

Review

Deep Learning–Assisted Automated Diagnosis of Osteoporosis Based on Computed Tomography Scans: Systematic Review and Meta-Analysis

Aobo Wang*, MD; Ziqian Ma*, MD; Tianyi Wang, MD; Ruiyuan Chen, MM; Yu Xi, MM; Qichao Wu, MD; Shuo Yuan, MD; Ning Fan, MD; Peng Du, MM; Lei Zang, MD

Department of Orthopedics, Beijing Chaoyang Hospital, Capital Medical University, Beijing, China

*these authors contributed equally

Corresponding Author:

Lei Zang, MD

Department of Orthopedics, Beijing Chaoyang Hospital

Capital Medical University

5 JingYuan Road, Shijingshan District

Beijing 100043

China

Phone: 86 51718268

Email: zanglei@ccmu.edu.cn

Abstract

Background: Osteoporosis is a prevalent skeletal disorder characterized by decreased bone mass and increased fracture risk; however, it frequently remains underdiagnosed due to limited health care resources and its asymptomatic progression. Deep learning (DL) provides a promising solution for automated screening using computed tomography (CT) scans, enabling earlier detection and improved management.

Objective: This systematic review and meta-analysis aimed to investigate the diagnostic performance of DL models in diagnosing osteoporosis based on CT scans.

Methods: This study was conducted under the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines using articles extracted from PubMed, Scopus, Web of Science (Core), and Embase (Ovid). Studies involving adult participants who underwent CT and in which DL was applied for osteoporosis diagnosis were included. The QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool was used to estimate the risk of bias in each study. The confusion matrices from the included studies were extracted to summarize the diagnostic performance of DL models for osteoporosis. Within a bivariate random-effects framework, sensitivity and specificity were jointly synthesized to yield the summary estimates. Heterogeneity was quantified with Higgins I^2 statistics. Subgroup analyses were performed to explore potential sources of heterogeneity among the included studies.

Results: This review included 24 studies, encompassing CT images from 29,808 participants. All studies used conventional CT scans and used DL-based architectures. Fifteen, 6, and 3 studies were assessed as having a low, uncertain, and high risk of bias, respectively. The meta-analysis included 20 studies. The pooled sensitivity and specificity were 0.88 (95% CI 0.85-0.91; $I^2=83.69\%$) and 0.94 (95% CI 0.91-0.96; $I^2=95.07\%$) for osteoporosis diagnosis; 0.81 (95% CI 0.76-0.85; $I^2=82.38\%$) and 0.92 (95% CI 0.90-0.94; $I^2=79.05\%$) for osteopenia identification; and 0.95 (95% CI 0.92-0.97; $I^2=98.28\%$) and 0.93 (95% CI 0.91-0.95; $I^2=94.93\%$) for normal case identification. The area under the curve of the DL models for identifying osteoporosis, osteopenia, and normal cases was 0.96 (95% CI 0.93-0.97), 0.94 (95% CI 0.92-0.96), and 0.98 (95% CI 0.96-0.99), respectively. Subgroup analyses revealed that models based on DenseNet variants ($P<.01$), multislice input ($P<.01$), 3D architecture ($P<.01$), and CT as the reference standard ($P<.01$) demonstrated superior diagnostic performance.

Conclusions: This study indicated that CT-based DL models achieve promising diagnostic performance for osteoporosis. However, substantial heterogeneity among the included studies, limited external validation, and incomplete end-to-end pipelines constrain the generalizability of the proposed models. Further research is warranted to support their clinical translation and standardized application.

Trial Registration: PROSPERO CRD42024601713; <https://www.crd.york.ac.uk/> and <https://www.crd.york.ac.uk/PROSPERO/view/CRD42024601713>

J Med Internet Res 2025;27:e77155; doi: [10.2196/77155](https://doi.org/10.2196/77155)

Keywords: deep learning; artificial intelligence; osteoporosis; computed tomography; meta-analysis

Introduction

Osteoporosis is a systemic skeletal disorder characterized by low bone mass, deteriorated bone microarchitecture, and increased bone fragility [1]. It affects individuals across diverse ethnic backgrounds, with postmenopausal women being particularly vulnerable. The global prevalence of osteoporosis is estimated at 19.7%, increasing to over 25% among individuals older than 60 years [2], and is expected to increase further with the aging population. In addition, a considerable proportion of individuals are diagnosed with osteopenia, a condition considered a precursor to osteoporosis [3].

Osteoporosis is frequently referred to as a “silent disease” because of its lack of noticeable symptoms in the early stages. However, patients may experience chronic pain and functional impairment as the disease progresses. The most prevalent and severe consequence of osteoporosis is fragility fractures, which significantly reduce patients’ quality of life. In particular, hip fractures are associated with increased mortality. The incidence of osteoporosis-related fractures in China is estimated to reach 4 million by 2030, resulting in over \$20 billion in health care costs [4]. Considering the large, affected population, early diagnosis and treatment of osteoporosis are crucial. Interventions such as calcium and vitamin D supplementation and fall prevention strategies have been shown to effectively reduce the risk of fractures [5].

The World Health Organization (WHO) recommends that a T-score of ≤ -2.5 measured using dual-energy X-ray absorptiometry (DXA) is indicative of osteoporosis [1]. In recent years, computed tomography (CT) has been recognized as a method that provides more precise and detailed trabecular bone information, making quantitative CT (QCT) a widely accepted diagnostic approach for osteoporosis [6]. However, osteoporosis remains significantly underdiagnosed in clinical practice despite the availability of well-defined diagnostic criteria. This is partly due to limited health care resources, with only 0.46 DXA machines available per million people on average in China [6], and partly due to its asymptomatic nature, with most patients being diagnosed only after a fracture event [7]. Therefore, developing a more accessible and widely applicable screening method for osteoporosis is essential.

The automated analysis of medical images is being actively developed and applied in clinical settings with the rapid advancement of artificial intelligence, particularly deep learning (DL) technologies. As a state-of-the-art approach, DL can directly process raw medical images and leverage large-scale data for training, thereby reducing the reliance on manual feature extraction required in traditional machine learning methods. Moreover, DL has demonstrated

superior performance in complex pattern recognition tasks, where subtle imaging features are difficult to predefine [8]. Compared with dedicated osteoporosis screening devices, DL enables the development of scalable screening approaches using existing medical images. For instance, routine chest and abdominal CT scans frequently contain sufficient bone information. DL can emulate the principles of QCT by leveraging vertebral imaging data to enable automated bone mineral density (BMD) assessment, thereby expanding the scope of opportunistic osteoporosis detection. Considering the substantial overlap between patients undergoing CT scans for other medical conditions and those at risk of osteoporosis, this approach may represent a cost-effective strategy for improving osteoporosis management.

However, DL technologies for osteoporosis assessment remain in the developmental and validation phases. Evidence from existing studies must be synthesized to assess the progress in this field and to identify the gaps between model development and clinical application. Therefore, a systematic review and meta-analysis were conducted. This study primarily aimed to evaluate the performance of DL technology in diagnosing osteoporosis and osteopenia based on CT scans and, secondarily, to determine the potential factors affecting the capability of automated diagnosis.

Methods

Study Design and Registration

The review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and flowchart [9,10] and the PRISMA of Diagnostic Test Accuracy (PRISMA-DTA) checklist [11]. The protocol for this systematic review was registered in the Prospective Register of Systematic Reviews (PROSPERO; CRD42024601713).

Search Strategy

A comprehensive and exhaustive search was conducted in PubMed, Scopus, Web of Science (Core Collection), and Embase (via Ovid) up to September 28, 2025, to determine relevant articles that used DL techniques for diagnosing osteoporosis based on CT images [12]. Further, the reference lists of the included articles were manually screened to identify additional eligible studies. The following terms were used for the PubMed search: “osteoporosis,” “computed tomography,” “deep learning,” and “neural networks, computer.” [Multimedia Appendix 1](#) provides the details of the search strategy.

Inclusion and Exclusion Criteria

The inclusion criteria for the studies were (1) full-text articles in peer-reviewed journals; (2) studies involving adult patients who underwent conventional CT scans for routine clinical indications, including chest, abdomen, lumbar spine, or pelvic CT, with or without contrast enhancement; (3) the use of DL methods for osteoporosis detection (classification) or bone density estimation (regression); (4) availability of test dataset information, particularly studies that reported or enabled reconstruction of a 2×2 or 3×3 confusion matrix based on sensitivity, specificity, precision, and recall.

The exclusion criteria were (1) non-English or non-peer-reviewed publications; (2) conference articles, preprints, reviews, letters, guidelines, editorials, or errata; (3) studies with fewer than 30 participants in either the training or test dataset; (4) DL used only applied to image segmentation, while traditional machine learning, radiomics, or HU-to-BMD conversion formulas were used for osteoporosis detection or bone density estimation; (5) studies used specific, nonconventional CT scanning protocols to acquire the input images for the DL model, such as QCT or dual-energy CT protocols (while QCT can serve as the reference standard for osteoporosis diagnosis, the input data for all included DL models were derived from routine CT scans rather than dedicated QCT images).

Review Process

Two reviewers (ZM and AW) independently performed the initial screening of the titles and abstracts of the included articles to identify potential eligibility after removing duplicates with EndNote (Clarivate). The full texts of the remaining articles were then reviewed, and those not meeting the inclusion criteria were excluded from the study. Any discrepancies were resolved through discussion or adjudicated by a third reviewer (LZ) when necessary. An email was sent to the corresponding authors for the acquisition of the necessary data for studies included in the systematic review but lacking sufficient data for meta-analysis.

Quality Assessment

The risk of bias and applicability were assessed with the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2), a tool designed to evaluate the quality of primary diagnostic accuracy studies [13]. The QUADAS-2 criteria were used to examine the risk of bias in 4 domains: Patient Selection, evaluating whether the studies reported general characteristics of the cohorts used for model development and whether the selection of participants was appropriate; Index Test, assessing whether the design and implementation of the models contained any obvious flaws, whether an independent test set was used, and whether the model outputs involved any manual intervention; Reference Standard, examining whether the diagnostic criteria for osteoporosis adhered to internationally recommended guidelines; and Flow and Timing, considering whether the study procedures were appropriate and whether there were excessive delays between the index test and the reference standard. Each domain was assessed in

terms of the risk of bias, and the first 3 were assessed in terms of concerns regarding applicability.

