

# Deep learning in the radiologic diagnosis of osteoporosis: a literature review

Journal of International Medical Research

2024, Vol. 52(4) 1–18

© The Author(s) 2024

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/03000605241244754

[journals.sagepub.com/home/imr](https://journals.sagepub.com/home/imr)

Yu He<sup>1,\*</sup>, Jiayi Lin<sup>2,\*</sup>, Shiqi Zhu<sup>2,\*</sup>,  
Jinzhou Zhu<sup>2</sup>  and Zhonghua Xu<sup>3</sup>

## Abstract

**Objective:** Osteoporosis is a systemic bone disease characterized by low bone mass, damaged bone microstructure, increased bone fragility, and susceptibility to fractures. With the rapid development of artificial intelligence, a series of studies have reported deep learning applications in the screening and diagnosis of osteoporosis. The aim of this review was to summarize the application of deep learning methods in the radiologic diagnosis of osteoporosis.

**Methods:** We conducted a two-step literature search using the PubMed and Web of Science databases. In this review, we focused on routine radiologic methods, such as X-ray, computed tomography, and magnetic resonance imaging, used to opportunistically screen for osteoporosis.

**Results:** A total of 40 studies were included in this review. These studies were divided into three categories: osteoporosis screening (n = 20), bone mineral density prediction (n = 13), and osteoporotic fracture risk prediction and detection (n = 7).

**Conclusions:** Deep learning has demonstrated a remarkable capacity for osteoporosis screening. However, clinical commercialization of a diagnostic model for osteoporosis remains a challenge.

## Keywords

Osteoporosis, screening, deep learning, bone mineral density, literature review, convolutional neural network

Date received: 1 August 2023; accepted: 26 February 2024

<sup>3</sup>Department of Orthopedics, Jintan Affiliated Hospital to Jiangsu University, Changzhou, China

\*These authors contributed equally to this work.

### Corresponding author:

Zhonghua Xu, Department of Orthopedics, Jintan Affiliated Hospital to Jiangsu University, Changzhou 213200, China.

Email: [xuzhonghua1985@163.com](mailto:xuzhonghua1985@163.com)

<sup>1</sup>Suzhou Medical College, Soochow University, Suzhou, Jiangsu, China

<sup>2</sup>Department of Gastroenterology, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative

Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

## Introduction

Osteoporosis is a systemic bone disease characterized by low bone mass, damaged bone microstructure, increased bone fragility, and susceptibility to fractures.<sup>1</sup> The clinical consequences of osteoporosis include bone fracture, which poses various health, economic, and social problems.<sup>2,3</sup> Dual-energy X-ray absorptiometry (DXA) or quantitative CT (QCT) are considered the gold standard for diagnosing osteoporosis by measuring bone mineral density (BMD).<sup>4-8</sup> Osteoporosis is a multifactorial disease with clinical risk factors that include age, sex, weight, previous fracture, smoking, excessive drinking, and use of glucocorticoids.<sup>9</sup> Individual risk assessment of osteoporosis can greatly benefit the early prevention and treatment of the disease. The World Health Organization (WHO) Fracture Risk Assessment Tool (FRAX) is the most recommended tool for predicting fractures.<sup>10,11</sup> This tool combines clinical risk factors with or without BMD to predict the probability of experiencing a fracture within 10 years.

Artificial intelligence (AI) uses computers and algorithms to emulate the decision-making and problem-solving capabilities of the human mind. Machine learning (ML), a subfield of AI, can be defined as the process of learning rules from data using statistical methods.<sup>12</sup> ML can be roughly divided into two categories: supervised learning and unsupervised learning. Supervised learning uses given data to learn toward a specific goal whereas unsupervised learning aims to discover relationships within the data. Deep learning (DL), a subfield of ML, can be considered an extension of ML that applies multilayered model architecture to uncover the underlying rules and representative layers of sample data.

With the rapid development of AI, these approaches have led to applications in medical image recognition, image data

management, and AI-assisted screening and diagnosis. The aim of the present study was to review the application of DL methods in the radiologic diagnosis of osteoporosis.

## Methods

### *Search strategy and selection process*

We developed this review based on the PRISMA guideline.<sup>13</sup> A literature search was conducted using the PubMed and Web of Science databases with the following search terms: (“artificial intelligence” OR “deep learning” OR “neural network”) AND osteoporosis AND (“CT” OR “X-ray” OR “MRI” OR “panoramic radiography”). In the process of searching, we excluded review articles. This study included literature published from 2018 to 2023. Specifically, studies after 2018 were selected because researchers have made significant advancements in the application of DL since then whereas substantial contributions of ML and AI can be found in previous research. Furthermore, we focused on routine radiologic methods, such as X-ray, CT, and magnetic resonance imaging (MRI), used to opportunistically screen for osteoporosis. Based on the categorization criteria proposed by the WHO, a T-score lower than or equal to  $-2.5$  is indicative of osteoporosis whereas a T-score between  $-1$  and  $-2.5$  denotes osteopenia. This standard was widely accepted in the included studies.

After the initial search, a two-step screening scheme was implemented. Titles and abstracts were screened for inclusion, and the following exclusion criteria were applied: (1) duplicated studies, and (2) studies that were irrelevant to ML and osteoporosis. Additionally, we restricted studies to only those reports written in the English language. The full texts of each study were then collated for further screening to ensure

study quality. In cases where there was disagreement between the two reviewers (He and Zhu) during the initial screening, the corresponding author (Xu) facilitated discussion and made the final decision. Owing to the diversity of methods and objectives, quantitative analysis was limited.

### Data extraction

Data extraction from the included studies was performed by two independent reviewers (He and Zhu) using a standardized data extraction form. Any disagreements were resolved through consensus or with the corresponding author (Xu) making the final judgment. The extracted information included the country, year, author, model selection, data amount, study objective, and performance metrics of the model such as area under the curve (AUC), accuracy (ACC), sensitivity, and specificity, among others.

This study was a literature review regarding deep learning in osteoporosis rather than an original study. Thus, there was no observational or interventional process in the study in terms of epidemiological design. For this reason, the requirement for ethics approval and informed consent was waived.

