and osteoporotic fractures have become global health issues of major concern with the growth in the aging population [1]. By 2020, approximately 12.3 million individuals in the United States older than 50 years are expected to have osteoporosis [2]. One in three women aged over 50 years will have an osteoporosis-related fracture [3]. As a precursor of osteoporosis, osteopenia also deserves attention because most fractures in postmenopausal women occurred in those with osteopenia [4,5]. Hence, screening of osteoporosis and osteopenia is clinically desirable for fracture prevention. The US Preventive Services Task Force (USPSTF) recommends that women aged ≥65 years should be routinely screened [6].

Central dual-energy X-ray absorptiometry (DXA) is globally accepted as the reference standard for diagnosing osteoporosis and osteopenia [7]. However, the application of DXA is limited by its low availability, which typically requires patients to travel to a referral centre [8]. Other barriers to DXA screening include knowledge deficits and declining financial incentives for screening. As a result, nearly half of female Medicare beneficiaries in the United States do not undergo DXA testing [9], and certain high-risk populations have screening rates of < 10% [10]; while in China, only 4.3% women aged ≥50 years have undergone testing, particularly in rural areas, the rate is only 1.9% [11]. The measurement of DXA assumes the presence of only bones and muscles, which would be inevitably influenced by fat [12]. In addition, as a two-dimensional projection technique, it cannot fully consider bone geometry, size and microstructure [12]. Accordingly, DXA is underutilized, and osteoporosis remains underdiagnosed. Safe and cost-effective alternatives to improve these conditions are needed. Conventional X-ray devices are widely available in almost any hospital worldwide, which carry potentially useful information about BMD. Retrieval of BMD data available on lumbar spine X-ray scans ordered for other indications requires no additional cost, patient time, or radiation exposure, and these data can be retrospectively acquired. It could, therefore, expand population screening effort for osteoporosis. However, it is a challenging task to assess BMD on lumbar spine X-ray images by inspection.

Recently, a deep learning technique, known as the deep convolutional neural network (DCNN), has gained significant ground in the field of computer vision. Deep learning takes raw image pixels and corresponding class labels from medical imaging data as inputs and automatically learns feature representation with multiple levels of abstraction [13]. Continuous improvements of DCNN architectures coupled with a rapid increase in hardware computational power have enabled DCNN to achieve human-level performance in layer tasks, such as facial recognition, game playing, and natural language processing [14]. Numerous early studies have also shown the promising results of DCNN used in a variety of medical imaging, including radiology [15–18], pathology [19,20], dermatology [21,22], and ophthalmology [23,24].

In this study, we aimed to determine the feasibility and performance of DCNN models based on lumbar spine X-ray imaging data obtained for other clinical indications on diagnosing osteoporosis and osteopenia in postmenopausal women, as compared to DXA-defined BMD measurements. The practicable and low-cost deep learning approach could be used as an adjunct to DXA screening, especially in community hospitals in which DXA machines are inadequate.

2. Materials and methods

2.1. Dataset and study population

We performed a retrospective, multicohort, diagnostic study using lumbar spine X-ray images sets from three large tertiary centres in China. This multicentre study was approved by the institutional review board of the principal investigator's hospital. Informed consent from patients was exempted due to the retrospective nature of this study. The inclusion criteria were as follows: (1) postmenopausal women aged ≥50 years, the age of menopause was confirmed by medical records or patients' self-report; (2) women that had undergone both lumbar spine X-ray and DXA examinations within a 3-month period, and during this interval, patients had not received any treatments that would affect BMD; and (3) lumbar spine X-ray images including anteroposterior and lateral images, comprising at least the first through the fourth lumbar vertebrae (L1-L4). The exclusion criteria were as follows: (1) patients had undergone lumbar spine (L1-L4) operation, such as internal fixation and bone cement filling; (2) lumbar spine (L1-L4) presenting lesions, including tumours (e.g., multiple myeloma and metastatic tumor), inflammatory diseases (e.g., ankylosing spondylitis, tuberculosis), serious scoliosis or deformity; (3) region of interests (ROIs) could not map to the raw images; and (4) images with low signal to noise ratio. As a result, we obtained lumbar spine X-ray Digital Imaging and Communications in Medicine (X-ray-DICOM) images from the picture archiving and communication systems repositories: for the primary dataset, 1820 images from 910 patients with imaging performed between January 1, 2014 and December 31, 2017 were collected. The primary dataset was randomly divided into training and internal validation datasets at a ratio of 8:1; for the test dataset 1, 396 images from 198 patients with imaging performed between October 1 and December 31, 2018 were collected; and for the test dataset 2, 294 images from 147 patients with imaging performed between June 1, 2018 and August 1, 2019 were collected. Fig. 1 demonstrates the flowchart of patient inclusion from the different participating centres.

2.2. Design and overview

As shown in Fig. 2, by using cloud database of lumbar spine X-ray images and DXA-derived BMD as a reference standard, we developed artificial intelligence (AI) models to classify patients into three categories according to the World Health Organization criteria: normal (T-score ≥ −1.0), osteopenia (−2.5 < T-score < −1.0), and osteoporosis (T-score ≤ −2.5) [25]. T-scores were calculated by using the BMD dataset of young Chinese female aged 20–40 years as reference. The AI model based on DCNN contained two channels: anteroposterior channel and lateral channel (Fig. 3). The final decision was based on a combination of the two channels.

2.3. Lumbar X-ray examinations

In the primary dataset, the lumbar X-ray scan was performed by the DMC GmbH Digital Radiographic (DR) machine (Philips Medical Systems, Germany), with parameters set at 75 kVp, 32 mAs for anteroposterior imaging, and 85 kVp, 32 mAs for lateral imaging. In the test dataset 1, the lumbar X-ray scan was conducted by the AXIOM Aristos VX DR machine (Siemens, Germany), and 70 kVp, 20-mAs was set for anteroposterior and lateral imaging. In the test dataset 2, the lumbar X-ray scan was also performed by the AXIM Aristos DR machine (Siemens, Germany), with settings of 66 kVp, 35 mAs for anteroposterior imaging and 70 kVp, 80 mAs for lateral imaging.

2.4. Image preprocessing

For each patient, the ROIs were delineated on the trabecular bone of lumbar vertebrae (L1–L4) from anteroposterior and lateral view by four radiologists with 10–20 years' experience. Cortical bone was excluded from the ROIs. Considering the differences in X-ray scan parameters, a series of greyscale normalization was performed, including window width, window level, and window pixel normalization.