Data Extraction

Data extraction was independently conducted by 2 reviewers (AW and ZM) following the PRISMA-DTA guidelines [11]. Discrepancies were resolved through discussion or adjudication by a third reviewer (LZ). Variables from 4 key aspects were extracted and documented: (1) study design, including the first author, publication year, country of authors, number of participants, participant demographics, testing strategy, and reference standard; (2) characteristics of CT imaging, including the type of CT scans, CT vendor, acquisition parameters, scan region, scan plane, and target vertebrae; (3) details of DL algorithms, including name or architecture of the proposed model, network dimensionality, region of interest (ROI) acquisition method, Dice similarity coefficient of automated segmentation, type of input data, and whether end-to-end processing was performed; (4) diagnostic performance metrics, including confusion matrix (for articles using a 3×3 confusion matrix based on the diagnostic classification criteria for osteoporosis, the matrix was transformed into three 2×2 confusion matrices, each corresponding to a comparison between one category and other categories), sensitivity, specificity, area under the receiver operator characteristic curve, and whether the BMD values were predicted.

The meta-analysis involved both internal and external test results from all eligible studies. Each test result was considered an independent observation for studies reporting multiple test results based on different DL models, datasets, labeling, or data input strategies. A temporal test was considered an internal test because the samples were obtained from the same center and the same CT scanner [14]. Six studies assessed the diagnostic performance of multiple existing models. To prevent these studies from disproportionately affecting the pooled results, only the test results of ResNet and DenseNet for osteoporosis diagnosis were recorded in the literature [15-17], whereas the best-performing 2D or 3D ResNet and DenseNet models were documented in the literature [18], as these architectures are among the most commonly used and effective in medical image detection and classification, owing to their efficient training mechanisms and strong feature representation [19]. Two studies reported model performance on subdistrict datasets from the same center. The relevant results were not included as these do not qualify as standard internal or external tests [20,21].

Statistical Analysis

The MIDAS and METAN modules in Stata (version 17.0; StataCorp) software were used for statistical analyses [22]. These 2 modules can generate pooled results based on the confusion matrix reported in each included study and perform further statistical analyses. Forest plots were generated to visualize the pooled sensitivity and specificity of DL models in osteoporosis diagnosis. Considering the heterogeneity in study design, model architecture, and reference standards among the included studies, a bivariate random-effects model was used. The overall diagnostic accuracy was

assessed through summary receiver operator characteristic curve analysis. Heterogeneity between studies was assessed using the Q test and Higgins I^2 statistics, with the following classification thresholds: 0%-40%, 30%-60%, 50%-90%, and 75%-100% indicating negligible, moderate, substantial, and considerable heterogeneity, respectively.

Subsequently, further analyses were conducted to evaluate the diagnostic performance of DL models specifically for osteoporosis. A Fagan nomogram was applied to estimate posttest probabilities, facilitating clinical interpretation of model performance. The likelihood ratio (LR) dot plots were stratified into 4 quadrants according to predefined evidence strength thresholds, guiding the decision-making process for model exclusion or confirmation. The Deeks funnel plot symmetry test was used to assess publication bias. Subgroup analyses were conducted to investigate potential sources of heterogeneity across studies. A $P < .05$ indicated statistical significance.

Ethical Considerations

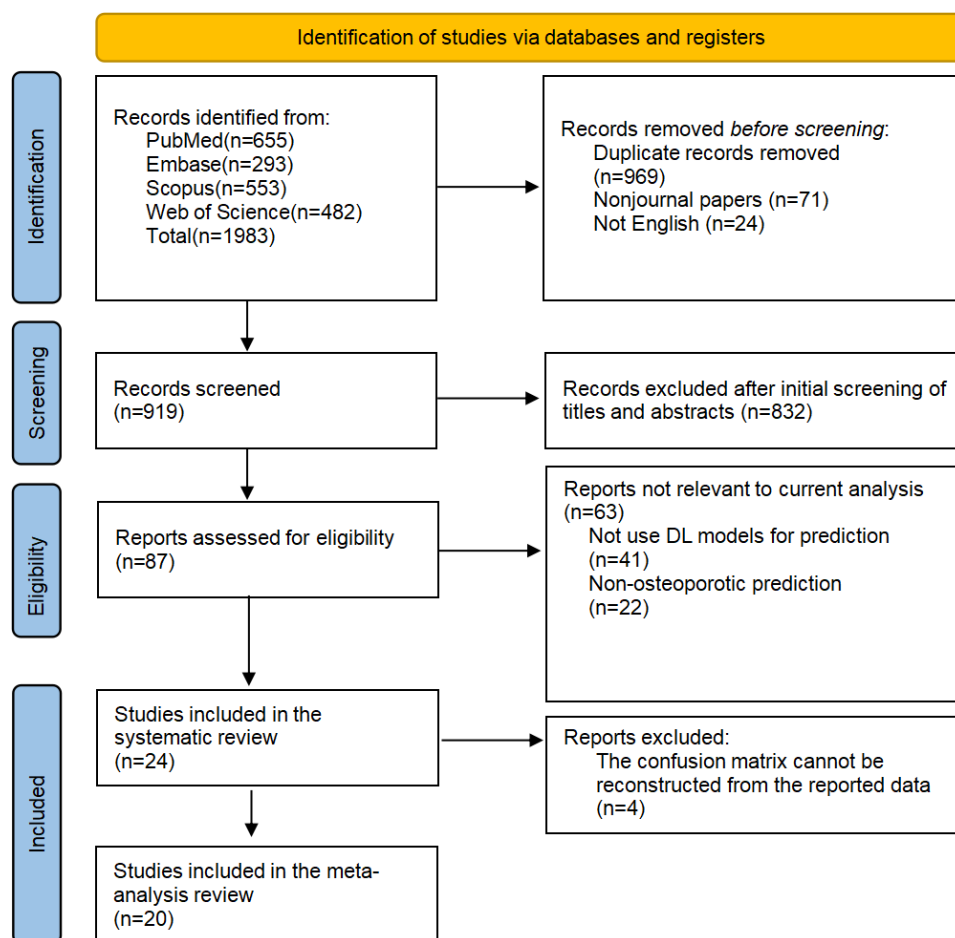
Ethical approval and informed consent were not required from the participants considering the nature of the systematic review and meta-analysis.

Results

Study Overview

The literature search was performed according to the protocol outlined in the PRISMA flowchart (Figure 1). The initial search across various databases yielded 1983 studies, including 655, 293, 553, and 482 from PubMed, Embase, Scopus, and Web of Science, respectively. After removing 969 duplicates, 71 nonjournal articles, and 24 non-English publications, 919 publications remained for screening. Of these, 832 were excluded based on titles and abstracts due to lack of relevance. Further, 87 studies were reviewed for full text, of which 24 were included in the systematic review [14-18,20,21,23-39], and 20 were ultimately incorporated in the quantitative meta-analysis [14-16,18,23,25-29]. The other remaining 4 studies were excluded from the meta-analysis because the reported test set sample size and diagnostic parameters were insufficient to reconstruct a complete confusion matrix [24,33,34,39]. Attempts to contact the corresponding authors of the studies did not obtain the necessary data.

Figure 1. Flowchart depicting the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) search strategy. DL: deep learning.



The studies included populations from China [14,16-18,20,21,23-26,28,29,32,33,37-39], Japan [27,30], Korea [31,34], the

United States [35], Poland [15], and Cyprus [36] geographically. The 24 studies included in the systematic review were

published between 2020 and 2025, involving a total of 29,808 participants. Further, 20 studies used single-center data [14,16,18,20,21,23,24,26-29,31-36,38,39], whereas 4 studies used multicenter data [17,25,30,35]. The development, validation, and testing of the DL models included 34,908 samples. In 5 included studies, each participant contributed more than 1 sample, including CT scans from different dates, slices, or vertebral levels [15,31,32,35,36], leading to a total sample size larger than the number of patients. Significant variation

in sample sizes was observed, with test sets ranging from 45 [30] to 2867 [23]. The classification forms for osteoporosis diagnosis were summarized into three types: (1) osteoporosis versus osteopenia versus normal cases [14,16,20,21,23-27,29,30,32,33,37-39], (2) osteoporosis versus nonosteoporosis cases [15,17,18,28,34], and (3) low bone mass versus normal cases [31,35,36]. Table 1 shows the detailed information of the included studies.

Table 1. Characteristics of the included studies in the systematic review and meta-analysis.

Study	Data source	Population characteristics	Diagnostic classification	Age (years), mean (SD) or median (IQR)	Participants, n (% female) or %	Test strategy and sample size (n)	Reference standard	Deep learning model networks
Wu et al ^a [23]	Single-center	Adult, excluded: previous spinal surgery and spinal tumors	Osteoporosis vs osteopenia vs normal cases	54 (IQR 47-61)	7713 (46.50)	Internal: 975; external: 4401	QCT ^b	DenseNet-121
Wang et al [24]	Single-center	Adult, excluded: fractures, metal implants, severe degenerative changes, deformities, and spinal tumors	Osteoporosis vs osteopenia vs normal cases	65 (IQR 57-70)	2274 (44.40)	Internal: 267; temporal: 347	QCT	ResNet-18
Tong ^a et al [14]	Single-center	Adult, excluded: metal implants, spinal tumors, and abnormal vertebral morphology	Osteoporosis vs osteopenia vs normal cases	62.89 (SD 11.55)	687 (47.20)	Internal: 101	QCT	ResNet
Peng ^a et al [25]	Multicenter	Adult, excluded: previous spinal surgery, fractures, and tumors	Osteoporosis vs osteopenia vs normal cases	—	1219 (59.80)	Internal: 176; external: 340	QCT	DenseNet
Pan ^a et al [26]	Single-center	Adult, excluded: metal implants, severe degenerative changes, fractures, and deformity	Osteoporosis vs osteopenia vs normal cases	51 (SD 14.5)	1048 (42.30)	Internal: 418	QCT	ResNet-101
Zhang ^a et al [16]	Single-center	— ^c	Osteoporosis vs osteopenia vs normal cases	—	—	Internal: 418	QCT	Joint framework
Fang et al ^a [18]	Single-center	Adult, excluded: severe scoliosis, fractures, and implants	Osteoporosis vs nonosteoporosis cases	—	488 (71.80)	Internal: 96	DXA ^d	Multiple algorithms
Yoshida et al ^a [27]	Single-center	Adult, excluded: fractures, severe scoliosis, severe spondylosis, prior spinal surgery, and enhanced CT ^e	Osteoporosis vs osteopenia vs normal cases	—	402 (77.40)	Internal: 52	DXA	ResNet-50
Niu ^a et al [28]	Single-center	Adult, excluded: metal or bone cement implant, secondary osteoporosis, Schmorl's nodes, severe scoliosis, and vascular calcification.	Osteoporosis vs nonosteoporosis cases	66.03 (SD 9.71)	—	Internal: 100	QCT	DenseNet