## Results

### Search results

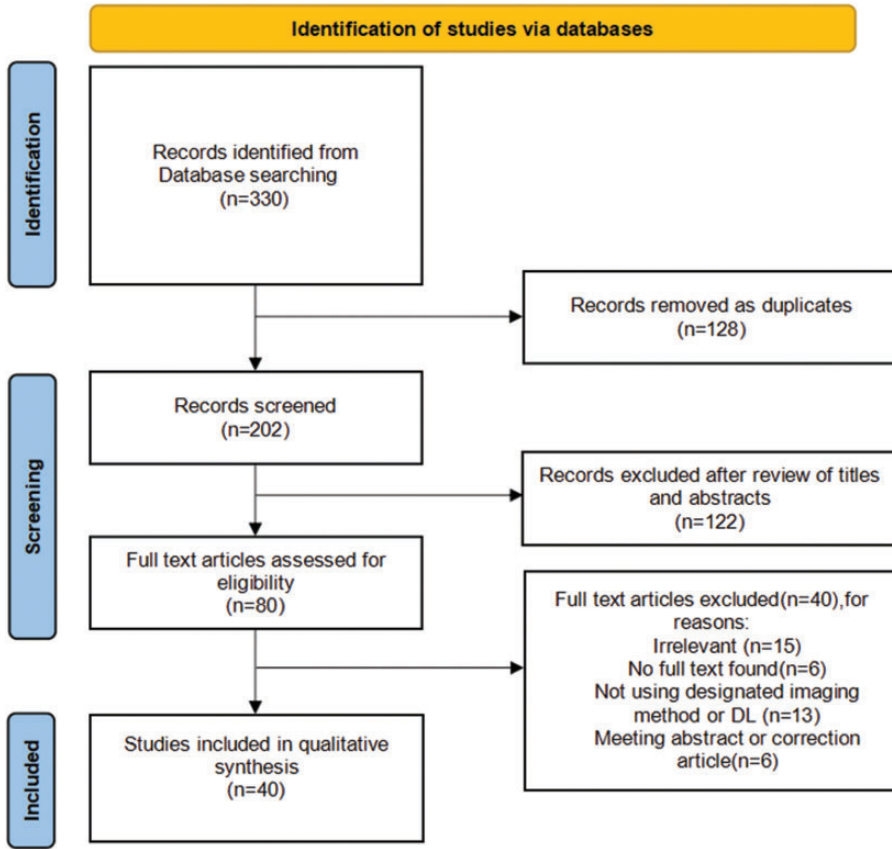
In the initial search of PubMed and Web of Science, 330 records were identified. After the exclusion of duplicates, defined as citations found in both databases, the titles and abstracts of 202 records were screened, and 122 of these studies were excluded after screening. Next, we conducted a full-text review of the remaining 80 articles. A total of 40 articles were finally included in the review. The entire screening process is illustrated in Figure 1.

The characteristics of the included studies are presented in Table 1 and Figure 2. All included studies were published between 2018 and 2023, with most being published after 2020. X-ray (17/40)<sup>14–25</sup> and CT (16/40)<sup>26–41</sup> images were the most frequently used modalities; digital projection radiography (DPR)<sup>42–44</sup> was used in three studies and MRI<sup>45,46</sup> in two studies. Two studies<sup>47,48</sup> using vertebral fracture assessment (VFA) images were considered in this review. In total, 75% of studies were conducted in Asia, with China (13/40) and South Korea (9/40) contributing most to these studies. The included studies mainly relied on unstructured data (i.e., imaging) as the original input. Five studies<sup>18,19,24,43,49</sup> combined the original image input with clinical covariates, which were structured data. Some studies used images as raw input; a series of structured feature data was obtained for DL.

Overall, more complex DL algorithms have been proposed to achieve better performance through the use of different hyperparameter tuning methods. In recent years, data pre-processing and data augmentation techniques have been applied to enhance the reliability of the results and reduce the risk of overfitting. We observed a trend in research moving from single-center studies to multi-center studies, which indicates stronger generalization capabilities and potential clinical applications.

### Studies on osteoporosis screening

Twenty studies<sup>18,19,24,26–29,34,42,44,50–52</sup> focused on osteoporosis screening and classification. The convolutional neural network (CNN) was widely used in all of these studies. Osteoporosis was identified based on opportunistic imaging from CT, X-ray, DPR, and MRI, although MRI is not recommended for opportunistic osteoporosis screening. The incorporation of clinical covariates into the model



**Figure 1.** Flow chart of the literature selection using PubMed and Web of Science.

construction process has resulted in improved model construction, with an improvement in the AUC by 2%–4%. Furthermore, the use of pre-trained models, pre-data augmentation, and image standardization has enabled studies with relatively limited data to achieve good performance. Visualization of models was performed using gradient-weighted class activation mapping (Grad-CAM). One study showed a perfect match with bone fracture,<sup>51</sup> but another study did not.<sup>18</sup>

Among the studies on osteoporosis, it is worth noting that several studies<sup>14,35,38</sup> used U-Net for image segmentation. U-Net is a well-known CNN semantic segmentation algorithm. Liu et al.<sup>14</sup> proposed the use of

U-Net for diagnosing osteoporosis. They compared the performance of U-Net with a back propagation network and support vector machine (SVM), and U-Net demonstrated the best performance. The results indicated that the U-Net algorithm could better distinguish between normal bone mass and osteoporosis, achieving an AUC of 0.872. However, U-Net had poorer recognition of lower bone mass and osteoporosis. The proposed U-Net model effectively dealt with image interference, thereby improving accuracy.

U-Net is commonly included in automatic systems for medical image segmentation. However, in some cases, the segmentation performance falls below

**Table 1.** Characteristics of the included studies.

Author	Country	Year	Objective	Material	Dataset	Algorithm	Result
Yamamoto et al. <sup>18</sup>	Japan	2020	Osteoporosis classification	X-ray	Total: 1223 patients	ResNet-18, ResNet-34, GoogleNet, EfficientNet-b3, EfficientNet-b4 U-net	ACC: 0.89 Recall: 0.899 F1 score: 0.89 AUC: 0.94 ACC: 0.96 Sensitivity: 0.96 Specificity: 0.92 F1-score: 0.98 AUC: 0.91 ACC: 0.81
Zhang et al. <sup>35</sup>	China	2023	Osteoporosis classification	CT	Total: 1048 patients Training vs. validation vs. test dataset = 5:1:4		
Yamamoto et al. <sup>19</sup>	Japan	2021	Osteoporosis classification	X-ray	Total: 1699 patients	ResNet-18, 34, 50, 101, and 152	AUC: 0.91
Liu et al. <sup>14</sup>	China	2019	Osteoporosis diagnosis	X-ray	Total: 89 patients	BP network, SVM, U-net	AUC: 0.89
Zhang et al. <sup>15</sup>	China	2020	Osteoporosis diagnosis	X-ray	Training and internal validation: 910 patients Test dataset 1: 198 patients Test dataset 2: 147 patients Total: 240 patients	DCNN	AUC: 0.77 Sensitivity: 0.74
Wani et al. <sup>16</sup>	India	2022	Osteoporosis diagnosis	X-ray	Total: 240 patients	ResNet-18, AlexNet, VGG-16, VGG-19	ACC: 0.91
Lee et al. <sup>44</sup>	South Korea	2018	Osteoporosis diagnosis	DPR	Total: 1268 patients Training and validation: 1068 Test: 200 patients	MC-DCNN SC-DCNN	ACC: 0.99 AUC: 1.00 F1-score: 0.99
Jang et al. <sup>50</sup>	South Korea	2022	Osteoporosis diagnosis	X-ray	Total: 1089 chest radiographs Training vs. validation vs. test = 7:1:2	CNN	AUC: 0.91 ACC: 0.82 Sensitivity: 0.84 Specificity: 0.82 AUROC: 0.93
Hong et al. <sup>20</sup>	South Korea	2023	Osteoporosis diagnosis	X-ray	Total: 9276 patients	EfficientNet-B4	AUC: 0.86 ACC: 0.84
Lee et al. <sup>42</sup>	South Korea	2020	Osteoporosis diagnosis	DPR	Total: 680 patients Test dataset: 20%	CNN-3 VGG-16 VGG-16_TF_FT VGG-16_TF	Sensitivity: 0.90 Specificity: 0.82 r > 0.98
Fang et al. <sup>29</sup>	China	2020	Osteoporosis diagnosis	CT	Total: 1449 patients Training: 586 patients Test: 863 patients	DenseNet-121 U-net	

(continued)

Table 1. Continued.

