PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis

Jongyun Jung[1], Jingyuan Dai[1], Bowen Liu[2], Qing Wu[1]*

1 Department of Biomedical Informatics (Dr. Qing Wu, Jongyun Jung, and Jingyuan Dai), College of Medicine, The Ohio State University, Columbus, Ohio, United States of America, **2** Department of Mathematics and Statistics, Division of Computing, Analytics, and Mathematics, School of Science and Engineering (Bowen Liu), University of Missouri-Kansas City, Kansas City, Missouri, United States of America

* Qing.Wu@osumc.edu

Abstract

Artificial Intelligence (AI), encompassing Machine Learning and Deep Learning, has increasingly been applied to fracture detection using diverse imaging modalities and data types. This systematic review and meta-analysis aimed to assess the efficacy of AI in detecting fractures through various imaging modalities and data types (image, tabular, or both) and to synthesize the existing evidence related to AI-based fracture detection. Peer-reviewed studies developing and validating AI for fracture detection were identified through searches in multiple electronic databases without time limitations. A hierarchical meta-analysis model was used to calculate pooled sensitivity and specificity. A diagnostic accuracy quality assessment was performed to evaluate bias and applicability. Of the 66 eligible studies, 54 identified fractures using imaging-related data, nine using tabular data, and three using both. Vertebral fractures were the most common outcome (n = 20), followed by hip fractures (n = 18). Hip fractures exhibited the highest pooled sensitivity (92%; 95% CI: 87–96, $p< 0.01$) and specificity (90%; 95% CI: 85–93, $p< 0.01$). Pooled sensitivity and specificity using image data (92%; 95% CI: 90–94, $p< 0.01$; and 91%; 95% CI: 88–93, $p < 0.01$) were higher than those using tabular data (81%; 95% CI: 77–85, $p< 0.01$; and 83%; 95% CI: 76–88, $p < 0.01$), respectively. Radiographs demonstrated the highest pooled sensitivity (94%; 95% CI: 90–96, $p < 0.01$) and specificity (92%; 95% CI: 89–94, $p< 0.01$). Patient selection and reference standards were major concerns in assessing diagnostic accuracy for bias and applicability. AI displays high diagnostic accuracy for various fracture outcomes, indicating potential utility in healthcare systems for fracture diagnosis. However, enhanced transparency in reporting and adherence to standardized guidelines are necessary to improve the clinical applicability of AI.

Review Registration: PROSPERO (CRD42021240359).

Author summary

Artificial Intelligence (AI) is increasingly employed to detect fractures by using various imaging modalities and data types. Our search of Medline (via PubMed), Web of Science,

OPEN ACCESS

Citation: Jung J, Dai J, Liu B, Wu Q (2024) Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. PLOS Digit Health 3(1): e0000438. https://doi.org/10.1371/journal.pdig.0000438

Editor: Martin G. Frasch, University of Washington, UNITED STATES

Received: May 3, 2023

Accepted: December 25, 2023

Published: January 30, 2024

Copyright: © 2024 Jung et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data generated or analyzed during the study are included in the published paper.

Funding: The research and analysis described in the current publication were supported by a grant (R21MD013681 to QW) from the National Institute on Minority Health and Health Disparities and a grant (R01AG080017 to QW) from the National Institute of Aging. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and IEEE revealed numerous primary studies demonstrating AI's superior performance in fracture detection. This systematic review and meta-analysis is the first to assess and compare the diagnostic accuracy of AI models across different imaging modalities and data types for various fracture outcomes. We found that AI models achieve high accuracy in fracture detection, particularly with radiograph images. However, we identified significant flaws in study design and reporting, limiting real-world applicability. Few studies provided patient characteristics, and only half reported the hyperparameter selection process. Our findings underscore the benefits of using AI models with radiographs for fracture detection, as they outperform other imaging modalities. Despite similar results across modalities, inadequate methodology and reporting in AI model evaluations call for improvement. Considering AI's high diagnostic performance, integrating it into existing fracture risk assessment tools could enhance patient identification and enable early intervention.