Study	Data source	Population characteristics	Diagnostic classification	Age (years), mean (SD) or median (IQR)	Participants, n (%; female) or %	Test strategy and sample size (n)	Reference standard	Deep learning model networks
Dzierżak and Omiotek ^a [15]	Single-center	—	Osteoporosis vs nonosteoporosis cases	—	100 (59)	Internal: 100	HU ^f	VGG ^g -16
Fang et al ^a [29]	Single-center	Adult, excluded: secondary osteoporosis and hyperparathyroidism; fractures and implants	Osteoporosis vs osteopenia vs normal cases	53.8	45.60%	Internal: 398; external: 294	QCT	DenseNet-121
Yasaka et al ^a [30]	Multicenter	Adult, excluded: previous spinal surgery, severe scoliosis, fractures, and deformity	Osteoporosis vs osteopenia vs normal cases	—	278 (50.80)	Internal: 45; external: 50	DXA	CNN ^h
Kang ^a et al [31]	Single-center	Adult, the date gap between CT ^e with a complete L1 axial cut and DXA scan was <1 month	Low bone mass vs normal cases	—	—	Internal: 457	DXA	Residual CNN
Li et al ^a [32]	Single-center	Patients with cough and epigastric pain as the main symptoms, excluded: malformations, fractures, and abnormal bone metabolism	Osteoporosis vs osteopenia vs normal cases	—	801 (48.80)	Internal: 404	QCT	ResNet
Tang et al [33]	Single-center	Adult, excluded: metastases and compression fractures	Osteoporosis vs osteopenia vs normal cases	—	82%	Internal: 63	DXA	BMDC-Net
Oh et al [34]	Single-center	Adult who underwent routine cancer screening, excluded: implants, and z score >3.3 or <−3.3	Osteoporosis vs nonosteoporosis cases	58.86 (SD 12.56)	286 (54.40)	Internal: 98	DXA	DenseNet-169
Tariq et al [35]	Multicenter	Adult, excluded: implants and anatomical variations	Low bone mass vs normal cases	66.9 (SD 9.2)	65.30%	Internal: 1205	DXA	DenseNet-121
Küçükçiloğlu ^a et al [36]	Single-center	Adult, excluded: severe scoliosis or deformity, spondylarthrosis, inflammatory diseases, tumors, and previous spinal surgery	Low bone mass vs normal cases	—	100 (67)	Internal: 68	DXA	CNN
Zhou ^a et al [37]	Single-center	Adult, excluded: prior spinal surgery, vertebral fracture, and tumors	Osteoporosis vs osteopenia vs normal cases	47.7	46.1%	Internal: 137	QCT	Resnet-101
Kuo ^a et al [38]	Single-center	Adult, excluded: prior spinal surgery, vertebral fracture, and tumors	Osteoporosis vs osteopenia vs normal cases	—	507 (66.1)	Internal: 186	DXA	ViT-CNN

Study	Data source	Population characteristics	Diagnostic classification	Age (years), mean (SD) or median (IQR)	Participants, n (% female) or %	Test strategy and sample size (n)	Reference standard	Deep learning model networks
Li ^a et al [20]	Single-center	Adult, excluded: history of previous spinal surgery, severe compression fractures, and tumors	Osteoporosis vs osteopenia vs normal cases	55.11 (SD 13.72)	51.8%	Internal: 245 external: 258	QCT	DenseNet and ResNet
Li ^a et al [21]	Single-center	Adult, excluded: prior spinal surgery, vertebral fractures, tumors, implants, and BMI >35 kg/m ²	Osteoporosis vs osteopenia vs normal cases	54.07 (SD 9.90)	987 (59.9)	Internal: 112 external: 137	QCT	DenseNet and ResNet
Zhang et al [39]	Single-center	—	Osteoporosis vs osteopenia vs normal cases	—	—	Internal: 575	QCT	DeepmdQCT
Huang ^a et al [17]	Multicenter	Inclusion criteria: age ≥50 years, with complete medical records	Osteoporosis vs nonosteoporosis cases	—	1126 (68.7)	External: 545	DXA	Multiple algorithms

^aStudies included in meta-analysis (the confusion matrix was either directly provided in the literature or could be reconstructed based on the reported sample size and diagnostic performance metrics, such as sensitivity and specificity).

^bQCT: quantitative computed tomography.

^cNot available.

^dDXA: dual-energy X-ray absorptiometry.

^eCT: computed tomography.

^fHU: Hounsfield unit.

^gVGG: Visual Geometry Group.

^hCNN: convolutional neural network.

Characteristics of CT Images

The types of CT scans used in this systematic review consisted of routine CT [15,16,18,20,24-27,29-33], low-dose CT [14,20,21,23,25,28,37,38], and contrast-enhanced CT [34,35]. The scan regions included chest CT [14,16-18,20,21,23-26,28,31-33,37,38], abdominal CT [25,28-32,34,35], lumbar spine CT [15,20,21,25,29,31,36], and pelvic CT [35]. All used CT images included a complete display of the target vertebral body for osteoporosis diagnosis. Seven studies used sagittal images [20,21,23-25,27,36], 19 used axial images [14-18,20,21,26,28-39], and 3 incorporated coronal images [20,21,35]. Table S1 in [Multimedia Appendix 2](#) provides detailed information on the CT scans.

DL Model Characteristics

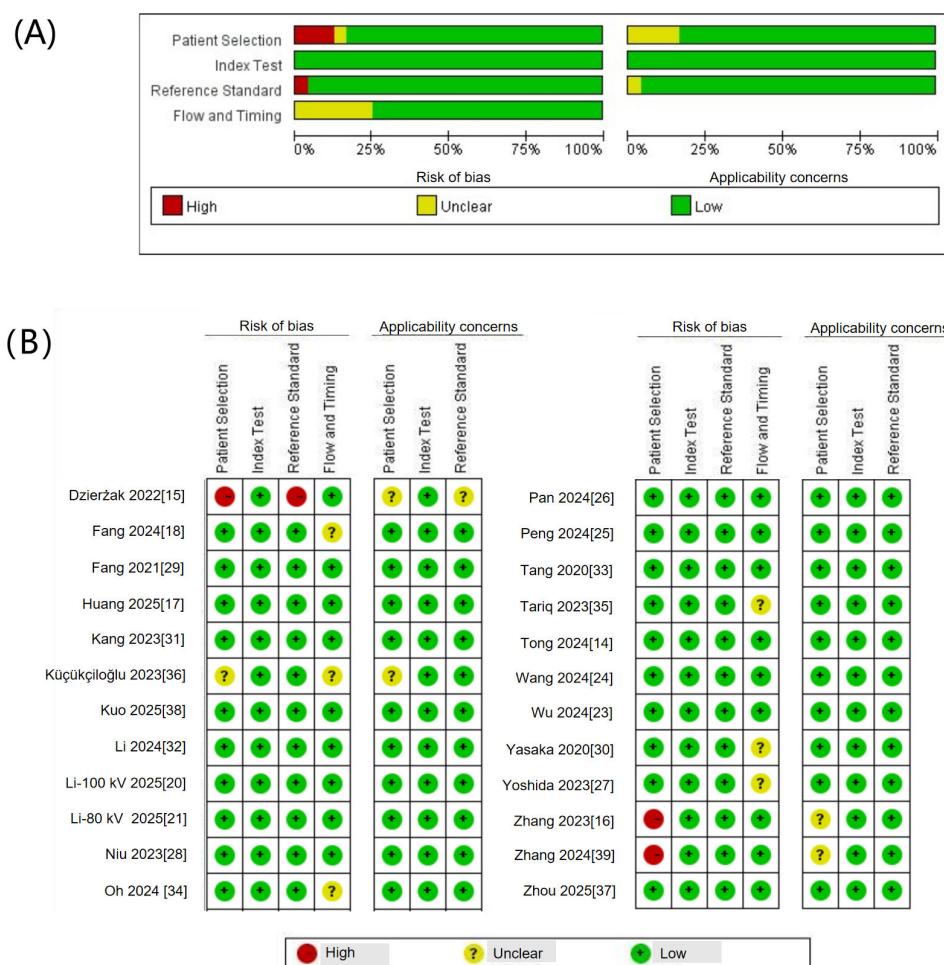
The most commonly used DL architectures in the included studies were ResNet variants [14,16-18,20,21,24,26,27,32,37] and DenseNet variants [17,20,21,23,25,28,29,33-35]. Seven studies used a 3D architecture to predict osteoporosis [14,18,20,21,23,28,31]. Automated segmentation was adopted in 18 studies for ROI localization, in contrast to manual segmentation in 5 studies [15,17,18,27,30]. An end-to-end approach was theoretically feasible in 12 studies among the 24 enrolled studies [14,20,21,23-25,28,29,32,34,35,38]. Osteoporosis was diagnosed directly in 11 studies [14-18,24,26,33,35,36,38]

and through BMD-based prediction in 13 studies [20,21,23,25,27-32,34,37,39]. Table S2 in [Multimedia Appendix 3](#) documents further details regarding the DL models.

Methodological Quality

Figure 2A provides an overview of the quality assessments of the included studies using the QUADAS-2 tool. Figure 2B provides the results of the nuanced analysis. For the risk of bias, 3 studies had a high risk of bias [15,16,39], 6 had an unclear risk of bias [18,27,30,34-36], and 15 had a low risk of bias [14,17,20,21,23-26,28,29,31-33,37,38]. Regarding the applicability, 4 studies had an unclear risk of concern [15,16,36,39]. Regarding the patient selection, 3 of the included studies did not report details of patient selection [15,16,39], and 1 was based on a potentially inappropriate patient population [36], causing a high and unclear bias, respectively. Regarding the index test, all studies enrolled demonstrated a low risk of bias. Regarding the reference standard, both DXA and QCT are recognized tools for osteoporosis assessment [40]. One study adopted HU values as the reference standard, contributing to a high risk of bias [15]. Regarding the flow and timing, 6 studies exhibited an unclear risk of bias due to either an excessive time interval between the index test and the reference standard (>3 months) [27,30,34-36] or insufficient information related to the image input [18].

Figure 2. Quality assessment by QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2). (A) The proportion of risk of bias and applicability concerns, and (B) summary of the risk of bias for each study. Green, yellow, and red circles indicate low, unclear, and high risk of bias, respectively.