Author	Country	Year	Objective	Material	Dataset	Algorithm	Result
Löffler et al. <sup>37</sup>	Germany	2021	Osteoporosis diagnosis	CT	Total: 192 patients	CNN	AUC: 0.89
Chen et al. <sup>27</sup>	Taiwan (China)	2023	Osteoporosis diagnosis	CT	Total: 197 External validation: 397	ResNet-50, SVM	AUC: 0.98 ACC: 0.94 Sensitivity: 0.95 Specificity: 0.93 AUC: 0.94
Mao et al. <sup>24</sup>	China	2022	Osteoporosis diagnosis	X-ray	Total: 5652 patients Training vs. validation vs. test set 1 vs. test set 2 = 8:1:1	DenseNet	Sensitivity: 0.75 Specificity: 0.92 AUC: 0.93
Zhao et al. <sup>45</sup>	China	2022	Osteoporosis diagnosis	MRI	Training: 142 patients Validation: 64 patients External validation: 25 patients	U-Net, LASSO	Sensitivity: 0.92 Specificity: 0.82
Jang et al. <sup>51</sup>	South Korea	2022	Osteoporosis diagnosis	X-ray	Total: 1001 patients Training vs. validation vs. Test = 8:1:1 External validation: 117 patients	NLNN	Accuracy: 0.81 Sensitivity: 0.91 Specificity: 0.69 PPV: 0.79 NPV: 0.86 AUC: 0.87
Sukegawa et al. <sup>43</sup>	Japan	2022	Osteoporosis diagnosis	DPR	Total: 778 patients	EfficientNet-b0, b3, and b7 ResNet-18, 50, and 152	Accuracy: 0.85 Specificity: 0.89 AUC: 0.92
Pickhardt et al. <sup>38</sup>	USA	2022	Osteoporosis diagnosis	CT	Total: 11035 patients	TernausNet	AUC: 0.93 Sensitivity: 0.94 Specificity: 0.84 AUROC: 0.86
Tariq et al. <sup>34</sup>	USA	2022	Osteoporosis diagnosis	CT	Total: 6083 images Prospective test group: 344 patients	DenseNet-121, RF	AUC: 0.98 ACC: 0.95 TPR: 0.96 TNR: 0.95 AUC: 0.93 Sensitivity: 0.86 Specificity: 1.00
Dzierzak et al. <sup>28</sup>	Poland	2022	Osteoporosis diagnosis	CT	Total: 100 patients Training vs. validation vs. test = 2:1:1	VGG-16, VGG-19, MobileNetV2, Xception, ResNet-50, InceptionResNetV2 U-net	AUC: 0.98 ACC: 0.95 TPR: 0.96 TNR: 0.95 AUC: 0.93 Sensitivity: 0.86 Specificity: 1.00
Pan et al. <sup>30</sup>	China	2020	BMD prediction	CT	Total: 374 patients	U-net	AUC: 0.93 Sensitivity: 0.86 Specificity: 1.00

(continued)

**Table 1.** Continued.

Author	Country	Year	Objective	Material	Dataset	Algorithm	Result
Nguyen et al. <sup>21</sup>	South Korea	2021	BMD prediction	X-ray	Total: 330 patients (660 hip X-ray images) Training: 510 images Test dataset: 150 images Total: 213 patients Training: 150 patients Test: 63 patients	VGGNet	Coefficient: 0.81
Tang et al. <sup>33</sup>	China	2020	BMD prediction	CT		Net DenseNet	AUC: 0.92 ACC: 0.77
Sato et al. <sup>17</sup>	Japan	2022	BMD prediction	X-ray	Total: 10,102 patients Training vs. validation vs. test = 7:2:1	ResNet-50	AUC: 0.84 Accuracy: 0.78 Sensitivity: 0.77 Specificity: 0.79 ACC: 0.98
Rühling et al. <sup>31</sup>	Germany	2021	BMD prediction	CT	Total: 193 patients Training vs. test = 8:2	2D DenseNet, 3D DenseNet	Sensitivity: 0.98 Specificity: 0.99 AUC: 0.86
Sollmann et al. <sup>22</sup>	Germany	2022	BMD prediction	CT	Total: 144 patients	CNN	Sensitivity: 0.92
Uemura et al. <sup>36</sup>	Japan	2022	BMD prediction	CT	Total: 75 patients	U-net	Specificity: 1.00 ACC: 0.75
Breit et al. <sup>26</sup>	Switzerland	2023	BMD prediction	CT	Total: 109 patients	D2iN	AUC: 0.80 Sensitivity: 0.93 Specificity: 0.61
Yasaka et al. <sup>39</sup> Ho et al. <sup>23</sup>	Japan Taiwan (China)	2020 2021	BMD prediction BMD prediction	CT X-ray	Total: 183 patients Total: 3472 images	CNN ResNet-18	AUC: 0.97 Sensitivity: 0.76 Specificity: 0.87 $r^2$ : 0.84
Hsieh et al. <sup>53</sup>	Taiwan (China)	2021	BMD prediction	X-ray	Total: 10,197 hip radiographs, 25,482 spine radiographs Hip testing set: 5164 images Spine testing set: 18,175 images External validation: pelvis: 2060 images, spine: 3346 images	PelviXNet-34, DAG, VGG-11, VGG-16, ResNet-18, ResNet-34	RMSE: 0.06 AUROC: 0.97 AUPRC: 0.89 Accuracy: 0.92 Sensitivity: 0.80 Specificity: 0.95 AUC: 0.74
Lee et al. <sup>52</sup>	South Korea	2019	BMD prediction	X-ray	Total: 334 patients Training-validation set vs. test set = 7:3	AlexNet, VGGNet, Inception-V3, ResNet-50, KNINC, SVMC, RFC	ACC: 0.71 Sensitivity: 0.81 Specificity: 0.60