$$window\_level = \frac{1}{N}\sum_{i=1}^{N} ct_i \quad (1)$$

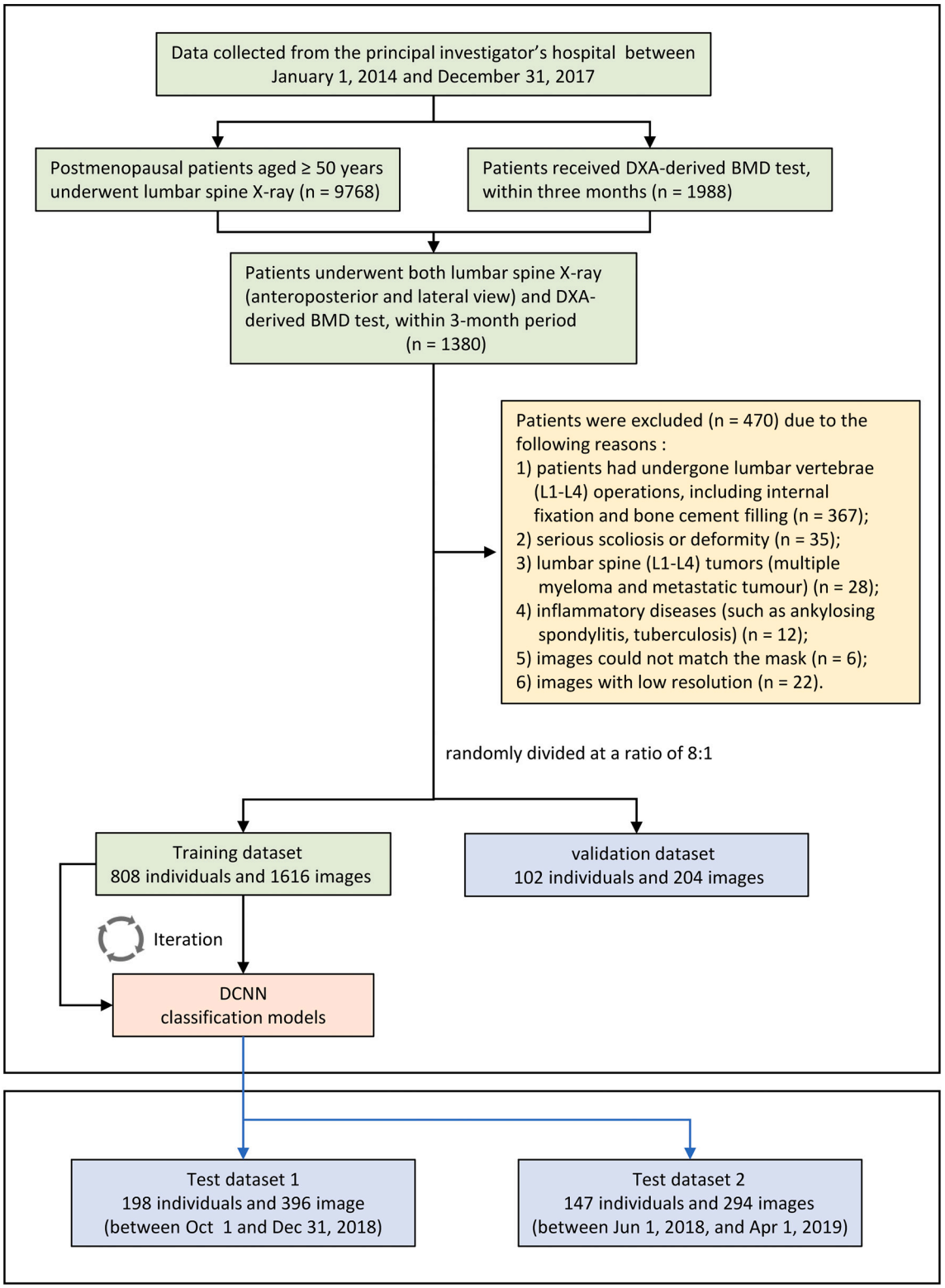

Fig. 1. Study flowchart illustrating multicentre patient inclusion. A total of 1255 postmenopausal females aged 50 years and older from three Chinese hospitals were included, and were allocated to the training, validation and test datasets, respectively. BMD, bone mineral density; DXA, dual energy X-ray absorptiometry; DCNN, deep convolutional neural network; AUC, area under the receiver operating characteristic curve.

$$window\_width = 2 * \sqrt{\frac{1}{N}\sum_{i=1}^{N}(ct_i - window\_level)} \quad (2)$$

$$ct_i' = \min(ct_i, window\_level + window\_width) \quad (3)$$

$$ct_i' = \max(ct_i, window\_level - window\_width) \quad (4)$$

$$pixel = \frac{(ct_i' - (window\_level - window\_width))}{(ct_i' - (window\_level + window\_width))} \quad (5)$$

where, N denotes total pixel value, i denotes pixel coordinates, ct

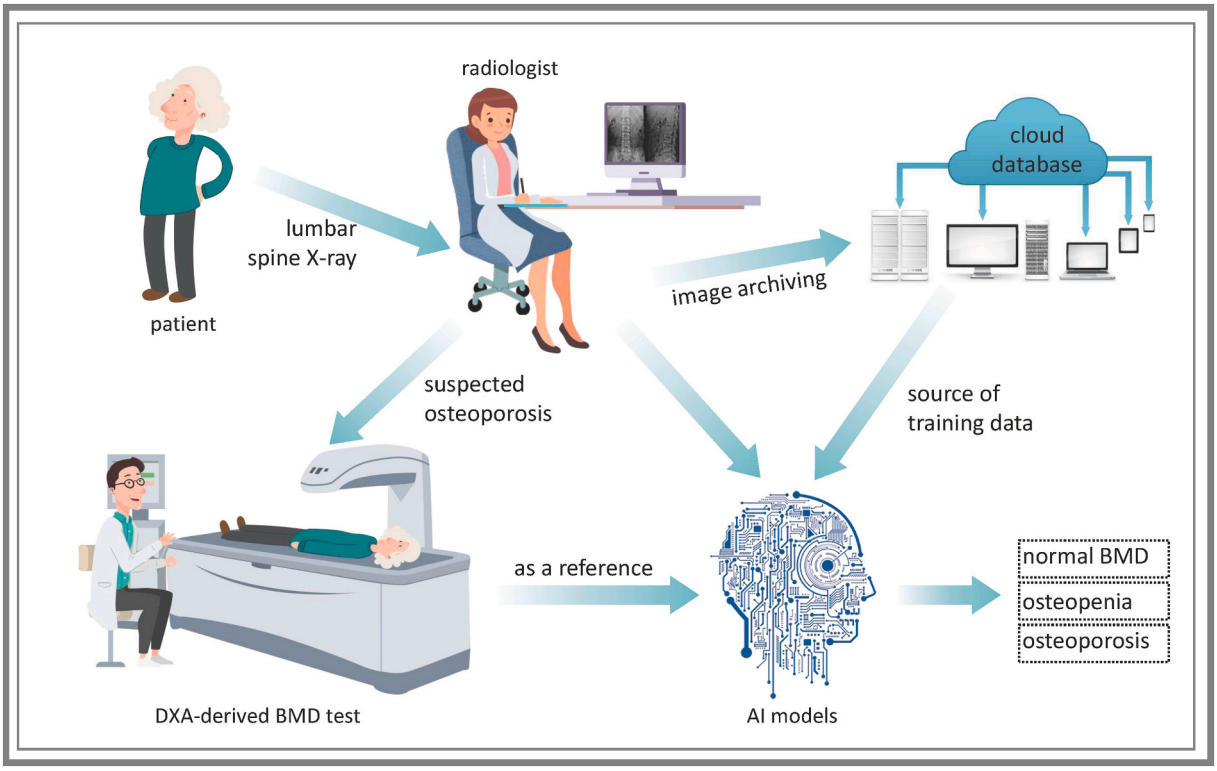

Fig. 2. Cloud-based multihospital AI platform. Whencoming to hospital, patients will be informed to firstly undergo lumbar spine X-ray examination and then DXA-derived BMD test if are suspected with osteoporosis. In this study, we used DCNN to train our AI models based on lumbar spine radiographs, as compared to DXA-derived BMD measures. As a result, we could provide additional assessment of BMD (normal, osteopenia, and osteoporosis) besides basic imaging findings reported by radiologists.

denotes original CT value, and *pixel* denotes value after adjusting for window width and window level.