Diagnostic Performance of the DL Models

Among the 20 studies included in the meta-analysis, 15 reported diagnostic performance based on multiple model architectures, input data, or test sets [15-18,20,21,25,26,29-31,35-38]. According to the principles described in the “Methods” section, 2-6 test results from each of these studies were treated as independent observations. Osteoporosis diagnosis involved 39 models. The pooled sensitivity and specificity of osteoporosis diagnosis were 0.88 (95% CI 0.85-0.91; $I^2=83.69\%$) and 0.94 (95% CI 0.91-0.96; $I^2=95.07\%$), respectively (Figure 3). Osteopenia identification involved 19 models, with pooled sensitivity and specificity of 0.81 (95% CI 0.76-0.85; $I^2=82.38\%$) and 0.92 (95% CI

0.90-0.94; $I^2=79.05\%$), respectively (Figure 4). Thirty models were included for the identification of normal cases. The pooled sensitivity and specificity for identifying normal cases were 0.95 (95% CI 0.92-0.97; $I^2=98.28\%$) and 0.93 (95% CI 0.91-0.95; $I^2=94.93\%$), respectively (Figure 5). Considering the substantial heterogeneity across studies, leave-one-out analyses were performed to assess the influence of individual studies on the pooled results. The findings are provided in Figures S1, S2, and S3 in Multimedia Appendices 4-6. The summary receiver operator characteristic curves indicate that the area under the receiver operator characteristic curve of the DL models for identifying osteoporosis, osteopenia, and normal cases was 0.96 (95% CI 0.93-0.97), 0.94 (95% CI 0.92-0.96), and 0.98 (95% CI 0.96-0.99), respectively (Figure 6).

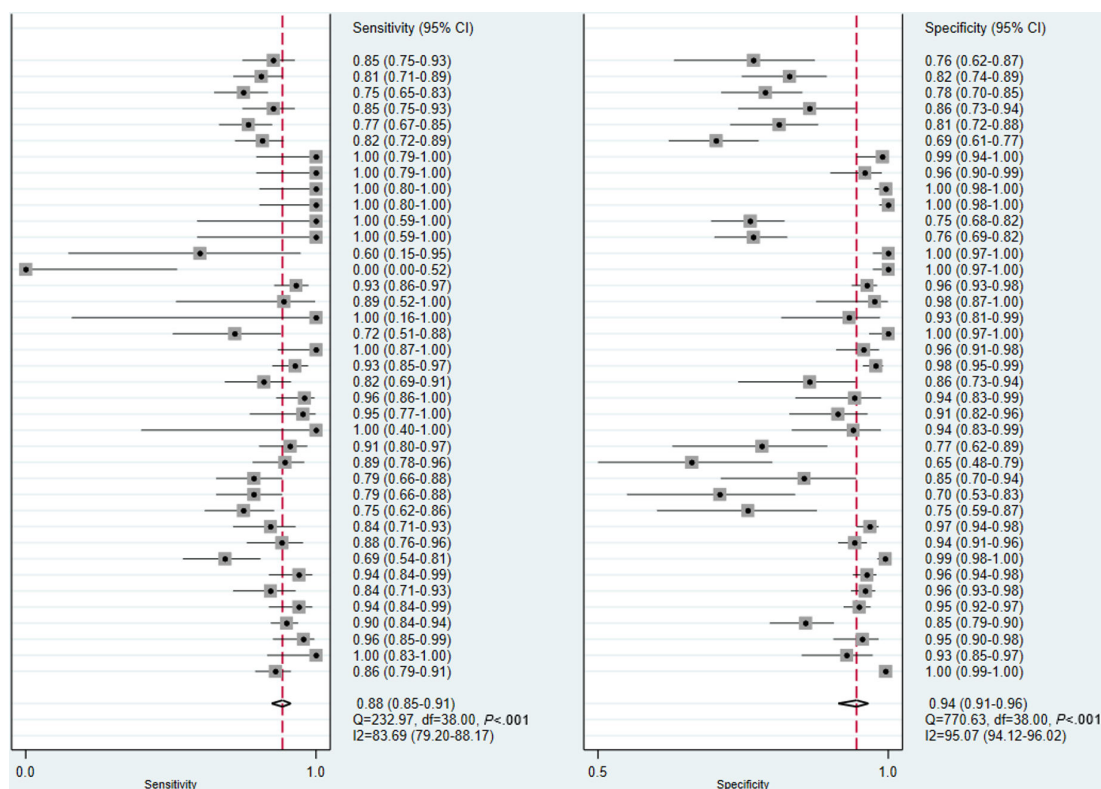
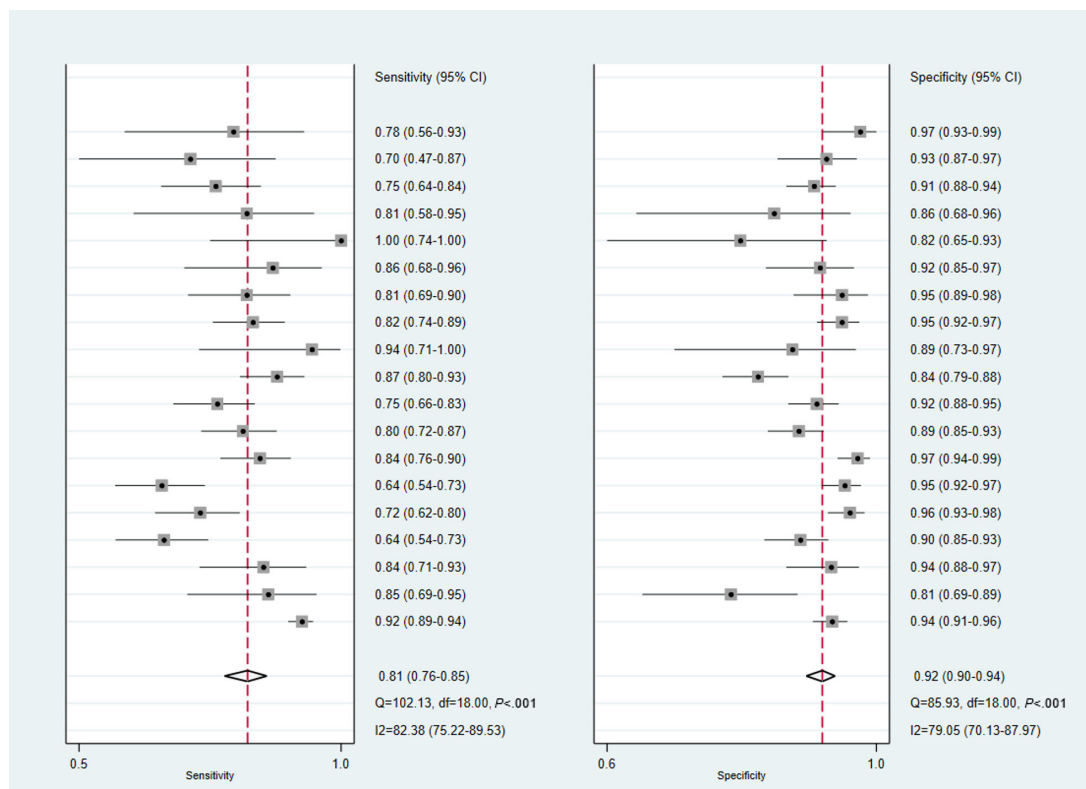
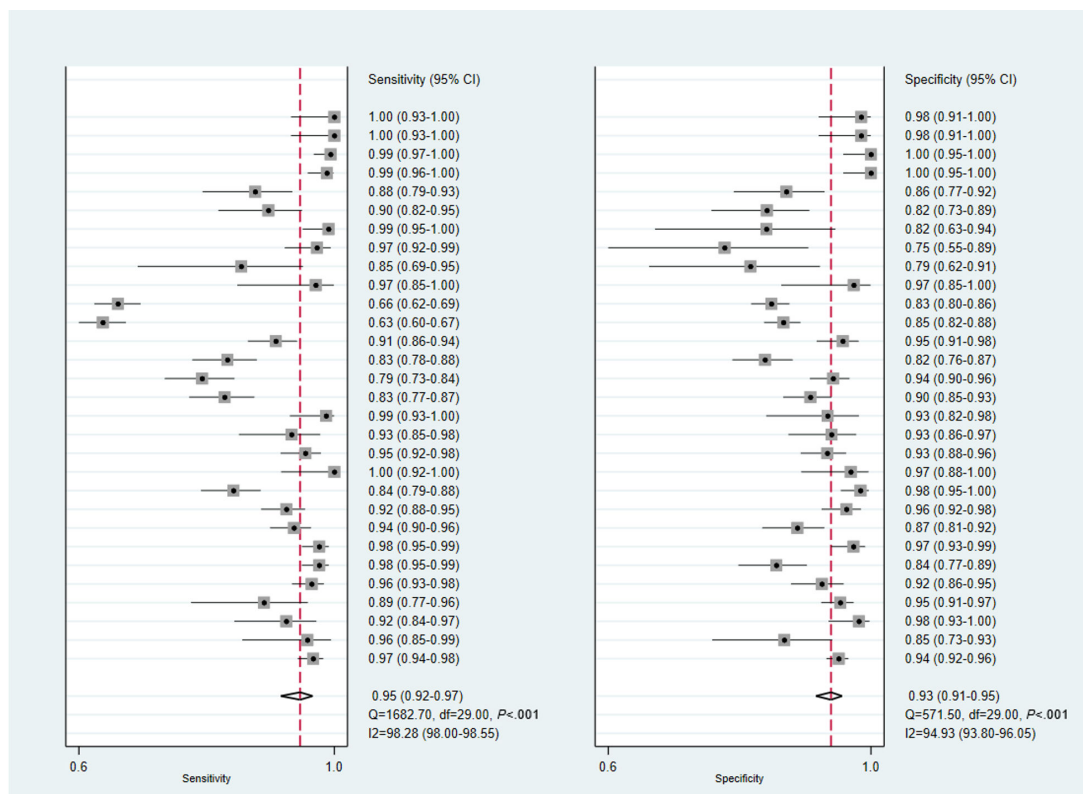
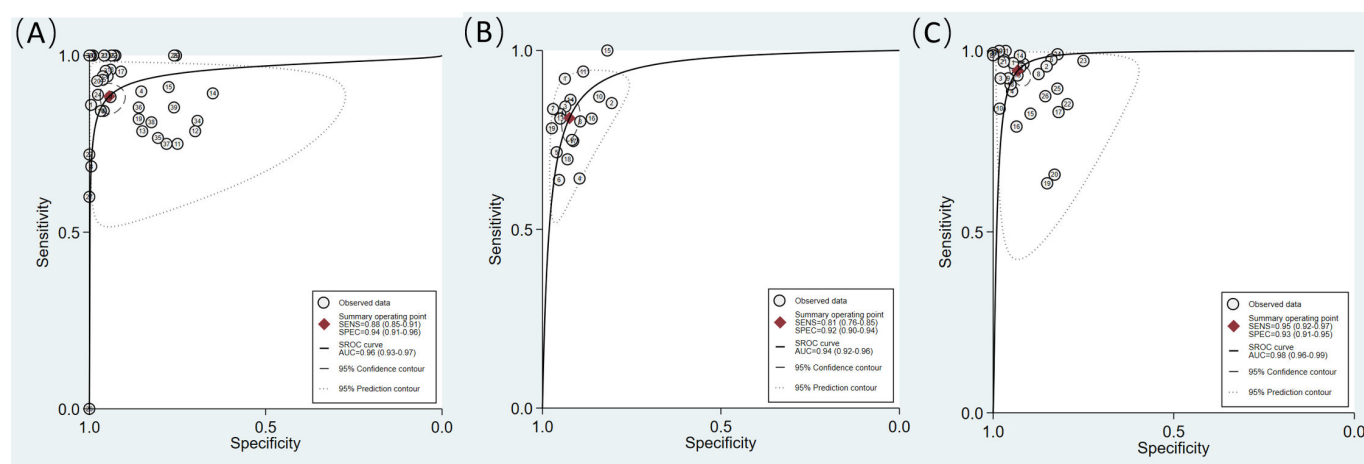
Figure 3. Forest plots in the sensitivity and specificity of deep learning (DL) models in diagnosing osteoporosis.**Figure 4.** Forest plots in the sensitivity and specificity of deep learning (DL) models in diagnosing osteopenia.