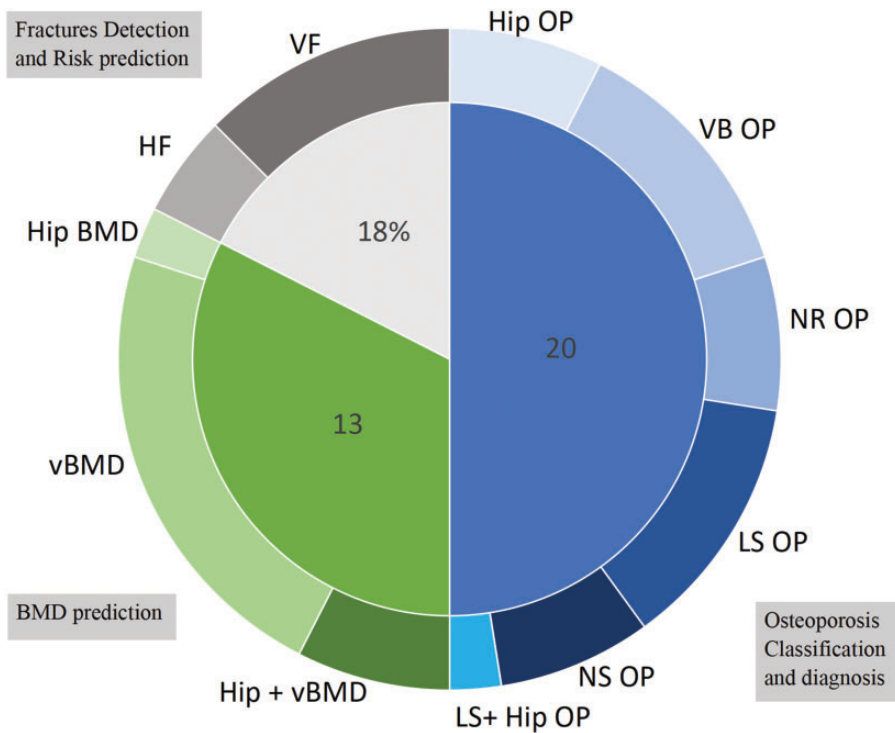
(continued)

Table 1. Continued.

Author	Country	Year	Objective	Material	Dataset	Algorithm	Result
Kang et al. <sup>40</sup>	South Korea	2023	BMD prediction	CT	Total: 547 patients (2696 CT images) Training: 2239 images Test: 457 images Total: 120 patients	U-net, CNN	ACC: 0.862 Sensitivity: 0.897 Specificity: 0.827
Du et al. <sup>22</sup>	China	2022	Fracture risk	X-ray		SVM, RF, GBDT, AdaBoost, ANIN, XGBoost, R2U-Net, DeepSurv	ACC: 0.96 Recall: 1.00
Kong et al. <sup>49</sup>	South Korea	2022	Fracture risk	X-ray	Total: 1595 participants Training: 1416 participants Test: 179 participants		C-index values: 0.612 95% CI: 0.571 – 0.653
Yabu et al. <sup>46</sup>	Japan	2021	Fracture detection	MRI	Total: 814 patients (1624 slices)	VGG16, VGG19, DenseNet-201, ResNet-50	AUC: 0.949 ACC: 0.88
Xiao et al. <sup>25</sup>	Hong Kong (China)	2022	Fracture detection	X-ray	Total: 6674 cases Training: 5970 cases Test: 704 cases	Not mentioned	Sensitivity: 0.881 Specificity: 0.879 ACC: 0.939
Tomita et al. <sup>41</sup>	USA	2018	Fracture detection	CT	Total: 1432 cases Training: 1168 cases Validation: 135 cases	ResNet-34, LSTM	Sensitivity: 0.971 Specificity: 0.892
Derkatch et al. <sup>48</sup>	Canada	2019	Fracture detection	VFA	Adjudicated test set: 129 cases Total: 12742 patients Training vs. validation vs. test = 6:1:3	InceptionResNetV2, DenseNet	Sensitivity: 0.852 Specificity: 0.958 F1 score: 0.908 AUC: 0.94
Monchka et al. <sup>47</sup>	Canada	2021	Fracture detection	VFA	Total: 12742 images Training vs. validation vs. test = 6:1:3	InceptionResNetV2, DenseNet	Sensitivity: 0.874 Specificity: 0.884 AUC: 0.95 Sensitivity: 0.824 Specificity: 0.943

BMD, bone mineral density; RF, random forest; RFC, random forest classifier; SVM, support vector machine; SVMC, support vector machine classifier; KNNC, k-nearest neighbor classifier; DPR, digital projection radiography; DAG, directed acyclic graph; LASSO, least absolute shrinkage and selection operator; CNN, convolutional neural network; ANIN, artificial neural network; CT, computed tomography; DXA, dual-energy X-ray absorptiometry; r, Pearson correlation coefficient; BP network, back propagation network; D2I2IN, deep image-to-image network; SC-DCNN: single-column deep convolutional neural network; MC-DCNN: multicolored deep convolutional neural network; GBDT, gradient boosting decision tree; NLNN, non-local neural network; ACC, accuracy; AUC, area under curve; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision recall curve; TPR, true positive rate; TNR, true negative rate; PPV, positive predictive value; NPV, negative predictive value; RMSE, root mean square error; VFA, vertebral fracture assessment; MRI, magnetic resonance imaging.





**Figure 2.** Study objectives related to osteoporosis included in this review. Osteoporosis classification and diagnosis was divided into lumbar spine (LS), vertebral body (VB), hip OP (osteoporosis), lumbar spine and hip OP, non-standard assessment technique (NS), or not reported (NR). Fracture detection and risk prediction was performed for the vertebrae (VF) and hip (HF). BMD prediction was conducted for hip BMD, vertebral BMD (vBMD), and hip and vBMD.

expectations. This can be attributed to factors such as the complex structures of vertebral bodies, label noise, semantic segmentation, and variations in CT scanners. Pickhardt et al.<sup>38</sup> developed TeraNet, a variant of U-Net with a VGG-11 encoder. Among the CNN-based studies, the development of an automatic system for fully automated detection of osteoporosis has potential clinical application and commercial value. Zhang et al.<sup>35</sup> proposed an end-to-end multitask joint learning framework that integrated positioning, segmentation, and classification. In this framework, U-Net has a prerequisite role. The framework achieved the best performance for diagnosing osteoporosis (ACC: 0.957, sensitivity: 0.962, specificity: 0.922,

F1-score: 0.975) with a learning rate of  $10^{-3}$ , surpassing state-of-the-art DL models.

Transfer learning has gained popularity in the field of DL owing to its ability to leverage knowledge from previous tasks, improving generalization and achieving more accurate outcomes with limited data input. Wani et al.<sup>16</sup> used transfer learning to develop four CNN networks: AlexNet, ResNet, VggNet-16, and VggNet-19. To mitigate the risk of overfitting, data augmentation techniques were applied, considering the limited amount of data available. The results showed that the pre-trained CNN networks consistently outperformed the normal CNN networks in all cases. Notably, the pre-trained AlexNet achieved

the best classification performance with an accuracy of 91%, surpassing the performance of the normal AlexNet by 15.3%. The classification accuracy for ResNet, VggNet-16, and VggNet-19 was 74.3%, 78.9%, and 73.68% for the normal CNN and 86.4%, 86.3%, and 84.2% for the pre-trained CNN, respectively.

In another study by Lee et al.,<sup>42</sup> transfer learning and fine-tuning were used to build transfer learning models based on VGG-16 (VGG-16\_TF), and further fine-tuning was performed with the transfer learning model (VGG-16\_TF\_FT). To prevent overfitting, a five-fold cross-validation approach was followed. The VGG-16\_TF and VGG-16\_TF\_FT models outperformed the VGG-16 and CNN-3 models, achieving an AUC of 0.858 and 0.782, respectively. These results highlight the important contribution of transfer learning and fine-tuning in improving the performance of DL models. By combining a pre-trained CNN network with fine-tuning, the screening performance using small image datasets is comparable to that of previous studies.