A patch-based data augmentation strategy was essential to ensure sufficient images to train the DCNN model [26]. The patch-based vote strategy refers to that the final predictive probability of a subject is obtained by voting on the output of patches from the same subject. Specifically, each ROI was resized to 128 × 128 square and averagely divided into four patches with at least 50% effective area. The four

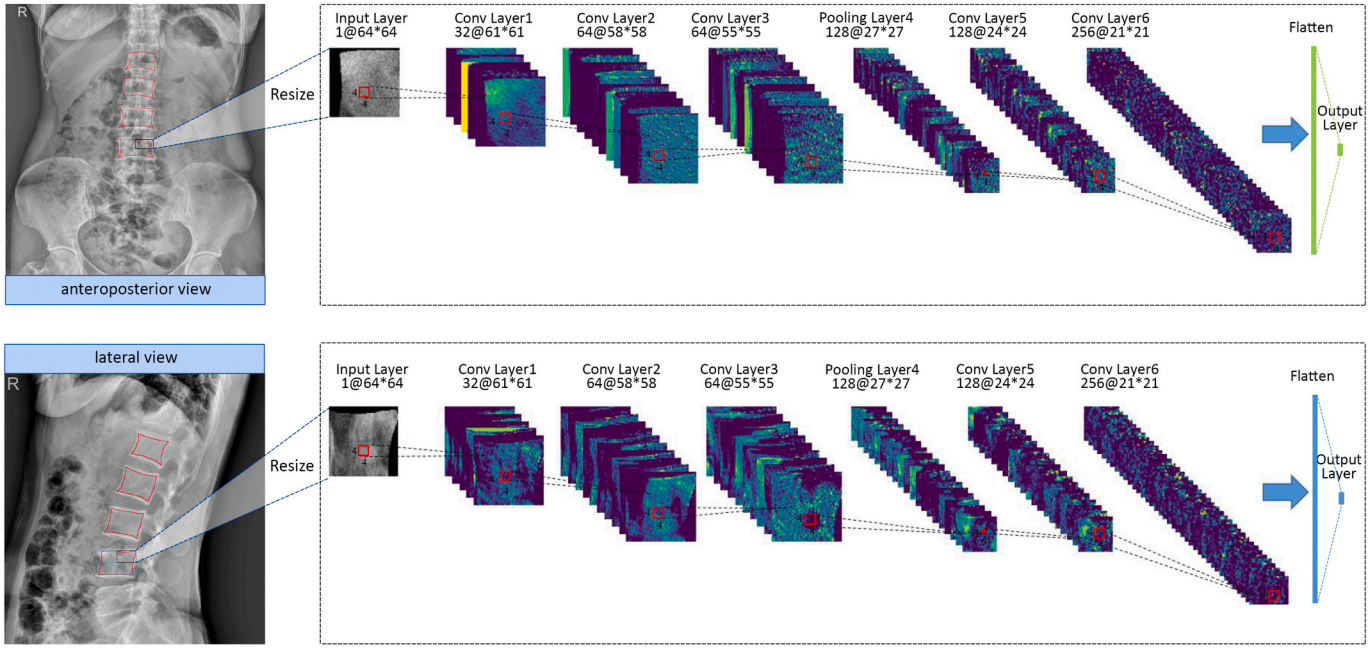

Fig. 3. Illustration of the measurements of anteroposterior and lateral X-rays and deep learning model flowchart. The left part represents the anteroposterior and lateral images and ROIs, while the right part represents the input layer, convolutional layers, pooling layer, and output layer of anteroposterior and lateral channels, respectively. The input of the framework includes the image patch from anteroposterior and lateral X-ray images. The image patches are first computed at the individual channel, and then integrated for the combined diagnosis. ROIs, region of interests.

patches were randomly cropped from the ROI without overlap. The effective area refers to the non-zero pixels in each patch.

2.5. DCNN architecture

The proposed DCNN classification model consisted of two channels to perform automatic analysis of the anteroposterior and lateral lumbar spine (L1–L4) images. The network parameters were determined by the leave-out method in the training process for all the networks (classifiers). Specifically, one-ninth of the primary dataset were stratified and sampled out to verify the parameters of the network and the convergence, so as to select the optimal model finally. The hyperparameters are shown in the Supplementary Table.

Both channels presented with the same structure and mainly included the convolutional layer, pooling layer, and dense layer. The input of each channel was 64 × 64 patches from the ROI. To capture highly relevant textures and edge information and improve model discriminability for the patches within the receptive field, five convolutional layers (kernel size of 4 × 4) and one pooling layer (4th layer with a kernel size of 1 × 1) were placed to extract features from the input patches or feature maps by a convolution operation. To ensure low-level representation and high-level understanding of images, the number of kernels in the five convolutional layers was increased with depth, namely, 32, 64, 64, 128 and 256, respectively. For each convolutional layer, a non-linear 'ReLU' operator was employed as an activation function and batch normalization was performed to avoid gradient vanishing and accelerate convergence. In addition, a max-pooling layer was utilized after the 3rd convolutional layer to preserve discriminable features and decrease the number of parameters and redundant information obtained. Finally, the feature maps were flattened and a dense layer with three neurons was used to perform prediction.

Considering this was a three-category task (normal, osteopenia, and osteoporosis), we trained a 3-class DCNN model to perform classification from anteroposterior, lateral, and anteroposterior + lateral views. The DCNN models generated by using the training dataset of this study are available at: https://github.com/zhang-de-lab/zhang-lab?from=singlemessage. Before training the models, all the neuron biases were set to zero, and all the weights were initialized as random uniform distribution ranging from −0.5 to 0.5. The stochastic gradient descent optimizer with a learning rate of 0.0005 was used to minimize categorical cross-entropy loss of function. After training, the patch-based vote strategy was applied to perform classification. The result of the combined anteroposterior and lateral channels was calculated by weighting and summing the results of a separate anteroposterior and lateral channels. If a patient had an only anteroposterior or lateral image, the result was output from a single channel.

2.6. Assessing the overall classification performance of the DCNN models

All the patients from the primary dataset were randomly and equiprobably allocated to training (1616 images from 808 patients) and validation sets (204 images from 102 patients) at a ratio of approximately 8:1. Data from the other two participating centres (396 images from 198 patients and 294 images from 147 patients respectively) were used as test datasets. The training set was used for training, the validation dataset was used for choosing hyperparameters and identifying a stopping condition, and the test datasets were used to report actual predictive results. The extracted patches were fed into the model to distill low level and high level features automatically. The model was finally structured for classifying cases of normal, osteopenia, and osteoporosis.