Figure 5. Forest plots in the sensitivity and specificity of deep learning (DL) models in identifying normal cases.**Figure 6.** The summary receiver operating characteristic (SROC) curves for deep learning (DL) models in the diagnosis of bone status categories. (A) Osteoporosis, (B) osteopenia, and (C) normal cases. AUC: area under the receiver operator characteristic curve; SENS: sensitivity; SPEC: specificity.

According to the prevalence of osteoporosis [2] and the distribution of patients in the included studies, the pretest probability in the Fagan nomogram was set at 20%. At this point, a positive test result of the DL model raises the post-test probability of osteoporosis to 80%, whereas a

negative result reduces it to 3% (Figure 7). It should be noted that post-test probabilities depend on the assumed pretest probability. Accordingly, the model's utility may differ across populations with different baseline prevalence.

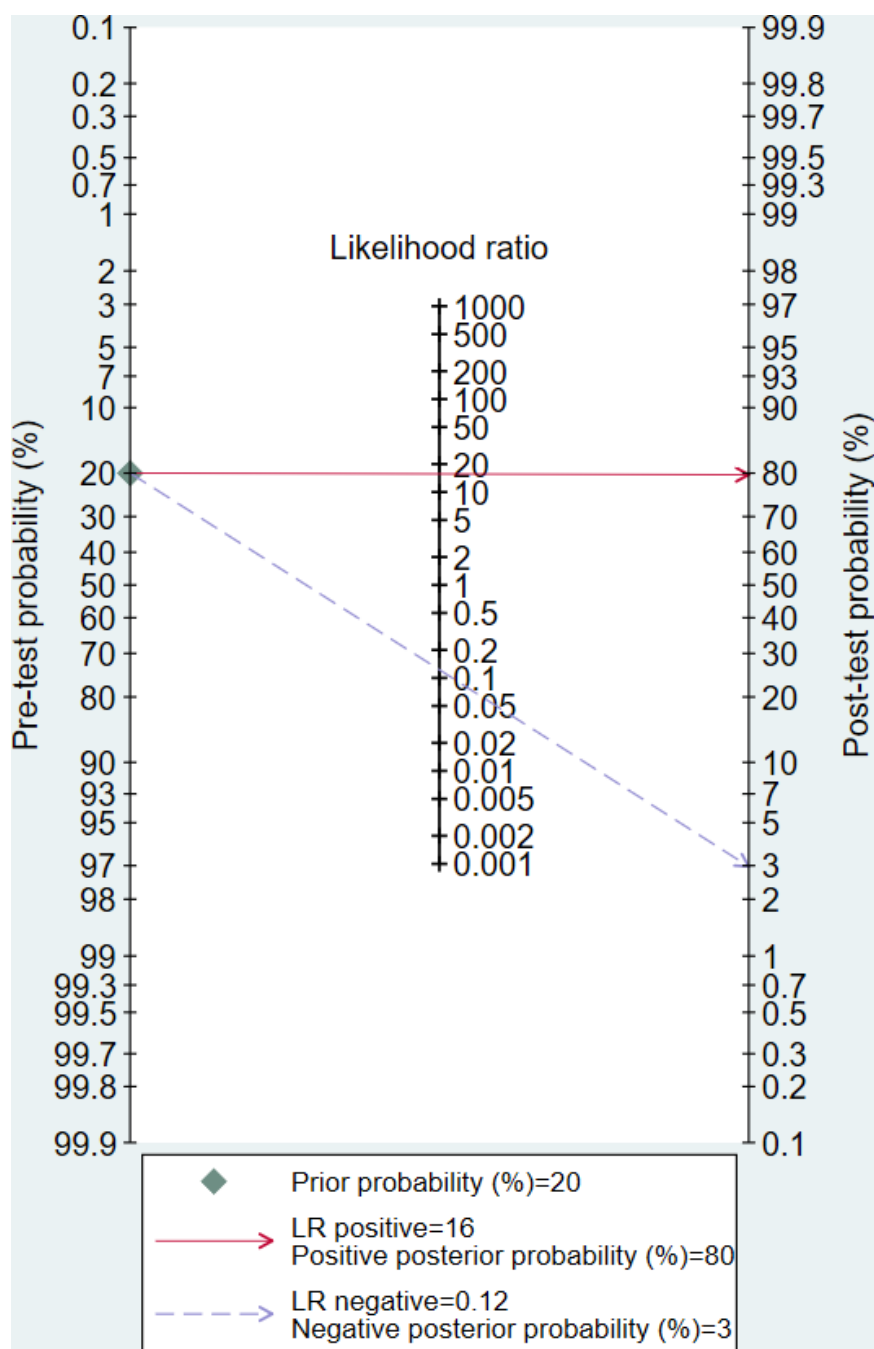
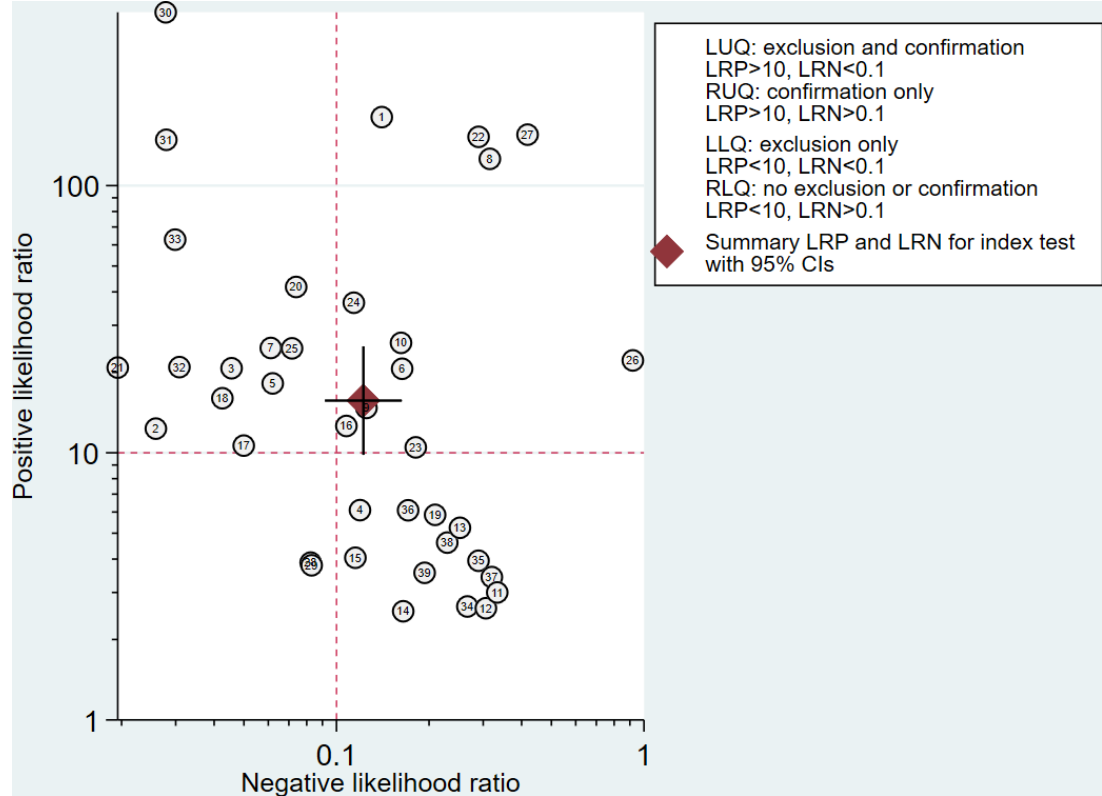
Figure 7. Fagan nomogram of deep learning (DL) models for diagnosing osteoporosis.

Figure 8 provides the LR scatter plots for osteoporosis diagnosis. Most data points are located in the upper left and upper right quadrants, with the summary LR plot of DL models positioned in the upper right quadrant, indicating the

value of DL models in confirming osteoporosis. However, several points fall in the lower right quadrant, demonstrating limited diagnostic use for certain models.

Figure 8. Likelihood ratio (LR) dot plot of deep learning (DL) models. LLQ: lower left quadrant; LRN: negative likelihood ratio; LRP: positive likelihood ratio; LUQ: upper left quadrant; RLQ: lower right quadrant; RUQ: upper right quadrant.



Publication Bias

The funnel plots and symmetry test were used to evaluate publication bias, indicating whether the results might be skewed due to selective reporting of positive or significant findings (Figure S4 in Multimedia Appendix 7). The funnel plots indicated no publication bias, and the asymmetry test was not significant ($P=.87$).

Subgroup Analysis

Subgroup analyses were conducted to explore whether specific study characteristics influenced the pooled diagnostic performance and to identify sources of heterogeneity. The

included covariates were validation strategy (internal test vs external test), backbone network architecture (ResNet vs DenseNet), scan plane (axial vs sagittal or coronal), image window setting (bone window vs soft tissue window), input data format (single-slice vs multislice), reference standard (CT vs DXA), and model dimensionality (2D vs 3D). The results are provided in Table 2. Models based on DenseNet variants, sagittal and coronal scans, multislice input, CT as the reference standard, and 3D architecture demonstrated superior diagnostic performance. Table S3 in Multimedia Appendix 8 provides detailed sensitivities and specificities for the heterogeneous subgroups.

Table 2. Results of subgroup analyses.

Subgroup	Joint model (<i>P</i> value)	I ^{2a} (%)	LRT ^b (χ ²) ^c
Validation strategy	.08	61	5.15
Backbone network architecture	<.001 ^d	96	46.90
Scan plane	<.001 ^d	95	41.19
Image window setting	.63	0	0.93
Input data format	<.001 ^d	89	18.65
Reference standard	<.001 ^d	94	35.71
Model dimensionality	<.001 ^d	87	15.19

^aI²: I-squared.