Dzierzak et al.<sup>28</sup> used six pre-trained deep convolutional neural network (DCNN) architectures for model construction. Among these models, the VGG-16 model demonstrated superior performance, achieving an AUC of 0.985 and an ACC of 95%. Additionally, three models, (VGG-16, VGG-19, and InceptionResNetV2) achieved an ACC of over 90%. The remarkable improvement in performance can be attributed to the application of pre-training and fine-tuning techniques, which have proven to be highly effective despite the limited amount of available data. Notably, with its relatively shallow architecture consisting of 19 layers, VGG-16 emerged as the most effective model for handling small datasets.

In a study conducted by Lee et al.,<sup>53</sup> a combination of VGG-16 and random forest classifier (RFC) yielded the best overall performance, with an AUC of 0.74, ACC of

0.71, sensitivity of 0.81, and specificity of 0.6. These findings further support the notion that VGG-16 is a popular and influential model in the field. With its 13 convolutional layers and three fully connected layers, VGG-16 offers a compact architecture. The use of small  $3 \times 3$  convolutional layer filters allows for a reduction in parameters, resulting in decreased computational effort and shorter model-construction time.

In 2020, Yamamoto et al.<sup>18</sup> developed an ensemble model by incorporating clinical covariates with hip radiographic images. This combined model outperformed those using only radiographic images or clinical covariates. The ensemble EfficientNet-b3 network was the most successful among the five ensemble models, surpassing the EfficientNet-b3 network that relied solely on images. Those authors used guided Grad-CAM for visualization of model classification. However, the heatmap was not accurately distributed to the proximal femoral trabeculae but rather to the exterior femoral bone, indicating that osteoporosis location was not adequately addressed. However, in the study by Jang et al.,<sup>51</sup> the visualization results were closely aligned with the proximal femur structure. In the following year, Yamamoto et al.<sup>19</sup> pursued a related study, which combined image features with patient variables and statistically analyzed the differentiation caused by the inclusion of these variables. Whereas the study showed performance improvements, some CNN models did not show similar improvements. The effect size, a measure of the true effect size of an experiment or the strength of the association of variables, was evaluated in the model, yielding a value of 0.871, indicating a large effect size. Osteoporosis screening was also conducted using DPR and six ensemble CNN models, including EfficientNet-b0, b3, and b7 and ResNet-18, 50, and 152. The EfficientNet-b7 and ResNet-152 models demonstrated comparable performance. However, the

visualization produced by Grad-CAM varied, with ResNet focusing on the cortical bone of the lower edge of the mandible, and EfficientNet extending its focus to the area above the cortical bone.

Hsieh and colleagues<sup>53</sup> reported that image-only models of VGG-16 and ResNet-34 performed well, and the addition of sex and age did not enhance model performance. The extent to which clinical characteristics can aid in performance improvement remains a topic for future investigation. Some studies have also explored the impact of CT channel on ML outcomes, in addition to considering clinical covariates.

Multi-center studies conducted by both Mao et al.<sup>24</sup> and Zhang et al.<sup>15</sup> aimed to explore performance improvements associated with the incorporation of CT channel and clinical covariates. The combination of anteroposterior and lateral channels was superior to each single channel. Mao's study achieved higher AUC and sensitivity in diagnosing osteoporosis compared with Zhang's study, potentially owing to data imbalance and the absence of cortical bone in X-ray images. Both studies highlighted potential biases in BMD measurements using DXA and the CNN models owing to factors such as aortic sclerosis, bowel gas, and osteophytic spurs. Tariq et al.<sup>34</sup> developed a fusion model, which combined coronal and sagittal images from enhanced CT scans with basic patient information to screen for low bone density. This model achieved a high area under the receiver operating characteristic curve of 0.86, indicating its potential in clinical application. The model performed well in a prospective group with and without the use of contrast material and demonstrated limited disparities when trained and validated on four sites with different CT scanners.

In their research, Chen et al.<sup>27</sup> developed a two-level classifier that combines ML techniques with radiomics texture analysis.

The proposed classifier used SVM binary classifiers, which were trained using nested cross-validation. This approach involved five outer and inner iterations to optimize the hyperparameters. The vertebral body segmentation model achieved a Sorenson–Dice coefficient of 0.87, which is consistent with the findings of Pan et al.<sup>30</sup> Furthermore, the two-level classifier demonstrated excellent performance, with an AUC of 0.98, ACC of 0.94, sensitivity of 0.95, and specificity of 0.93.

Jang et al.<sup>51</sup> addressed the limitations of CNNs, which primarily focus on local features rather than global features, by developing a nonlocal neural network (NLNN) based on the VGG-16 architecture. The NLNN model demonstrated promising results, achieving an overall ACC of 81.2%, sensitivity of 91.1%, and specificity of 68.9%. Notably, the Grad-CAM visualization technique revealed that the NLNN model accurately captured the bone structure, further validating its effectiveness in analyzing global features.

### *Studies on BMD prediction*

A total of 13 studies<sup>17,21,23,26,30–33,36,39,53,54</sup> aimed to measure BMD. Among these, nine studies focused on predicting vertebral BMD,<sup>17,26,30–33,39,40,53</sup> three studies assessed hip BMD,<sup>21,23,36</sup> and only one study examined both vertebral and hip BMD.<sup>54</sup> It is worth noting that whereas most studies used quantitative assessment methods, one study<sup>33</sup> conducted qualitative assessment.

Pan et al.<sup>30</sup> used U-Net for fully automated segmentation and labeling of vertebral bodies, which is a crucial step in BMD prediction. The developed system demonstrated a strong correlation with BMD measurements, with a coefficient of determination ( $r^2$ ) ranging from 0.964 to 0.968. The system also exhibited high diagnostic performance for osteoporosis, with an

AUC of 0.927, sensitivity of 85.71%, and specificity of 99.68%.

In another study by Fang et al.,<sup>29</sup> U-Net was applied for automated segmentation of vertebral bodies, and DenseNet-121 was used to calculate BMD. The results from three independent testing cohorts showed that the average Dice similarity coefficients for L1–L4 vertebral bodies were consistently near or above 0.8. Moreover, the BMD values calculated using the CNN exhibited a strong correlation with BMD derived from DXA, with  $r > 0.98$ .