2.7. Bone mineral density measurement

For all patients, we measured the body weight (kg) and height (m) by a sensitive digital scale (Electronic Body Scale, TCS-200-RT, China) and a stadiometer respectively, and the body mass index (BMI) was expressed as weight/height^2 (kg/m^2). Patients' age, weight, height, and BMI were collected from BMD testing reports. In the training and validation datasets, lumbar spine BMD was measured by the DXA total-body densitometer (Lunar Prodigy, GE Healthcare, Madison, WI); in the test dataset 2, lumbar spine BMD was measured by the same DXA total-body densitometer; however, in the test dataset 1, lumbar spine BMD was measured by a total-body DXA densitometer (Discovery A, HOLOGIC, USA).

2.8. Statistical analysis

Descriptive statistics and continuous variables were expressed as numbers (percentages) and mean (standard deviation, SD), respectively. For diagnosis purposes, we used the receiver operating characteristic (ROC) curve to indicate the performance of the models in classifying BMD in postmenopausal women. The ROC curve was created by plotting the true positive rate (sensitivity) against the false positive rate (1-sensitivity). By varying the predicted probability threshold, we calculated the area under the curve (AUC) values. We calculated 95% confidence intervals (CIs) for sensitivity and specificity with the bootstrap (1000 iterations) method. Sensitivity/specificity values were taken from a 0.5 threshold on our model outputs. DeLong's test was performed to statistically evaluate the difference in AUC between the deep learning models.

We also calculated the positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (PLR), and negative likelihood ratio (NLR). PPV was the probability that the disease was present when the test was positive (expressed as a percentage). NPV was the probability that the disease was not present when the test was negative (expressed as a percentage). PLR was the ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease (i.e., the true positive rate/false positive rate = sensitivity/[1-specificity]). The NLR was the ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease (i.e. false negative rate/true negative rate = [1-Sensitivity]/Specificity). The confusion matrix in our study was given as a 2 × 2 contingency table that reported the number of true positives, false positives, false negatives, and true negatives.

All the deep convolutional models were complemented by PYTHON (3.6.7, Guido van Rossum, Netherlands). All statistical analyses were carried out by R software (3.0.2, R Core Team, 2013) and MedCalc software (15.6.1, Microsoft Partner, 2015). All experiments were performed under Windows on a machine with Central Processing Unit (CPU) of Intel (R) Core (TM) Processor i5-4590 @ 3.30 GHz, Graphics Processing Unit (GPU) of NVIDIA Pascal Titan X, and RAM of 128 GB. All statistical tests were two-tailed, and P values < 0.05 indicated a significant difference. We reported our findings in accordance with the Guidelines for Standards for Reporting Diagnostic accuracy studies, Developing and Reporting Machine Learning Predictive Models in Biomedical Research [27], and Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis [28].

3. Results

3.1. Patient characteristics

A total of 1255 postmenopausal women (mean age, 65.8 ± 9.1 years; range, 50–92 years) with 2510 lumbar spine X-ray images (anteroposterior: n = 1255; lateral: n = 1255) were included in this retrospective and multicentre cohort study. The baseline characteristics of the training, validation, and two test datasets are shown in Table 1. Age and body mass index in the training, validation and test datasets were comparable. The DXA-defined BMD screening reference

Table 1
Demographic characteristics of 1255 postmenopausal women.

Characteristics	Training dataset	Validation dataset	Test dataset 1	Test dataset 2
Patients	808	102	198	147
Age, yr., mean (SD)	65.3 (8.9)	65.4 (8.6)	65.9 (9.1)	68.9 (9.8)
Height, cm, mean (SD)	154.7 (6.0)	154.2 (6.6)	155.0 (5.8)	153.9 (6.6)
Weight, kg, mean (SD)	58.1 (11.2)	58.8 (14.4)	59.4 (9.9)	56.1 (9.0)
BMI, kg/m^2, mean (SD)	24.0 (4.1)	24.0 (4.4)	24.7 (3.9)	23.6 (3.3)
Lumbar spine images				
Anteroposterior	808	102	198	147
Lateral	808	102	198	147
T-score, mean L1-L4	−1.64	−1.68	−2.00	−1.91
BMD, mean (SD) L1-L4	0.91 (0.20)	0.91 (0.19)	0.82 (0.15)	0.88 (0.18)
BMD categories, n (%)				
Normal	246 (30.4)	30 (29.4)	44 (22.2)	34 (23.1)
Osteopenia	316 (39.1)	39 (38.2)	77 (38.9)	56 (38.1)
Osteoporosis	246 (30.4)	33 (32.4)	77 (38.9)	57 (38.8)

Note: categorical and continuous data were expressed as n (%) and mean (standard deviation, SD), respectively. BMI, body mass index; BMD, bone mineral density.

standard categorized 30.4%, 32.4%, 39.9%, and 38.8% patients as osteoporotic, and 28.2%, 25.5%, 26.3%, and 25.2% patients as osteopenic in the training dataset, validation dataset, and two test datasets, respectively.

3.2. Overall diagnostic performance of the DCNN models

Fig. 4, Tables 2 and 3 show the performance of DCNN model in classifying osteoporosis and osteopenia based on lumbar spine X-ray images. Based on the anteroposterior + lateral channels, the model diagnosed osteoporosis with an AUC of 0.990 (95% CI: 0.982–0.996), sensitivity of 98.0% (95% CI: 95.7–99.3), and specificity of 92.4% (95% CI: 90.0–94.4) (Table 2); while the model diagnosed osteopenia with an AUC of 0.995 (95% CI: 0.988–0.998), sensitivity of 96.2% (95% CI: 93.0–98.1), and specificity of 98.6% (95% CI: 97.3–99.3) (Table 3). Fig. 5 illustrates the training process and how accuracy increases with iteration, which suggests that the DCNN models are gradually convergent.

In the validation dataset, based on the anteroposterior + lateral channels, the model diagnosing osteoporosis achieved the highest AUC of 0.786 (95% CI: 0.693–0.861), sensitivity of 60.6% (95% CI: 42.1–77.1), and specificity of 85.5% (95% CI: 75.0–92.8) (Table 2); while the model diagnosing osteopenia achieved the highest AUC of 0.743 (95% CI: 0.647–0.824), sensitivity of 46.7% (95% CI: 28.3–65.7), and specificity of 88.9% (95% CI: 79.3–95.1) (Table 3).

In the test dataset 1, based on the anteroposterior channel, the model diagnosing osteoporosis yielded the highest AUC of 0.767 (95% CI: 0.701–0.824), with sensitivity of 73.7% (95% CI: 62.3–83.1) (Table 2). Based on the anteroposterior + lateral channels, the model diagnosing osteopenia yielded the highest AUC of 0.787 (95% CI: 0.723–0.842), with sensitivity of 81.8% (95% CI: 67.3–91.8) (Table 3).

In the test dataset 2, based on the anteroposterior + lateral channels, the model diagnosing osteoporosis yielded the highest AUC of 0.726 (95% CI: 0.646–0.796), with sensitivity of 68.4% (95% CI: 54.8–80.1) (Table 2). Based on the anteroposterior channel, the model diagnosing osteopenia yielded the highest AUC of 0.810 (95% CI: 0.737–0.870), with sensitivity of 85.3% (95% CI: 68.9–95.0) (Table 3).