^bLRT: likelihood ratio test.

^cAll covariates in the subgroup analyses were binary. Therefore, the degrees of freedom for the LRT test were 2.

^dStatistically significant.

Discussion

Principal Findings

The rapid development of artificial intelligence, particularly DL, holds the potential to bring significant improvement to health care. DL facilitates opportunistic screening from routine CT scans in osteoporosis diagnosis, enabling early detection and intervention for asymptomatic patients while providing a cost-effective alternative for bone health assessment. However, translating these advancements into clinical practice remains challenging, particularly in terms of generalizability and clinical validation. Therefore, this study aimed to evaluate the diagnostic performance of DL models in osteoporosis detection, emphasizing both progress and limitations. The pooled sensitivity for diagnosing osteoporosis, osteopenia, and normal cases was 0.88, 0.81, and 0.95, respectively, whereas the pooled specificity was 0.94, 0.92, and 0.93, respectively. To our knowledge, this is the first meta-analysis focusing on the use of DL for osteoporosis detection. However, considerable heterogeneity across studies remains, and individual covariates evaluated in the subgroup analysis could not fully account for the observed heterogeneity. In addition, the meta-analysis excluded 4 studies due to insufficient data. Therefore, the results should be interpreted with caution. The pooled estimates should not be interpreted as a single, generalizable performance metric for all DL models, but rather as an overall summary of performance across diverse models, datasets, and clinical settings.

Among studies indexed with “Deep Learning” as a keyword, this systematic review and meta-analysis only included research using DL technology as a classifier. Beyond classification, DL has been used in other components of the automated diagnostic pipeline. A common approach involves using a DL-based localization module to determine ROI, followed by HU- or radiomics-based diagnosing methods [41-45]. HU-based diagnosis estimates BMD using CT density values [46,47]. However, its reliance on a single parameter limits adaptability. Radiomics-based approaches that extract predefined imaging features for prediction face challenges in both feature selection and generalizability. Recent studies indicate that DL-based models may outperform HU- and radiomics-based methods, which may be associated with their capability of automatically learning complex imaging features and leveraging larger datasets [14, 18,30]. In addition, DL enables direct feature extraction from raw imaging data, thereby eliminating the reliance on manual feature selection. Compared to previous meta-analyses on HU-based (pooled sensitivity: 63% and specificity: 91%) and radiomics-based (pooled sensitivity: 87% and specificity: 87%) diagnostic methods [48,49], our results indicate a potentially superior performance of DL-based models. However, further direct comparative studies are required to validate this advantage.

Subgroup analysis in terms of model construction revealed that models based on DenseNet demonstrated slightly better performance in osteoporosis diagnosis than those using ResNet. This may be due to DenseNet’s dense connectivity,

which improves feature reuse and facilitates stable gradient propagation, enhancing fine-grained image analysis. Studies on viral pneumonia and Alzheimer disease diagnosis reported similar results [50,51], indicating that DenseNet may be better suited for tasks requiring detailed texture and structural assessment. In addition, the use of deep architectures such as ResNet-50 carries an inherent risk of overfitting and suboptimal training performance when applied to small datasets [24]. However, although the 2 architectures showed statistically distinguishable performance, the absolute differences in pooled sensitivity and specificity were minimal, suggesting that the observed significance may not reflect a clinically meaningful superiority. This discrepancy may be attributed to the use of various optimization strategies in most included studies. Future research should conduct more rigorous comparative analyses to identify the optimal model architecture and training strategies for large-scale clinical deployment. Furthermore, it is not surprising that models using multislice input and 3D architecture demonstrated improved diagnostic performance, as these approaches can leverage more comprehensive CT imaging information. Models using sagittal or coronal scans showed superior performance compared with those based on axial scans, possibly because the former provide bone quality information from multiple vertebrae simultaneously. However, the reliability of this finding may be constrained, as studies in the sagittal or coronal subgroup were very limited (n=4).

The results of this study showed no significant difference between the internal and external test performance of DL models. However, this may be attributed to the limited number of eligible external test datasets (n=10), potentially affecting the stability of the pooled estimates. Considering that osteoporosis diagnosis and BMD estimation primarily depend on trabecular bone texture, it is reasonable to assume that factors, such as CT image contrast and clarity, could affect the diagnostic results. Several studies [24,27,39] have demonstrated variations in the performance of DL models when tested on CT images acquired from different vendors and tube voltage settings, which is consistent with our hypothesis.

Models using CT as the reference standard exhibited superior diagnostic performance compared with those using DXA. This is not unexpected, as the outputs tend to be more consistent when the index test and reference standard were derived from the same imaging modality. Although DXA remains the globally recognized gold standard for osteoporosis diagnosis [1], it has known limitations in clinical practice, including the risk of both false negatives and false positives. Lin et al [52] reported that the detection rate of osteoporosis was lower with DXA (73.2%) than with QCT (84.4%) among patients with vertebral fractures—clinically confirmed osteoporosis cases. This discrepancy is caused by DXA being susceptible to osteophytes, degenerative changes, and vascular calcifications [53]. When the reference standard itself is prone to measurement errors, the performance of DL models trained on such data is inherently constrained. Therefore, establishing a more accurate reference standard for

osteoporosis diagnosis remains a crucial challenge for future research.

Regarding the diagnostic workflow of DL models, existing studies generally adopt 2 approaches: prediction of BMD values followed by diagnosis, or direct classification. Although these 2 approaches theoretically should not cause differences in diagnostic accuracy, the former may provide practical advantages in clinical applications. First, BMD predictions provide clinicians and patients with a more precise assessment of osteoporosis severity. Second, Peng et al [25] indicated that a primary source of misclassification in DL-based diagnosis involved cases where BMD values are close to classification boundaries. By outputting continuous BMD values instead of discrete diagnostic categories, this approach helps mitigate the effect of misclassification on clinical decision-making.

Future Directions

Based on the above results and discussion, DL models face practical and ethical challenges in real-world applications beyond diagnostic performance, which should be further addressed in future research and clinical implementation.

First, the considerable heterogeneity observed across the included studies suggests that the performance of current DL models may be influenced by multiple factors, including model architecture and CT acquisition parameters. Therefore, particular caution is warranted when considering their deployment across different scanners and institutions. An alternative strategy for clinical implementation may involve tailoring automated diagnostic models to specific CT vendors and scanner types to ensure diagnostic performance. Future studies on model development should also comprehensively and transparently report details of model design and study data, and conduct more rigorous and extensive cross-institutional and prospective validation to enhance model robustness and generalizability.

DL models encounter practical and ethical challenges in real-world applications beyond diagnostic performance. Only a small proportion of the DL models included in this review were capable of end-to-end processing, as mentioned in the “Results” section. Moreover, Fang et al [29] acknowledged that 14%-34.5% of cases had invalid CT segmentation results. Similar problems in localization and segmentation further complicate the feasibility of end-to-end automation. These limitations hinder the feasibility of complete automation, requiring human oversight for reviewing and adjusting CT images as warranted. Balancing sensitivity and specificity remains another critical consideration for DL model application. The LR scatter plot indicates that the current DL models predominantly exhibit strong confirmatory capability, posing concerns about their suitability as screening tools. Future research should focus on optimizing the sensitivity of DL models to minimize missed diagnoses while maintaining diagnostic specificity as much as possible.

From an ethical perspective, current DL models primarily serve as diagnostic aids, with physicians integrating DL-generated outputs into their clinical judgment and assuming

responsibility for the final diagnosis. However, there is currently no radiological diagnostic standard for osteoporosis that human physicians can directly reference. This indicates that the outputs of DL models may directly constitute the final radiological interpretation, posing ethical concerns regarding accountability and trustworthiness in clinical applications. Therefore, replacing existing osteoporosis diagnostic methods with DL models is not a rational approach in the short term. When DL indicates a risk of osteoporosis, it is necessary to recommend established diagnostic examinations such as DXA and arrange appropriate longitudinal follow-up and treatment. Leveraging the large volume of routine CT examinations for preliminary osteoporosis screening, followed by standardized diagnostic confirmation and appropriate treatment, may improve the overall cost-effectiveness of osteoporosis management.

Limitations

This study has several limitations. First, although this study used a comprehensive search strategy as thoroughly as possible, the exclusion of non-English and non-peer-reviewed publications may still have resulted in the omission of certain valuable studies. Second, the included studies involved relatively small sample sizes, which may have limited the training effectiveness of DL models. Third, several included studies lacked sufficient reporting of key metrics, making the confusion matrix reconstruction impossible, which could introduce discrepancies between the pooled results and the actual performance. Fourth, the published papers did not always provide complete details of the proposed models, and some of the studies were assessed to have a risk of bias, which limited the interpretability and reproducibility of the results. The complexity of model design and CT parameters also limited our ability to perform further meta-regression analyses. Fifth, multiple test results from individual studies were treated as independent observations in this meta-analysis. Although this approach enabled a more comprehensive inclusion of model performances, it may have introduced potential bias by disproportionately weighting certain study designs. Focusing primarily on the diagnostic performance of ResNet and DenseNet may also represent a source of bias in this study. Finally, no independent validation studies of these models from other institutions were identified in the literature, emphasizing the need for further verification of their generalizability.

Conclusion

This systematic review and meta-analysis revealed that DL models exhibit promising sensitivity and specificity for osteoporosis diagnosis based on CT images. However, this study also highlights several limitations of existing DL models. First, the included studies showed substantial heterogeneity and lacked robust external validation, which restricts model generalizability; therefore, the pooled results should not be interpreted as a single and universal estimate. Second, some DL models have not achieved a fully end-to-end diagnostic pipeline. Third, the balance between sensitivity and specificity requires further optimization to better align with clinical screening requirements. Despite these

limitations, DL techniques hold considerable potential for integration into clinical practice, enabling broader osteoporosis screening and improving cost-effectiveness.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AW and ZM contributed equally to this work and should be regarded as co-first authors. LZ contributed to the conception and design of the study. ZM and AW contributed to the acquisition of data. TW, YX, and AW contributed to the analysis and interpretation of data. TW and NF contributed to drafting the article. RC, NF, SY, PD, and QW contributed to critically revising the article. LZ reviewed the submitted version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Full search strings.

[\[DOCX File \(Microsoft Word File\), 14 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Characteristics of computed tomography (CT) scans in the included studies.

[\[DOCX File \(Microsoft Word File\), 27 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Characteristics of computed tomography (CT) image input and deep learning (DL) algorithms in the included studies.