Breit et al.<sup>26</sup> used a deep image-to-image network (DI2IN) in their study. The proposed DI2IN is a multi-layer CNN that incorporates wavelet features and geometric constraints (AdaBoost) to accurately segment the vertebral bodies of the thoracic spine. Cylindrical volumes of interest (VOIs) were placed at the center of the segmented vertebral bodies. Mean attenuation, measured in Hounsfield units (HU), was calculated to assess bone density. The study showed that the mean HU values for T1–12 were significantly higher in individuals with normal bone density compared with those who had osteopenia or osteoporosis. Furthermore, a statistically significant correlation was observed between the mean HU values of T1–12 and absolute bone density. The mean HU values demonstrated a higher ACC (0.75), AUC (0.8), Youden index (0.54), sensitivity (0.93), and specificity (0.61) compared with clinical reports. However, Kang et al.<sup>40</sup> found that not only the vertebral body but also other bone areas contributed to BMD estimation. Therefore, putting VOIs in the center of the vertebral bodies may not enhance performance, although the center of the segmented vertebral bodies received greater attention in the Gard-CAM of that study. Similarly, Yasaka et al.<sup>39</sup> used supervised learning to develop a CNN model using axial CT images as input. The CNN

model estimated BMD with AUCs of 0.965 and 0.970 for the internal and external validation datasets, respectively, outperforming BMD measurements derived from CT values of the lumbar vertebrae. However, a limitation of that study was the use of unenhanced CT. Sollmann et al.<sup>32</sup> established an automated CNN model for segmentation, labeling, and extraction of vertebral bone mineral density (vBMD) from QCT scans. The study showed that vBMD derived from QCT (vBMD\_QCT) exhibited a stronger correlation with opportunistic CNN-based vBMD derived from routine CT data (vBMD\_OPP) than with noncalibrated HU values. Discrimination between patients with and without osteoporotic vertebral fractures was slightly better when using vBMD\_OPP compared with vBMD\_QCT or noncalibrated HU values. Considering the scanner-specific nature and susceptibility to contrast media, noncalibrated HU values were not recommended in this study. To address the bias introduced by contrast material, Rühling et al.<sup>31</sup> used three artificial neural network (ANN) models. Among these models, the 2D anatomy-guided DenseNet model aimed to minimize contrast medium-induced errors and achieved an ACC of 0.983, sensitivity of 0.983, and specificity of 0.991. Additionally, the 2D anatomy-guided DenseNet model, which used combined axial images from T8–L2, demonstrated excellent generalizability, performing well on the public dataset.

Tang et al.<sup>33</sup> developed a qualitative assessment method for diagnosing osteoporosis. The process involved extracting 2D CT image slices of the pedicle level in the lumbar vertebrae (L1) from 3D CT images. The lumbar vertebrae were then identified and segmented using MS-Net. Subsequently, qualitative classification of the lumbar vertebrae was performed using BMDC-net. The study results demonstrated

high precision of 80.57%, 66.21%, and 82.55% for predicting normal bone mass, low bone mass, and osteoporosis, respectively. However, it is worth noting that the recognition accuracy for low bone mass was significantly lower compared with that for normal bone mass and osteoporosis, indicating a challenge that still needs to be addressed. To visually represent the distribution of probabilities for the three categories, a pie chart was created, providing an overview of the bone condition. However, in cases where the probabilities of two categories are similar, providing a definitive clinical diagnosis becomes challenging. This highlights the need for further research and improvement in the diagnostic accuracy of the qualitative assessment method.

In the multi-center study by Sato et al.,<sup>17</sup> the proposed method uses chest X-ray, age, and sex to predict BMD and T-score. The predictive performances for true BMD are as follows: femoral BMD ( $r$ : 0.75,  $r^2$ : 0.54) and lumbar spine BMD ( $r$ : 0.63,  $r^2$ : 0.40). The predicted femoral BMD demonstrates an excellent correlation with BMD measured by DXA. This indicates that the model can accurately estimate the BMD values based on the provided inputs. Furthermore, the prediction for T-score, which is used to diagnose osteoporosis (T-score below  $-2.5$ ), exhibits strong performance with an AUC of 0.8. This suggests that the model can effectively identify individuals with osteoporosis based on their T-scores. In addition to the imaging data, clinical features such as age, height, and weight are normalized and included as input data for training an ensemble ANN. The model has achieved excellent performance in assessing BMD, with a high  $r$  value of 0.8075 ( $p < 0.0001$ ) when compared with DXA measurements. This indicates strong agreement between the predicted BMD values and BMD values obtained using DXA.

### *Studies on risk prediction and detection of osteoporotic fracture*

Two included articles concerned the risk prediction of osteoporotic fractures. Du et al.<sup>22</sup> first segmented the femur region of interest and the hard bone edge. They then conducted a quantitative analysis of the femoral trabecular score and femoral neck strength to assess the quality of the femoral neck. A risk prediction ML-based model of femoral neck osteoporosis in older adults was established, which makes sense in auxiliary diagnosis. In the test dataset, the neural network model showed the highest ACC of 95.83% and a recall of 100%.

Kong et al.<sup>49</sup> used DL techniques to demonstrate superior performance compared with traditional fracture risk assessment tools such as the FRAX and Cox proportional hazard (CoxPH) models. The DL model, specifically the DeepSurv model, achieved a higher C-index value of 0.612 compared with FRAX (0.547) and CoxPH (0.594). Notably, the authors found that using multiple slices in the spine region with DL yielded even better results than using a single slice. This highlights the importance of considering a broader range of information when using DL algorithms for fracture risk assessment. The findings of the above study suggest that DL models have the potential to achieve performance that is comparable to that of FRAX, even when using only a single X-ray image as input. This indicates the clinical potential of DL in improving fracture risk assessment and potentially enhancing patient care.

Another five studies were aimed at detecting osteoporotic fractures. Osteoporotic vertebral fracture (OVF) is the most common type of osteoporotic fracture. Tomita<sup>41</sup> and colleagues used CNN for the extraction of features and achieved an ACC of 89.2% in detecting OVF. Xiao et al.<sup>25</sup> designed



automated vertebral fracture detection software based on lateral chest radiographs of older women. In a study by Yabu,<sup>46</sup> data augmentation techniques and ensemble learning were applied to distinguish fresh OVFs from MRI images. However, that study did not use any images from normal individuals or patients with pathological fractures, a notable limitation.

VFA via DXA can be used to detect a vertebral fracture. VFA can be conducted using both dual-energy (DE) and single-energy (SE) mode scanners. Monchka et al.<sup>47</sup> and Derkatch et al.<sup>48</sup> used DL to detect osteoporotic fracture from VFA images. Derkatch et al. trained and tested their model on DE only. Monchka et al. constructed their model on the basis of a DE training set, an SE training set, and a composite dataset comprising both DE and SE VFAs. The authors found that CNN models are highly sensitive to the scan mode. The model trained with DE images performed poorly in the SE test set. Training on the composite dataset allows CNN to generalize to both scan modes and enhance the performance in DE VFAs. The performance of this CNN is on par with that of Derkatch et al., with a similar AUC, slightly lower sensitivity, and higher specificity. Thus, a composite training dataset including not only DE VFA images but also SE images is highly recommended. Notably, the study by Monchka has some limitations in that they did not try to identify external validation sets. However, there are few studies toward this direction. Further research is needed to reveal the best solution.

## Discussion

In this review, we summarized previous reports focusing on DL models for the radiologic diagnosis of osteoporosis. Our analysis is primarily centered around imaging methods commonly used in clinical

settings, such as X-ray, CT, and DXA. These methods offer a cost-effective approach to opportunistic osteoporosis screening.

Overall, the DL models discussed in the reviewed studies have demonstrated favorable results in both classification and regression tasks. To mitigate the risk of overfitting, researchers have used various techniques such as data augmentation, increasing the dataset size, and regularization methods.