4. Discussion

This is the first multicentre cohort study that aimed to classify osteoporosis and osteopenia in postmenopausal women (n = 1255) aged ≥ 50 years using developed DCNN models based on conventional X-rays performed for other clinical indications, with DXA measures. The results demonstrated that deep-learning approach might have potential in automated BMD classification in a real-world setting.

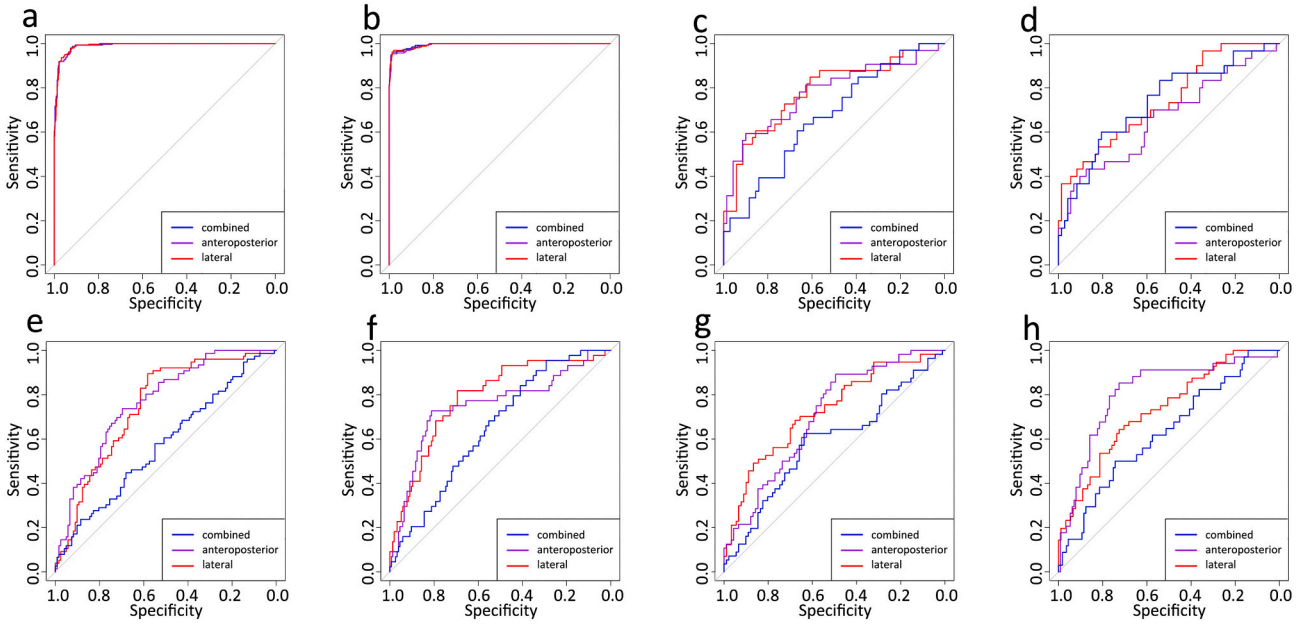

Fig. 4. Comparison of ROC curves between the DCNN model based on single and combined image projects. (a) model diagnosed osteoporosis in the training dataset; (b) model diagnosed osteopenia in the training dataset; (c) model diagnosed osteoporosis in the validation dataset; (d) model diagnosed osteopenia in the validation dataset; (e) model diagnosed osteoporosis in the test dataset 1; (f) model diagnosed osteopenia in the test dataset 1; (g) model diagnosed osteoporosis in the test dataset 2; and (h) model diagnosed osteopenia in the test 2. ROC, receiver operating characteristic; DCNN, deep convolutional neural network; BMD, bone mineral density.

Table 2
Performance metrics for the ensemble DCNN model screening osteoporosis, assessed on the validation and test datasets.

Dataset	Image projection	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	PLR	NLR
Training dataset	Anteroposterior	0.990 (0.981–0.996)	98.0 (95.7–99.3)	92.3 (89.8–94.3)	86.4 (82.8–89.4)	98.9 (97.6–99.5)	12.68 (9.6–16.7)	0.02 (0.01–0.05)
	Lateral	0.990 (0.982–0.996)	99.0 (97.1–99.8)	90.4 (87.8–92.7)	83.9 (80.2–86.9)	99.4 (98.3–99.8)	10.33 (8.1–13.2)	0.01 (0.004–0.03)
	Anteroposterior and lateral	0.990 (0.982–0.996)	98.0 (95.7–99.3)	92.4 (90.0–94.4)	86.7 (83.1–89.6)	98.9 (97.6–99.5)	12.96 (9.8–17.2)	0.02 (0.01–0.05)
Validation dataset	Anteroposterior	0.777 (0.684–0.854)	59.4 (40.6–76.3)	90.0 (80.5–95.9)	73.1 (56.0–85.3)	82.9 (76.0–88.1)	5.94 (2.8–12.7)	0.45 (0.3–0.7)
	Lateral	0.668 (0.568–0.759)	63.6 (45.1–79.6)	63.8 (51.3–75.0)	45.7 (35.9–55.8)	78.6 (69.3–85.6)	1.8 (1.2–2.6)	0.6 (0.4–0.9)
	Anteroposterior and lateral	0.786 (0.693–0.861)	60.6 (42.1–77.1)	85.5 (75.0–92.8)	66.7 (51.4–79.1)	81.9 (74.6–87.5)	4.2 (2.2–7.9)	0.5 (0.3–0.7)
Test dataset 1	Anteroposterior	0.767 (0.701–0.824)	73.7 (62.3–83.1)	69.7 60.7–77.7)	60.2 (52.8–67.2)	81.0 (74.1–86.3)	2.43 (1.8–3.3)	0.38 (0.3–0.6)
	Lateral	0.575 (0.502–0.644)	57.9 (46.0–69.1)	54.9 (45.7–63.9)	44.4 (37.8–51.3)	67.7 (60.6–74.0)	1.3 (1.0–1.7)	0.77 (0.6–1.0)
	Anteroposterior and lateral	0.747 (0.681–0.806)	89.5 (80.3–95.3)	58.2 (48.9–67.1)	57.1 (51.6–62.5)	89.9 (81.9–94.6)	2.14 (1.7–2.7)	0.18 (0.09–0.4)
Test dataset 2	Anteroposterior	0.704 (0.623–0.777)	89.3 (78.1–96.0)	49.5 (38.8–60.1)	52.1 (46.5–57.6)	88.2 (77.4–94.3)	1.77 (1.4–2.2)	0.22 (0.10–0.5)
	Lateral	0.586 (0.501–0.666)	62.5 (48.5–75.1)	63.7 (53.0–73.6)	51.5 (43.0–59.8)	73.4 (65.6–80.0)	1.72 (1.2–2.4)	0.59 (0.4–0.9)
	Anteroposterior and lateral	0.726 (0.646–0.796)	68.4 (54.8–80.1)	67.8 (57.1–77.2)	57.4 (48.7–65.6)	77.2 (69.3–83.6)	2.12 (1.5–3.0)	0.47 (0.3–0.7)