[\[DOCX File \(Microsoft Word File\), 24 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Influence of individual studies on the diagnosis of osteoporosis (leave-one-out analysis).

[\[DOCX File \(Microsoft Word File\), 93 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Influence of individual studies on the diagnosis of osteopenia (leave-one-out analysis).

[\[DOCX File \(Microsoft Word File\), 77 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Influence of individual studies on the diagnosis of normal cases (leave-one-out analysis).

[\[DOCX File \(Microsoft Word File\), 89 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Funnel plot of included studies.

[\[DOCX File \(Microsoft Word File\), 87 KB-Multimedia Appendix 7\]](#)

Multimedia Appendix 8

Sensitivities and specificities of the heterogeneous subgroups.

[\[DOCX File \(Microsoft Word File\), 17 KB-Multimedia Appendix 8\]](#)

Checklist 1

PRISMA-DTA (Preferred Reporting Items for Systematic Review and Meta-Analyses of Diagnostic Test Accuracy) checklist.

[\[DOCX File \(Microsoft Word File\), 31 KB-Checklist 1\]](#)

References

1. Ensrud KE, Crandall CJ. Osteoporosis. *Ann Intern Med.* Aug 1, 2017;167(3):ITC17-ITC32. [doi: [10.7326/AITC201708010](#)] [Medline: [28761958](#)]
2. Xiao PL, Cui AY, Hsu CJ, et al. Global, regional prevalence, and risk factors of osteoporosis according to the World Health Organization diagnostic criteria: a systematic review and meta-analysis. *Osteoporos Int.* Oct 2022;33(10):2137-2153. [doi: [10.1007/s00198-022-06454-3](#)] [Medline: [35687123](#)]
3. Karaguzel G, Holick MF. Diagnosis and treatment of osteopenia. *Rev Endocr Metab Disord.* Dec 2010;11(4):237-251. [doi: [10.1007/s11154-010-9154-0](#)] [Medline: [21234807](#)]

4. Si L, Winzenberg TM, Jiang Q, Chen M, Palmer AJ. Projection of osteoporosis-related fractures and costs in China: 2010-2050. *Osteoporos Int*. Jul 2015;26(7):1929-1937. [doi: [10.1007/s00198-015-3093-2](https://doi.org/10.1007/s00198-015-3093-2)] [Medline: [25761729](https://pubmed.ncbi.nlm.nih.gov/25761729/)]
5. Srivastava M, Deal C. Osteoporosis in elderly: prevention and treatment. *Clin Geriatr Med*. Aug 2002;18(3):529-555. [doi: [10.1016/s0749-0690\(02\)00022-8](https://doi.org/10.1016/s0749-0690(02)00022-8)] [Medline: [12424871](https://pubmed.ncbi.nlm.nih.gov/12424871/)]
6. Cheng X, Zhao K, Zha X, et al. Opportunistic screening using low-dose CT and the prevalence of osteoporosis in China: a nationwide, multicenter study. *J Bone Miner Res*. Mar 2021;36(3):427-435. [doi: [10.1002/jbmr.4187](https://doi.org/10.1002/jbmr.4187)] [Medline: [33145809](https://pubmed.ncbi.nlm.nih.gov/33145809/)]
7. Gates M, Pillay J, Nuspl M, Wingert A, Vandermeer B, Hartling L. Screening for the primary prevention of fragility fractures among adults aged 40 years and older in primary care: systematic reviews of the effects and acceptability of screening and treatment, and the accuracy of risk prediction tools. *Syst Rev*. Mar 21, 2023;12(1):51. [doi: [10.1186/s13643-023-02181-w](https://doi.org/10.1186/s13643-023-02181-w)] [Medline: [36945065](https://pubmed.ncbi.nlm.nih.gov/36945065/)]
8. Paderno A, Ataíde Gomes EJ, Gilberg L, et al. Artificial intelligence-enhanced opportunistic screening of osteoporosis in CT scan: a scoping review. *Osteoporos Int*. Oct 2024;35(10):1681-1692. [doi: [10.1007/s00198-024-07179-1](https://doi.org/10.1007/s00198-024-07179-1)] [Medline: [38985200](https://pubmed.ncbi.nlm.nih.gov/38985200/)]
9. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. Mar 29, 2021;372:n160. [doi: [10.1136/bmj.n160](https://doi.org/10.1136/bmj.n160)] [Medline: [33781993](https://pubmed.ncbi.nlm.nih.gov/33781993/)]
10. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst Rev*. Jun 2022;18(2):e1230. [doi: [10.1002/cl2.1230](https://doi.org/10.1002/cl2.1230)] [Medline: [36911350](https://pubmed.ncbi.nlm.nih.gov/36911350/)]
11. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-Analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA*. Jan 23, 2018;319(4):388-396. [doi: [10.1001/jama.2017.19163](https://doi.org/10.1001/jama.2017.19163)] [Medline: [29362800](https://pubmed.ncbi.nlm.nih.gov/29362800/)]
12. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res Synth Methods*. Mar 2020;11(2):181-217. [doi: [10.1002/jrsm.1378](https://doi.org/10.1002/jrsm.1378)] [Medline: [31614060](https://pubmed.ncbi.nlm.nih.gov/31614060/)]
13. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. Oct 18, 2011;155(8):529-536. [doi: [10.7326/0003-4819-155-8-201110180-00009](https://doi.org/10.7326/0003-4819-155-8-201110180-00009)] [Medline: [22007046](https://pubmed.ncbi.nlm.nih.gov/22007046/)]
14. Tong X, Wang S, Zhang J, Fan Y, Liu Y, Wei W. Automatic osteoporosis screening system using radiomics and deep learning from low-dose chest CT images. *Bioengineering (Basel)*. Jan 2, 2024;11(1):50. [doi: [10.3390/bioengineering11010050](https://doi.org/10.3390/bioengineering11010050)] [Medline: [38247927](https://pubmed.ncbi.nlm.nih.gov/38247927/)]
15. Dzierżak R, Omiotek Z. Application of deep convolutional neural networks in the diagnosis of osteoporosis. *Sensors (Basel)*. Oct 26, 2022;22(21):8189. [doi: [10.3390/s22218189](https://doi.org/10.3390/s22218189)] [Medline: [36365886](https://pubmed.ncbi.nlm.nih.gov/36365886/)]
16. 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(1):3018320. [doi: [10.1155/2023/3018320](https://doi.org/10.1155/2023/3018320)] [Medline: [36970245](https://pubmed.ncbi.nlm.nih.gov/36970245/)]
17. Huang C, Wu D, Wang B, et al. Application of deep learning model based on unenhanced chest CT for opportunistic screening of osteoporosis: a multicenter retrospective cohort study. *Insights Imaging*. Jan 10, 2025;16(1). [doi: [10.1186/s13244-024-01817-2](https://doi.org/10.1186/s13244-024-01817-2)]
18. Fang K, Zheng X, Lin X, Dai Z. A comprehensive approach for osteoporosis detection through chest CT analysis and bone turnover markers: harnessing radiomics and deep learning techniques. *Front Endocrinol*. 2024;15(1296047):38894742. [doi: [10.3389/fendo.2024.1296047](https://doi.org/10.3389/fendo.2024.1296047)]
19. Yousef R, Gupta G, Yousef N, Khari M. A holistic overview of deep learning approach in medical imaging. *Multimed Syst*. 2022;28(3):881-914. [doi: [10.1007/s00530-021-00884-5](https://doi.org/10.1007/s00530-021-00884-5)] [Medline: [35079207](https://pubmed.ncbi.nlm.nih.gov/35079207/)]
20. Li Y, Ye K, Liu S, et al. Deep learning-enhanced opportunistic osteoporosis screening in 100 kV low-voltage chest CT: a novel way toward bone mineral density measurement and radiation dose reduction. *Acad Radiol*. Nov 2025;32(11):6812-6822. [doi: [10.1016/j.acra.2025.07.060](https://doi.org/10.1016/j.acra.2025.07.060)] [Medline: [40835570](https://pubmed.ncbi.nlm.nih.gov/40835570/)]
21. Li Y, Liu S, Zhang Y, et al. Deep learning-enhanced opportunistic osteoporosis screening in ultralow-voltage (80 kV) chest CT: a preliminary study. *Acad Radiol*. Jul 2025;32(7):4254-4265. [doi: [10.1016/j.acra.2024.11.062](https://doi.org/10.1016/j.acra.2024.11.062)]
22. Nyaga VN, Arbyn M, Aerts M. Metaprop: a Stata command to perform meta-analysis of binomial data. *Arch Public Health*. 2014;72(1):39. [doi: [10.1186/2049-3258-72-39](https://doi.org/10.1186/2049-3258-72-39)] [Medline: [25810908](https://pubmed.ncbi.nlm.nih.gov/25810908/)]
23. Wu Y, Yang X, Wang M, et al. Artificial intelligence assisted automatic screening of opportunistic osteoporosis in computed tomography images from different scanners. *Eur Radiol*. Apr 2025;35(4):2287-2295. [doi: [10.1007/s00330-024-11046-2](https://doi.org/10.1007/s00330-024-11046-2)] [Medline: [39231830](https://pubmed.ncbi.nlm.nih.gov/39231830/)]
24. Wang S, Tong X, Cheng Q, et al. Fully automated deep learning system for osteoporosis screening using chest computed tomography images. *Quant Imaging Med Surg*. Apr 2024;14(4):2816-2827. [doi: [10.21037/qims-23-1617](https://doi.org/10.21037/qims-23-1617)]