It is worth noting that only one study, that of Lee et al. in 2018,<sup>44</sup> reported an AUC exceeding 0.99, indicating a potential risk of overfitting. When evaluating regression tasks, the Pearson correlation coefficient ( $r$ ) is commonly used as a performance metric. Additionally, some studies also present the mean absolute error to further assess the accuracy of the regression models.

However, it is important to consider the issue of class imbalance when interpreting accuracy results. Owing to the disproportionate representation of classes, accuracy alone may overestimate performance by favoring the majority class. Therefore, it is crucial to consider alternative evaluation metrics, such as AUC, that account for class imbalance and provide a more comprehensive assessment of model performance.

To evaluate the reliability of a model, sample size matters. To train a state-of-art DL model, hundreds to millions of samples are critical. Among the studies included in the review, the sample size ranged from 75 to 35,679. With respect to data collection, the main barrier is the lack of a large sample size and distinct annotations of training data. In 2017, Litjens et al.<sup>54</sup> conducted a survey of 308 studies, with most studies showing a preference for pre-trained CNN models to extract features. Because the pre-trained model is accessible on some websites and shows extremely good performance with a small amount of data,

it is the main choice when handling a small data set. Lee and colleagues<sup>42</sup> used transfer learning to develop several models, among which fine-tuning was proven highly successful. We also observed that some studies chose databases from other countries and regions as the test set, achieving a relatively good result and good generalizability. However, access to a labeled database that has accurate outcome for a certain disease remains a challenge.

Achieving accurate outcomes in disease prediction can be difficult in the field of health care. Whereas advancements in technology and ML techniques have improved the accuracy of predictions, certain factors can still affect the reliability of the results. One challenge in disease prediction is the availability and quality of data. Obtaining a large and diverse dataset that accurately represents the target population can be difficult. Additionally, the presence of confounding factors and individual variations within the population can affect the accuracy of predictions. Furthermore, the complexity and heterogeneity of diseases can make accurate prediction challenging. Diseases often have multiple contributing factors and can manifest differently in different individuals. This variability makes it difficult to develop a one-size-fits-all prediction model.

To address these challenges, ongoing research focuses on refining ML algorithms, incorporating more comprehensive and diverse datasets, and considering additional clinical and genetic factors. Collaborative efforts between researchers, health care professionals, and data scientists are crucial to improving the accuracy of disease prediction models.

Most studies included in this review underwent external validation, which enhances the reliability and generalizability of the findings. However, to further strengthen the predictive capabilities of the models, prospective cohort studies should

be conducted to closely monitor the impact of disease progression on prediction. This will provide valuable insights and enable the refinement of prediction models.

Whereas the studies in this review present certain challenges, they also provide valuable directions for future research. Notably, the inclusion of clinical covariates in ensemble models has demonstrated improvements in performance across various medical specialties. The incorporation of additional clinical characteristics into CNN model construction is a common practice. For instance, Tang et al.<sup>33,36</sup> reported the inclusion of electrocardiogram data as input for predicting atrial fibrillation recurrence, which yielded superior results compared with other models.

Moving forward, it is crucial to explore the types of clinical data input that can substantially enhance the prediction results, particularly those with practical implications for clinical treatment and prevention. Identifying and incorporating relevant clinical data in future efforts can contribute to the development of more accurate and impactful prediction models.

DL has demonstrated a remarkable capacity for osteoporosis screening. However, improving the quality of input requires enhancing the pre-processing of image output, managing class imbalance, and using large amounts of data. Sharing the data and code of these studies would be highly beneficial. Clinical commercialization of a diagnostic model for osteoporosis remains a challenge owing to the need for a high level of generalizability, low cost, and reliance on readily accessible medical images that are obtained daily.

### Author contributions

Xu Z. contributed to the design of the work. He Y. contributed to manuscript drafting. He Y. and Zhu S. contributed to data acquisition. Zhu J. evaluated the articles for eligibility. Lin J. reviewed and polished the manuscript.

All authors contributed to final approval of the completed version and all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work were appropriately investigated and resolved.

### Declaration of conflicting interest

The authors declare that there is no conflict of interest.

### Funding

This study was supported by the National Natural Science Foundation of China (82000540), Changzhou Science and Technology Program (CJ20210003, CJ20210005), Suzhou Health Committee Program (KJXW2019001), and Jiangsu Health Committee Program (Z2022077).

### ORCID iD

Jinzhou Zhu  <https://orcid.org/0000-0003-0544-9248>

### References

1. NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy. Osteoporosis prevention, diagnosis, and therapy. *JAMA* 2001; 285: 785–795.
2. Compston JE, McClung MR and Leslie WD. Osteoporosis. *Lancet* 2019; 393: 364–376.
3. Chandran M, Ebeling PR, Mitchell PJ, Executive Committee of the Asia Pacific Consortium on Osteoporosis (APCO), et al. Harmonization of Osteoporosis Guidelines: Paving the Way for Disrupting the Status Quo in Osteoporosis Management in the Asia Pacific. *J Bone Miner Res* 2022; 37: 608–615.
4. Link TM. Osteoporosis imaging: state of the art and advanced imaging. *Radiology* 2012; 263: 3–17.
5. Kanis JA. Diagnosis of osteoporosis and assessment of fracture risk. *Lancet* 2002; 359: 1929–1936.
6. Dimai HP. Use of dual-energy X-ray absorptiometry (DXA) for diagnosis and fracture risk assessment; WHO-criteria, T- and Z-score, and reference databases. *Bone* 2017; 104: 39–43.
7. Salzman SN, Shirahata T, Yang J, et al. 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* 2019; 19: 695–702.
8. Carey JJ, Chih-Hsing Wu P and Bergin D. Risk assessment tools for osteoporosis and fractures in 2022. *Best Pract Res Clin Rheumatol* 2022; 36: 101775.
9. Lane NE. Epidemiology, etiology, and diagnosis of osteoporosis. *Am J Obstet Gynecol* 2006; 194: S3–S11.
10. Kanis JA, Harvey NC, Johansson H, et al. FRAX Update. *J Clin Densitom* 2017; 20: 360–367.
11. Kanis JA, Cooper C, Rizzoli R, on behalf of the Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis (ESCEO) and the Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (IOF), et al. European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporos Int* 2019; 30: 3–44.
12. Rajkomar A, Dean J and Kohane I. Machine Learning in Medicine. Reply. *N Engl J Med* 2019; 380: 2589–2590.
13. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; 372: n71.
14. Liu J, Wang J, Ruan W, et al. Diagnostic and Gradation Model of Osteoporosis Based on Improved Deep U-Net Network. *J Med Syst* 2019; 44: 15.
15. Zhang B, Yu K, Ning Z, et al. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. *Bone* 2020; 140: 115561.
16. Wani IM and Arora S. Osteoporosis diagnosis in knee X-rays by transfer learning based on convolution neural network. *Multimed Tools Appl* 2023; 82: 14193–14217.
17. Sato Y, Yamamoto N, Inagaki N, et al. Deep Learning for Bone Mineral Density