Note: Statistical quantifications were demonstrated with 95% CI, when applicable. AUC, area under the receiver operating characteristic curve; DCCN, deep convolutional neural network; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

Currently, the only reference standard accepted by the WHO to measure BMD is central DXA of the hips and lumbar spine, other available screening strategies include calcaneal quantitative ultrasonography (QUS), quantitative computed tomography (QCT), and magnetic resonance imaging (MRI) [29]. QCT measures trabecular BMD in milligrams per cubic centimetre by indirectly quantifying hydroxyapatite in comparison to a reference phantom [30]. However, an obvious limitation of QCT is the significantly higher radiation dose associated with scanning central body sites as compared to DXA (0.02 mSv) [31,32]. Radiation exposure doses are in the order of 1.5 mSv for the spine and 2.5–3.0 mSv for the hips [33]. In addition, QCT requires sophisticated calibration and positioning techniques and careful technical monitoring. Although QUS is a low-cost technique for assessing fracture risk, it examines calcaneus in different skeletal sites, measures and reports instable bone parameters in various ways, and has differing levels of validation data for association with DXA-derived BMD, all of these reasons limit the application of QUS in clinical practice [34]. Dozens of early studies evaluated the use of routine CT examinations for assessing BMD [35,36], but the opinions were divided by the effects of the use of contrast media, region of interest selection, and threshold Hounsfield units. In addition, most studies were done in a single centre and lacked external validation. Machine-to-machine

Table 3
Performance metrics for the ensemble DCNN model screening osteopenia, assessed on the validation and test datasets.

Dataset	Image projection	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	PLR	NLR
Training dataset	Anteroposterior	0.994 (0.986–0.998)	95.0 (91.6–97.3)	99.1 (98.0–99.7)	97.6 (94.9–98.9)	98.0 (96.6–98.8)	100.38 (45.2–222.7)	0.05 (0.03–0.09)
	Lateral	0.995 (0.987–0.998)	97.3 (94.5–98.9)	90.1 (87.5–92.3)	80.1 (76.0–83.6)	98.8 (97.5–99.4)	9.8 (7.7–12.4)	0.03 (0.01–0.06)
	Anteroposterior and lateral	0.995 (0.988–0.998)	96.2 (93.0–98.1)	98.6 (97.3–99.3)	96.5 (93.6–98.2)	98.4 (97.1–99.1)	67.74 (35.4–129.6)	0.039 (0.02–0.07)
Validation dataset	Anteroposterior	0.663 (0.562–0.753)	43.3 (25.5–62.6)	87.5 (77.6–94.1)	59.1 (40.9–75.1)	78.7 (72.8–83.7)	3.5 (1.7–7.2)	0.65 (0.5–0.9)
	Lateral	0.733 (0.637–0.816)	60.0 (40.6–77.3)	80.6 (69.5–88.9)	56.2 (42.5–69.1)	82.9 (75.5–88.4)	3.1 (1.8–5.4)	0.5 (0.3–0.8)
	Anteroposterior and lateral	0.743 (0.647–0.824)	46.7 (28.3–65.7)	88.9 (79.3–95.1)	63.6 (45.1–78.9)	80.0 (73.9–85.0)	4.2 (2.0–9.0)	0.6 (0.4–0.8)
Test dataset 1	Anteroposterior	0.751 (0.685–0.809)	72.7 (57.2–85.0)	81.2 (74.1–87.0)	52.5 (43.1–61.6)	91.2 (86.5–94.4)	3.9 (2.7–5.6)	0.34 (0.2–0.5)
	Lateral	0.650 (0.580–0.717)	84.1 (69.9–93.4)	40.9 (33.1–49.1)	28.9 (25.3–32.8)	90.0 (81.6–94.8)	1.4 (1.2–1.7)	0.4 (0.2–0.8)
	Anteroposterior and lateral	0.787 (0.723–0.842)	81.8 (67.3–91.8)	69.5 (61.6–76.6)	43.4 (36.8–50.2)	93.0 (87.6–96.2)	2.7 (2.0–3.5)	0.3 (0.1–0.5)
Test dataset 2	Anteroposterior	0.810 (0.737–0.870)	85.3 (68.9–95.0)	72.6 (63.4–80.5)	48.3 (40.2–56.6)	94.3 (87.9–97.4)	3.1 (2.2–4.3)	0.2 (0.09–0.5)
	Lateral	0.629 (0.545–0.707)	50.0 (32.4–67.6)	74.3 (65.3–82.1)	37.0 (27.0–48.1)	83.2 (77.6–87.6)	2.0 (1.2–3.1)	0.7 (0.5–1.0)
	Anteroposterior and lateral	0.732 (0.652–0.801)	64.3 (50.4–76.6)	72.5 (62.2–81.4)	59.0 (49.4–67.9)	76.7 (69.4–82.7)	2.4 (1.6–3.4)	0.5 (0.3–0.7)

Note: Statistical quantifications were demonstrated with 95% CI, when applicable. AUC, area under the receiver operating characteristic curve; DCCN, deep convolutional neural network; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

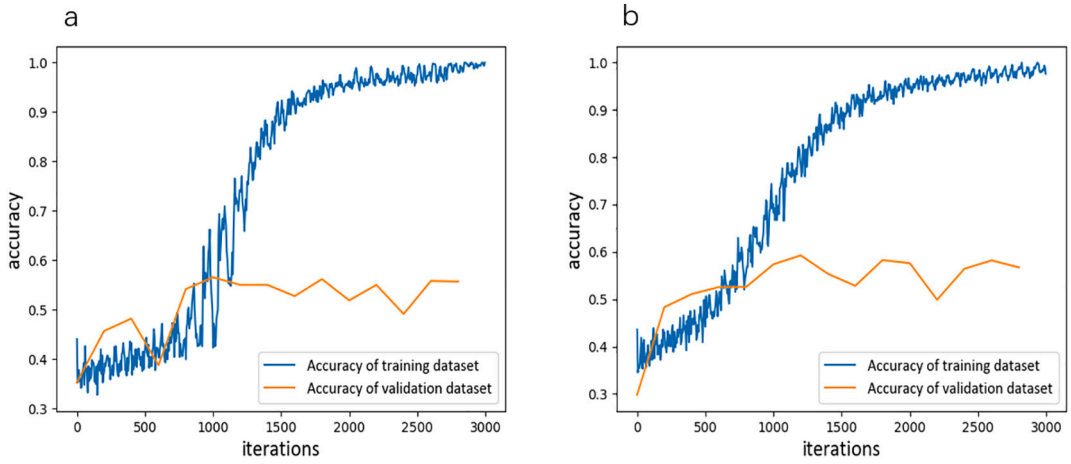

Fig. 5. The training processes of DCNN Model. (a) anteroposterior channel and (b) lateral channel. DCNN, deep convolutional neural network.

variability limited the broad applicability of this technique in its current state. MRI is a method that, while not providing information on BMD, provides some resolution on the internal structure of spongy bone [37]. MRI is currently used as an investigation tool due to high cost and complexity [38].