25. Peng T, Zeng X, Li Y, et al. A study on whether deep learning models based on CT images for bone density classification and prediction can be used for opportunistic osteoporosis screening. *Osteoporos Int. Jan 2024*;35(1):117-128. [doi: [10.1007/s00198-023-06900-w](https://doi.org/10.1007/s00198-023-06900-w)] [Medline: [37670164](#)]
26. Pan J, Lin PC, Gong SC, et al. Effectiveness of opportunistic osteoporosis screening on chest CT using the DCNN model. *BMC Musculoskelet Disord. Feb 27, 2024*;25(1):176. [doi: [10.1186/s12891-024-07297-1](https://doi.org/10.1186/s12891-024-07297-1)] [Medline: [38413868](#)]
27. Yoshida K, Tanabe Y, Nishiyama H, et al. Feasibility of bone mineral density and bone microarchitecture assessment using deep learning with a convolutional neural network. *J Comput Assist Tomogr. 2023*;47(3):467-474. [doi: [10.1097/RCT.0000000000001437](https://doi.org/10.1097/RCT.0000000000001437)] [Medline: [37185012](#)]
28. Niu X, Huang Y, Li X, et al. Development and validation of a fully automated system using deep learning for opportunistic osteoporosis screening using low-dose computed tomography scans. *Quant Imaging Med Surg. Aug 1, 2023*;13(8):5294-5305. [doi: [10.21037/qims-22-1438](https://doi.org/10.21037/qims-22-1438)] [Medline: [37581046](#)]
29. Fang Y, Li W, Chen X, et al. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. *Eur Radiol. Apr 2021*;31(4):1831-1842. [doi: [10.1007/s00330-020-07312-8](https://doi.org/10.1007/s00330-020-07312-8)] [Medline: [33001308](#)]
30. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. *Eur Radiol. Jun 2020*;30(6):3549-3557. [doi: [10.1007/s00330-020-06677-0](https://doi.org/10.1007/s00330-020-06677-0)] [Medline: [32060712](#)]
31. Kang JW, Park C, Lee DE, Yoo JH, Kim M. Prediction of bone mineral density in CT using deep learning with explainability. *Front Physiol. 2022*;13(1061911):1061911. [doi: [10.3389/fphys.2022.1061911](https://doi.org/10.3389/fphys.2022.1061911)] [Medline: [36703938](#)]
32. Li J, Zhang P, Xu J, et al. Prediction of bone mineral density based on computer tomography images using deep learning model. *Gerontology. 2025*;71(1):71-80. [doi: [10.1159/000542396](https://doi.org/10.1159/000542396)] [Medline: [39527924](#)]
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. May 2021*;32(5):971-979. [doi: [10.1007/s00198-020-05673-w](https://doi.org/10.1007/s00198-020-05673-w)] [Medline: [33165630](#)]
34. Oh J, Kim B, Oh G, Hwangbo Y, Ye JC. End-to-end semi-supervised opportunistic osteoporosis screening using computed tomography. *Endocrinol Metab (Seoul). Jun 2024*;39(3):500-510. [doi: [10.3803/EnM.2023.1860](https://doi.org/10.3803/EnM.2023.1860)] [Medline: [38721637](#)]
35. Tariq A, Patel BN, Sensakovic WF, Fahrenholtz SJ, Banerjee I. Opportunistic screening for low bone density using abdominopelvic computed tomography scans. *Med Phys. Jul 2023*;50(7):4296-4307. [doi: [10.1002/mp.16230](https://doi.org/10.1002/mp.16230)] [Medline: [36748265](#)]
36. Küçükçiloğlu Y, Şekeroğlu B, Adalı T, Şentürk N. Prediction of osteoporosis using MRI and CT scans with unimodal and multimodal deep-learning models. *Diagn Interv Radiol. Jan 8, 2024*;30(1):9-20. [doi: [10.4274/dir.2023.232116](https://doi.org/10.4274/dir.2023.232116)] [Medline: [37309886](#)]
37. Zhou K, Xin E, Yang S, et al. Automated fast prediction of bone mineral density from low-dose computed tomography. *Acad Radiol. Jul 2025*;32(7):4111-4120. [doi: [10.1016/j.acra.2025.02.041](https://doi.org/10.1016/j.acra.2025.02.041)] [Medline: [40082126](#)]
38. Kuo DP, Chen YC, Cheng SJ, et al. A vision transformer-convolutional neural network framework for decision-transparent dual-energy X-ray absorptiometry recommendations using chest low-dose CT. *Int J Med Inform. Jul 2025*;199:105901. [doi: [10.1016/j.ijmedinf.2025.105901](https://doi.org/10.1016/j.ijmedinf.2025.105901)] [Medline: [40187299](#)]
39. Zhang K, Lin PC, Pan J, et al. DeepmdQCT: a multitask network with domain invariant features and comprehensive attention mechanism for quantitative computer tomography diagnosis of osteoporosis. *Comput Biol Med. Mar 2024*;170(107916):107916. [doi: [10.1016/j.combiomed.2023.107916](https://doi.org/10.1016/j.combiomed.2023.107916)] [Medline: [38237237](#)]
40. Expert Panel on Musculoskeletal Imaging, Ward RJ, Roberts CC, Bencardino JT, et al. ACR appropriateness criteria® osteoporosis and bone mineral density. *J Am Coll Radiol. May 2017*;14(5S):S189-S202. [doi: [10.1016/j.jacr.2017.02.018](https://doi.org/10.1016/j.jacr.2017.02.018)] [Medline: [28473075](#)]
41. Oh S, Kang WY, Park H, et al. Evaluation of deep learning-based quantitative computed tomography for opportunistic osteoporosis screening. *Sci Rep. Jan 5, 2024*;14(1):363. [doi: [10.1038/s41598-023-45824-7](https://doi.org/10.1038/s41598-023-45824-7)] [Medline: [38182616](#)]
42. 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. Jul 2020*;30(7):4107-4116. [doi: [10.1007/s00330-020-06679-y](https://doi.org/10.1007/s00330-020-06679-y)] [Medline: [32072260](#)]
43. Park H, Kang WY, Woo OH, Lee J, Yang Z, Oh S. Automated deep learning-based bone mineral density assessment for opportunistic osteoporosis screening using various CT protocols with multi-vendor scanners. *Sci Rep. Oct 23, 2024*;14(1). [doi: [10.1038/s41598-024-73709-w](https://doi.org/10.1038/s41598-024-73709-w)]
44. Pan Y, Zhao F, Cheng G, et al. Automated vertebral bone mineral density measurement with phantomless internal calibration in chest LDCT scans using deep learning. *Br J Radiol. Dec 2023*;96(1152):20230047. [doi: [10.1259/bjr.20230047](https://doi.org/10.1259/bjr.20230047)] [Medline: [37751163](#)]
45. Pickhardt PJ, Nguyen T, Perez AA, et al. Improved CT-based osteoporosis assessment with a fully automated deep learning tool. *Radiol Artif Intell. Sep 2022*;4(5):e220042. [doi: [10.1148/ryai.220042](https://doi.org/10.1148/ryai.220042)] [Medline: [36204542](#)]

46. Perez AA, Pickhardt PJ, Elton DC, Sandfort V, Summers RM. Fully automated CT imaging biomarkers of bone, muscle, and fat: correcting for the effect of intravenous contrast. *Abdom Radiol (NY)*. Mar 2021;46(3):1229-1235. [doi: [10.1007/s00261-020-02755-5](https://doi.org/10.1007/s00261-020-02755-5)] [Medline: [32948910](https://pubmed.ncbi.nlm.nih.gov/32948910/)]
47. Agaronnik ND, Giberson-Chen C, Bono CM. Using advanced imaging to measure bone density, compression fracture risk, and risk for construct failure after spine surgery. *Spine J*. Jul 2024;24(7):1135-1152. [doi: [10.1016/j.spinee.2024.02.018](https://doi.org/10.1016/j.spinee.2024.02.018)] [Medline: [38437918](https://pubmed.ncbi.nlm.nih.gov/38437918/)]
48. Deng L, Shuai P, Liu Y, et al. Diagnostic performance of radiomics for predicting osteoporosis in adults: a systematic review and meta-analysis. *Osteoporos Int*. Oct 2024;35(10):1693-1707. [doi: [10.1007/s00198-024-07136-y](https://doi.org/10.1007/s00198-024-07136-y)] [Medline: [38802557](https://pubmed.ncbi.nlm.nih.gov/38802557/)]
49. Ahern DP, McDonnell JM, Riffault M, et al. A meta-analysis of the diagnostic accuracy of Hounsfield units on computed topography relative to dual-energy X-ray absorptiometry for the diagnosis of osteoporosis in the spine surgery population. *Spine J*. Oct 2021;21(10):1738-1749. [doi: [10.1016/j.spinee.2021.03.008](https://doi.org/10.1016/j.spinee.2021.03.008)] [Medline: [33722727](https://pubmed.ncbi.nlm.nih.gov/33722727/)]
50. Xiao B, Yang Z, Qiu X, et al. PAM-DenseNet: a deep convolutional neural network for computer-aided COVID-19 diagnosis. *IEEE Trans Cybern*. Nov 2022;52(11):12163-12174. [doi: [10.1109/TCYB.2020.3042837](https://doi.org/10.1109/TCYB.2020.3042837)] [Medline: [34428169](https://pubmed.ncbi.nlm.nih.gov/34428169/)]
51. Al Shehri W. Alzheimer's disease diagnosis and classification using deep learning techniques. *PeerJ Comput Sci*. 2022;8:e1177. [doi: [10.7717/peerj-cs.1177](https://doi.org/10.7717/peerj-cs.1177)] [Medline: [37346304](https://pubmed.ncbi.nlm.nih.gov/37346304/)]
52. Lin W, He C, Xie F, et al. Quantitative CT screening improved lumbar BMD evaluation in older patients compared to dual-energy X-ray absorptiometry. *BMC Geriatr*. Apr 17, 2023;23(1):231. [doi: [10.1186/s12877-023-03963-6](https://doi.org/10.1186/s12877-023-03963-6)] [Medline: [37069511](https://pubmed.ncbi.nlm.nih.gov/37069511/)]
53. Xu XM, Li N, Li K, et al. Discordance in diagnosis of osteoporosis by quantitative computed tomography and dual-energy X-ray absorptiometry in Chinese elderly men. *J Orthop Translat*. Jul 2019;18:59-64. [doi: [10.1016/j.jot.2018.11.003](https://doi.org/10.1016/j.jot.2018.11.003)] [Medline: [31508308](https://pubmed.ncbi.nlm.nih.gov/31508308/)]

Abbreviations

BMD: bone mineral density

CT: computed tomography

DL: deep learning

DXA: dual-energy X-ray absorptiometry

LR: likelihood ratio

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PRISMA-DTA: Preferred Reporting Items for Systematic Review and Meta-Analyses of Diagnostic Test Accuracy

PROSPERO: Prospective Register of Systematic Reviews

QCT: quantitative CT

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2

ROI: region of interest

WHO: World Health Organization

Edited by Andrew Coristine; peer-reviewed by Bernardo Cedeno-Veloz, Xiaolong Liang; submitted 08.May.2025; final revised version received 14.Oct.2025; accepted 14.Oct.2025; published 24.Nov.2025

Please cite as:

Wang A, Ma Z, Wang T, Chen R, Xi Y, Wu Q, Yuan S, Fan N, Du P, Zang L

Deep Learning–Assisted Automated Diagnosis of Osteoporosis Based on Computed Tomography Scans: Systematic Review and Meta-Analysis

J Med Internet Res 2025;27:e77155

URL: <https://www.jmir.org/2025/1/e77155>

doi: [10.2196/77155](https://doi.org/10.2196/77155)

© Aobo Wang, Ziqian Ma, Tianyi Wang, Ruiyuan Chen, Yu Xi, Qichao Wu, Shuo Yuan, Ning Fan, Peng Du, Lei Zang. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 24.Nov.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.