- and T-Score Prediction from Chest X-rays: A Multicenter Study. *Biomedicines* 2022; 10: 2323.
18. Yamamoto N, Sukegawa S, Kitamura A, et al. Deep Learning for Osteoporosis Classification Using Hip Radiographs and Patient Clinical Covariates. *Biomolecules* 2020; 10: 1534.
  19. Yamamoto N, Sukegawa S, Yamashita K, et al. Effect of Patient Clinical Variables in Osteoporosis Classification Using Hip X-rays in Deep Learning Analysis. *Medicina (Kaunas)* 2021; 57: 846.
  20. Hong N, Cho SW, Shin S, et al. Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography. *J Bone Miner Res* 2023; 38: 887–895.
  21. Nguyen TP, Chae DS, Park SJ, et al. A novel approach for evaluating bone mineral density of hips based on Sobel gradient-based map of radiographs utilizing convolutional neural network. *Comput Biol Med* 2021; 132: 104298.
  22. Du J, Wang J, Gai X, et al. Application of intelligent X-ray image analysis in risk assessment of osteoporotic fracture of femoral neck in the elderly. *Math Biosci Eng* 2023; 20: 879–893.
  23. Ho CS, Chen YP, Fan TY, et al. Application of deep learning neural network in predicting bone mineral density from plain X-ray radiography. *Arch Osteoporos* 2021; 16: 153.
  24. Mao L, Xia Z, Pan L, et al. Deep learning for screening primary osteopenia and osteoporosis using spine radiographs and patient clinical covariates in a Chinese population. *Front Endocrinol (Lausanne)* 2022; 13: 971877.
  25. Xiao BH, Zhu MSY, Du EZ, et al. A software program for automated compressive vertebral fracture detection on elderly women's lateral chest radiograph: Ofeye 1.0. *Quant Imaging Med Surg* 2022; 12: 4259–4271.
  26. Breit HC, Varga-Szemes A, Schoepf UJ, et al. CNN-based evaluation of bone density improves diagnostic performance to detect osteopenia and osteoporosis in patients with non-contrast chest CT examinations. *Eur J Radiol* 2023; 161: 110728.
  27. Chen YC, Li YT, Kuo PC, et al. Automatic segmentation and radiomic texture analysis for osteoporosis screening using chest low-dose computed tomography. *Eur Radiol* 2023; 33: 5097–5106.
  28. Dzierżak R and Omiotek Z. Application of Deep Convolutional Neural Networks in the Diagnosis of Osteoporosis. *Sensors (Basel)* 2022; 22: 8189.
  29. Fang Y, Li W, Chen X, et al. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. *Eur Radiol* 2021; 31: 1831–1842.
  30. Pan Y, Shi D, Wang H, et al. Automatic opportunistic osteoporosis screening using low-dose chest computed tomography scans obtained for lung cancer screening. *Eur Radiol* 2020; 30: 4107–4116.
  31. Rühling S, Navarro F, Sekuboyina A, et al. Automated detection of the contrast phase in MDCT by an artificial neural network improves the accuracy of opportunistic bone mineral density measurements. *Eur Radiol* 2022; 32: 1465–1474.
  32. Sollmann N, Löffler MT, El Hussein M, et al. Automated Opportunistic Osteoporosis Screening in Routine Computed Tomography of the Spine: Comparison With Dedicated Quantitative CT. *J Bone Miner Res* 2022; 37: 1287–1296.
  33. Tang C, Zhang W, Li H, et al. CNN-based qualitative detection of bone mineral density via diagnostic CT slices for osteoporosis screening. *Osteoporos Int* 2021; 32: 971–979.
  34. Tariq A, Patel BN, Sensakovic WF, et al. Opportunistic screening for low bone density using abdominopelvic computed tomography scans. *Med Phys* 2023; 50: 4296–4307.
  35. Zhang K, Lin P, Pan J, et al. End to End Multitask Joint Learning Model for Osteoporosis Classification in CT Images. *Comput Intell Neurosci* 2023; 2023: 3018320.
  36. Uemura K, Otake Y, Takao M, et al. Development of an open-source measurement system to assess the areal bone mineral density of the proximal femur from clinical CT images. *Arch Osteoporos* 2022; 17: 17.
  37. Löffler MT, Jacob A, Scharr A, et al. Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral

- fractures compared to DXA. *Eur Radiol* 2021; 31: 6069–6077.
38. Pickhardt PJ, Nguyen T, Perez AA, et al. Improved CT-based Osteoporosis Assessment with a Fully Automated Deep Learning Tool. *Radiol Artif Intell* 2022; 4: e220042.
  39. Yasaka K, Akai H, Kunimatsu A, et al. Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. *Eur Radiol* 2020; 30: 3549–3557.
  40. Kang JW, Park C, Lee DE, et al. Prediction of bone mineral density in CT using deep learning with explainability. *Front Physiol* 2022; 13: 1061911.
  41. Tomita N, Cheung YY and Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med* 2018; 98: 8–15.
  42. Lee KS, Jung SK, Ryu JJ, et al. Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs. *J Clin Med* 2020; 9: 392.
  43. Sukegawa S, Fujimura A, Taguchi A, et al. Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates. *Sci Rep* 2022; 12: 6088.
  44. Lee JS, Adhikari S, Liu L, et al. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. *Dentomaxillofac Radiol* 2019; 48: 20170344.
  45. Zhao Y, Zhao T, Chen S, et al. Fully automated radiomic screening pipeline for osteoporosis and abnormal bone density with a deep learning-based segmentation using a short lumbar mDixon sequence. *Quant Imaging Med Surg* 2022; 12: 1198–1213.
  46. Yabu A, Hoshino M, Tabuchi H, et al. Using artificial intelligence to diagnose fresh osteoporotic vertebral fractures on magnetic resonance images. *Spine J* 2021; 21: 1652–1658.
  47. Monchka BA, Kimelman D, Lix LM, et al. Feasibility of a generalized convolutional neural network for automated identification of vertebral compression fractures: The Manitoba Bone Mineral Density Registry. *Bone* 2021; 150: 116017.
  48. Derkatch S, Kirby C, Kimelman D, et al. Identification of Vertebral Fractures by Convolutional Neural Networks to Predict Nonvertebral and Hip Fractures: A Registry-based Cohort Study of Dual X-ray Absorptiometry. *Radiology* 2019; 293: 405–411.
  49. Kong SH, Lee JW, Bae BU, et al. Development of a Spine X-Ray-Based Fracture Prediction Model Using a Deep Learning Algorithm. *Endocrinol Metab (Seoul)* 2022; 37: 674–683.
  50. Jang M, Kim M, Bae SJ, et al. Opportunistic Osteoporosis Screening Using Chest Radiographs With Deep Learning: Development and External Validation With a Cohort Dataset. *J Bone Miner Res* 2022; 37: 369–377.
  51. Jang R, Choi JH, Kim N, et al. Prediction of osteoporosis from simple hip radiography using deep learning algorithm. *Sci Rep* 2021; 11: 19997.
  52. Lee S, Choe EK, Kang HY, et al. The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population. *Skeletal Radiol* 2020; 49: 613–618.
  53. Hsieh CI, Zheng K, Lin C, et al. Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nat Commun* 2021; 12: 5472.
  54. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; 42: 60–88.