AI models are increasingly used in medicine. However, few studies have used these models to assess the risk of osteoporosis and fracture. Yoo TK et al. identified the risk of osteoporosis in postmenopausal women by machine learning algorithms and found better prediction results than conventional clinical decision tools [39]. Ferizi U et al. trained and validated 15 machine learning algorithms to predict fragility fractures from quantitative MRI data [40]. Atkinson EJ et al. assessed fracture risk using CT-based gradient boosting machine models [41]. Cruz AS et al. reviewed the main machine learning models to identify groups at risk for osteoporosis or fractures [42]. All the analyzed studies predicted fracture risk based on clinical risk factors with or without considering BMD.

Unlike traditional machine learning, deep learning does not require engineered features designed by human experts [22]. Rather, deep learning allows computational models that are composed of multiple processing layers to learn data representations with multiple levels of abstraction. Therefore, deep learning overpass traditional machine learning in an end-to-end way. We hypothesized that routine lumbar spine radiographs in clinical setting are good options for the opportunistic classification of BMD by deep learning. For the first time, we trained a cost-effective and highly available AI-aided tool to assess BMD in postmenopausal women. Conferred by the high speed of a GPU, the DCNN models have the advantage of learning imaging biomarkers of BMD by an automatic procedure. Although regression task has some advantages over classification task, such as less likely to divide patients who have similar BMD T-score (around the thresholds) into different categories (osteopenia vs osteoporosis) and could also apply classification after regression using the usual T-score thresholds, regression modelling has challenge in complex non-linear problem. In this study, the relationship between input X-ray images and BMD T-score was non-linear, therefore, we chose a new classification model to evaluate BMD T-score. Unlike common deep learning models based on binary classification, our DCNN models were based on triple classification of the entire lumbar spine radiographs into normal, osteopenia, and osteoporosis. The high sensitivity of DCNN models results in a low fraction of false-negative classifications and therefore osteoporotic subjects will potentially be identified, and consequently treated. However, sensitivity of our models was somewhat low, which may due to data imbalance, the ROI used for the X-ray images (excluding cortical bone) was different from that used in DXA (including cortical bone), and the inclusion of lumbar fractures. DXA including cortical bone will result in overestimation of BMD and therefore increasing the sensitivity of screening osteoporosis, while our models did not include cortical bone may lead to reduced sensitivity when using DXA as a reference. Relatively low sensitivity of DCNN models is not conducive to osteoporosis screening, thus the main direction for future study is to improve the sensitivity of the models.

Our study also has some potential limitations. Firstly, the features and calculations that deep learning models used to make a classification are challenging to interpret. Therefore, when the judgment of physicians or radiologists differs from that of trained models, the discrepancy cannot be resolved by discussion. Secondly, patient positioning can lead to a significant variability on the BMD measurement for DXA analysis. In clinical practice, we have no effective methods to deal with this limitation except for using standard positioning. DXA measures BMD in a projectional image, the measurement is sensitive to degenerative changes of the spine and overlying structures such as aortic calcification, which will result in an overestimation of BMD. Similar to DXA, our DCNN models may also be affected by aortic, sclerotic, and osteophytic calcifications, and may produce falsely elevated BMD values. Moreover, lumbar spines exhibiting tumor, infection diseases, severe scoliosis, or deformation were also not suitable for the DCNN models. Thirdly, our study focused on women aged ≥ 50 years, the classification performance of DCNN models in women aged < 50 years or in men needs to be investigated in prospective studies. Fourthly, although we performed image normalization, the variability across images maybe not eliminated due to different X-ray scan settings. Finally, we could not formally evaluate the potential benefits and costs of the DCNN models, but we speculated increasing detection of osteopenia and osteoporosis with subsequent appropriate treatment to reduce fracture risk, combined with reducing the number of normal DXA studies, which is expected to yield substantial health care cost savings [43]. We are building a website or mobile application that will provide free access to the developed DCNN models. Medical resources in urban and rural areas in China -and in many other countries in the world-are unbalanced; the deep learning system developed in our study could contribute to reducing barriers and providing a convenient way for community hospitals to screen osteopenia and osteoporosis, particularly in rural areas.

In summary, we provided a deep learning model that achieves favourable performance in clinical tasks on opportunistic BMD classification in postmenopausal women. Routine lumbar spine X-ray images obtained for other reasons can be applied to identify patients at risk of osteoporosis and osteopenia without additional radiation exposure or cost, which would improve screening and availability of BMD classification. Although the deep learning models could not replace DXA for BMD screening, it could be used in a scenario in which lumbar spine X-ray is readily available but DXA has not been performed. Before

switching to prospective clinical trials, it might be necessary to use retrospective data to evaluate the ability of the proposed deep learning method in assessing fracture risk.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bone.2020.115561.

CRediT authorship contribution statement

Bin Zhang: Conceptualization, Formal analysis, Data curation, Writing - original draft, Writing - review & editing.**Keyan Yu:**Conceptualization, Formal analysis, Data curation.**Zhenyuan Ning:**Formal analysis, Data curation, Writing - original draft, Writing - review & editing.**Ke Wang:**Formal analysis, Data curation.**Yuhao Dong:**Formal analysis, Data curation.**Xian Liu:**Investigation.**Shuxue Liu:**Investigation.**Jian Wang:**Investigation.**Cuiling Zhu:**Investigation.**Qinqin Yu:**Investigation.**Yuwen Duan:**Investigation.**Siying Lv:**Investigation.**Xintao Zhang:**Investigation.**Yanjun Chen:**Investigation. **Xiaojia Wang:**Investigation.**Jie Shen:**Investigation.**Jia Peng:**Investigation. **Qiuying Chen:**Formal analysis, Data curation.**Yu Zhang:**Formal analysis, Data curation.**Xiaodong Zhang:**Conceptualization.**Shuixing Zhang:** Conceptualization.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgments

This work was supported by a grant of the National Natural Science Foundation of China (81571664, 81871323, 81801665, 81801653); the Natural Science Foundation of Guangdong Province (2018B030311024); the Science and Technology Planning Project of Guangdong Province (2017B090912006), the Scientific Research General Project of Guangzhou Science Technology and Innovation Commission (201707010328); and China Postdoctoral Science Foundation (2016M600145) and the Fundamental Research Funds for the Central Universities (21620447). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] A.M. Cheung, A. Papaioannou, S. Morin, Osteoporosis Canada Scientific Advisory Council, Postmenopausal osteoporosis, N. Engl. J. Med. 374 (2016) 2096.
[2] U.S. Preventive Services Task Force, S.J. Curry, A.H. Krist, D.K. Owens, M.J. Barry, A.B. Caughey, et al., Screening for osteoporosis to prevent fractures: US Preventive Services Task Force recommendation statement, JAMA. 319 (2018) 2521–2531.
[3] C. Brown, Osteoporosis: staying strong, Nature. 550 (2017) S15–S17.
[4] S.C. Schuit, M. van der Klift, A.E. Weel, C.E. de Laet, H. Burger, E. Seeman, et al., Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam study, Bone 34 (2004) 195–202.
[5] A. Wainwright Stacey, M. Marshall Lynn, E. Ensrud Kristine, A. Cauley Jane, M. Black Dennis, A. Hillier Teresa, et al., Hip fracture in women without osteoporosis, J. Clin. Endocrinol. Metab. 90 (2005) 2787–2793.
[6] U.S. Preventive Services Task Force, Screening for osteoporosis in postmenopausal women: recommendations and rationale, Am. Fam. Physician 66 (2002) 1430–1432.
[7] H.P. Dimai, Use of dual-energy X-ray absorptiometry (DXA) for diagnosis and fracture risk assessment; WHO-criteria, T- and Z-score, and reference databases, Bone. 104 (2017) 39–43.
[8] Centers for Medicare & Medicaid Services, National physician fee schedule, Available from: www.cms.hhs.gov/PFSlookup/ , Accessed date: 17 April 2019.
[9] A.B. King, D.M. Fiorentino, Medicare payment cuts for osteoporosis testing reduced use despite tests' benefit in reducing fractures, Health Aff (Millwood). 30 (2011) 2362–2370.
[10] A.L. Amarnath, P. Franks, J.A. Robbins, G. Xing, J.J. Fenton, Underuse and overuse of osteoporosis screening in a regional health system: a retrospective cohort study, J. Gen. Intern. Med. 30 (2015) 1733–1740.
[11] National Health Commission of China, Epidemiological investigation of osteoporosis in China, Available from: http://www.phsciencedata.cn/Share/jsp/PublishManager/foregroundView/1/9eaadbb3-bd64-4531-9d9f-753ec183f26d.html , Accessed date: 15 May 2019.
[12] Uran Ferizi, Stephen Honig, Gregory Chang, Artificial intelligence, osteoporosis and fragility fractures, Curr. Opin. Rheumatol. 31 (2019) 368–375.
[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444.
[14] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, et al., A guide to deep learning in healthcare, Nat. Med. 25 (2019) 24–29.
[15] P. Rajpurkar, J. Irvin, R.L. Ball, K. Zhu, B. Yang, H. Mehta, et al., Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLoS Med. 15 (2018) e1002686.
[16] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, et al., Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study, Lancet Oncol. 20 (2019) 193–201.
[17] A. Hosny, C. Parmar, T.P. Coroller, P. Grossmann, R. Zeleznik, A. Kumar, et al., Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study, PLoS Med. 15 (2018) e1002711.
[18] L. Lin, Q. Dou, Y.M. Jin, G.Q. Zhou, Y.Q. Tang, W.L. Chen, et al., Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma, Radiology 291 (3) (2019) 677–686, https://doi.org/10.1148/radiol.2019182012.
[19] J.N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.A. Weis, et al., Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study, PLoS Med. 16 (2019) e1002730.
[20] Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci. Transl. Med. 2011; 3: 108r-113r.
[21] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature. 542 (2017) 115–118.
[22] H.A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, et al., Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, Ann. Oncol. 29 (2018) 1836–1842.
[23] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA. 316 (2016) 2402–2410.
[24] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nat. Med. 24 (2018) 1342–1350.
[25] P.M. Camacho, S.M. Petak, N. Binkley, B.L. Clarke, S.T. Harris, D.L. Hurley, et al., American association of clinical endocrinologists and American college of endocrinology clinical practice guidelines for the diagnosis and treatment of postmenopausal osteoporosis, Endocr. Pract. 22 (2016) 1–42.
[26] Z. Ning, J. Luo, Y. Li, S. Han, Q. Feng, Y. Xu, et al., Pattern classification for gastrointestinal stromal tumors by integration of radiomics and deep convolutional features, IEEE J Biomed Health Inform. 23 (2019) 1181–1191.
[27] W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, et al., Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view, J. Med. Internet Res. 18 (2016) e323.
[28] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, Ann. Intern. Med. 162 (2015) 55–63.
[29] H.D. Nelson, M. Helfand, S.H. Woolf, J.D. Allan, Screening for postmenopausal osteoporosis: a review of the evidence for the U.S. Preventive Services Task Force, Ann. Intern. Med. 137 (2002) 529–541.
[30] J.E. Adams, Quantitative computed tomography, Eur. J. Radiol. 71 (2009) 415–424.
[31] E. Bezakova, P.J. Collins, A.H. Beddoe, Absorbed dose measurements in dual energy X-ray absorptiometry (DXA), Br. J. Radiol. 70 (1997) 172–179.
[32] J. Damilakis, J.E. Adams, G. Guglielmi, T.M. Link, Radiation exposure in X-ray-based imaging techniques used in osteoporosis, Eur. Radiol. 20 (2010) 2707–2714.
[33] M.A. Krieg, R. Barkmann, S. Gonnelli, A. Stewart, D.C. Bauer, L. Del Rio Barquero, et al., Quantitative ultrasound in the management of osteoporosis: the 2007 ISCD official positions, J. Clin. Densitom. 11 (2008) 163–187.
[34] E.B. Gausden, B.U. Nwachukwu, J.J. Schreiber, D.G. Lorich, J.M. Lane, Opportunistic use of CT imaging for osteoporosis screening and bone density assessment: a qualitative systematic review, J. Bone Joint Surg. Am. 99 (2017) 1580–1590.
[35] M.R.K. Mookiah, A. Rohrmeier, M. Dieckmeyer, K. Mei, F.K. Kopp, P.B. Noel, et al., Feasibility of opportunistic osteoporosis screening in routine contrast-enhanced multi detector computed tomography (MDCT) using texture analysis, Osteoporos. Int. 29 (2018) 825–835.
[36] S. Jang, P.M. Graffy, T.J. Ziemlewicz, S.J. Lee, R.M. Summers, P.J. Pickhardt, Opportunistic osteoporosis screening at routine abdominal and thoracic CT: normative L1 trabecular attenuation values in more than 20 000 adults, Radiology. 291 (2019) 360–367.
[37] T. Lang, P. Augat, S. Majumdar, X. Ouyang, H.K. Genant, Noninvasive assessment of bone density and structure using computed tomography and magnetic resonance, Bone. 22 (1998) 149S–153S.
[38] Y. Chen, Y. Guo, X. Zhang, Y. Mei, Y. Feng, X. Zhang, Eur. Radiol. 28 (2018) 5027–5034.
[39] T.K. Yoo, S.K. Kim, D.W. Kim, J.Y. Choi, W.H. Lee, E. Oh, et al., Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning, Yonsei Med. J. 54 (2013) 1321–1330.

[40] U. Ferizi, H. Besser, P. Hysi, J. Jacobs, C.S. Rajapakse, C. Chen, et al., Artificial intelligence applied to osteoporosis: a performance comparison of machine learning algorithms in predicting fragility fractures from MRI data, J. Magn. Reson. Imaging 49 (2019) 1029–1038.

[41] E.J. Atkinson, T.M. Therneau, L.R. Melton, J.J. Camp, S.J. Achenbach, S. Amin, et al., Assessing fracture risk using gradient boosting machine (GBM) models, J. Bone Miner. Res. 27 (2012) 1397–1404.

[42] A.S. Cruz, H.C. Lins, R. Medeiros, J. Filho, S.S. Da, Artificial intelligence on the identification of risk groups for osteoporosis, a general review, Biomed. Eng. Online 17 (2018) 12.

[43] S. Nayak, M.S. Roberts, S.L. Greenspan, Cost-effectiveness of different screening strategies for osteoporosis in postmenopausal women, Ann. Intern. Med. 155 (2011) 751–761.