Introduction

Bone fractures represent a significant public health concern globally [1], particularly for individuals with osteoporosis [2]. Fractures contribute to work absences, disability, reduced quality of life, health complications, and increased healthcare costs, affecting individuals, families, and societies [3,4]. A meta-analysis of 113 studies reported the pooled cost of hospital treatment for a hip fracture after 12 months as $10,075, with total health and social care costs amounting to $43,669 per hip fracture [5].

Artificial Intelligence (AI), encompassing Machine Learning (ML) and Deep Learning (DL), has been extensively employed for fracture outcome prediction due to technological advancements and accessibility. Various imaging modalities, including X-rays [6,7], computed tomography (CT) [8,9], and magnetic resonance imaging (MRI) [10,11], have been used in fracture diagnosis and detection. AI can also predict fractures using tabular data, such as electronic medical records (structured patient-level data). However, few studies [12–14] have applied AI with tabular data in fracture prediction despite its growing importance over the past decade. Recent systematic reviews and meta-analyses have reported high accuracy for AI in fracture detection and classification. Kuo et al. [15] summarized 42 studies with 115 contingency tables, finding pooled sensitivity of 92% (95% CI: 88, 94) and specificity of 91% (95% CI: 88, 93). Yang et al. [16] reviewed 14 studies on orthopedic fractures, reporting pooled sensitivity and specificity of DL models as 87% (95% CI: 78, 93) and 91% (95% CI: 85, 95), respectively.

However, existing systematic review and meta-analysis studies focused solely on image-based analyses, neglecting comprehensive examination of various imaging modalities and data types (image, tabular, or both). Despite the superior performance of AI for medical image analysis and using tabular data, a critical gap exists in the current literature concerning the optimal choice of image modalities and the choice between image, tabular, or combined data types. There is a lack of comprehensive guidance on the most effective selection of image modalities and data types for fracture diagnosis. This gap in knowledge underscores the need for systematic investigation to determine which image modality, and by extension, which data type, yields the highest diagnostic accuracy and clinical relevance in AL algorithms. Addressing this gap will not only optimize the design of AI-based diagnostic tools but also enable healthcare practitioners to make informed decisions when selecting appropriate imaging modalities and data types for improved patient care.

Thus, this study primarily aims to evaluate the diagnostic accuracy of AI in fracture detection using diverse imaging modalities and data types, reflecting AI's growing role in healthcare. Additionally, we seek to synthesize current evidence on AI-based fracture detection, offering a concise overview and discerning the strengths and limitations of various data types, whether image, tabular, or combined.

Materials and methods

Identification and selection of studies

This systematic review, registered with PROSPERO (CRD42021240359), follows PRISMA guidelines (**S1 PRISMA Checklist**) [17]. We searched Medline (via PubMed), Web of Science, and IEEE. The last search was conducted on December 15, 2022, and we manually searched bibliographies, citations, and related articles of included studies. **S1 Text** lists each search term. Two independent reviewers (JJ and JD) assessed study eligibility, resolving disagreements through discussion or involving a third author (BL) if necessary.

Eligible studies predicted fracture outcomes using structured patient-level health data (electronic health records and cohort studies data) and image-related data (MRI, DXA, and X-ray). We excluded reviews, gray literature, non-human subject studies, studies without machine learning or deep learning models, fracture outcomes, AUC, accuracy, sensitivity, specificity, validation, and insufficient algorithm development details. We only considered studies published in English without time restrictions.

Data extraction

All three categories of data were considered: image-related, tabular, and both. Image-type studies used MRI, DXA, CT, or X-ray; tabular-type studies used structured electronic health records data; image and tabular studies used both data types. Two investigators (JJ and JD) independently evaluated study eligibility, extracting relevant data for articles meeting inclusion criteria. A structured data collection form was used to capture general study characteristics, population, data preprocessing, clinical outcomes, analytical methods, and results. A third author (BL) resolved discrepancies if necessary. We constructed the contingency table (true positive, true negative, false positive, and false negative) based on the provided information of sensitivity, specificity, positive predictive value, and negative predictive value for each study (**S4 Table**). If the study reported multiple sensitivity and specificity, we used the highest sensitivity and specificity.

Statistical analysis

Meta-analyses were performed using a random-effects model to calculate the pooled sensitivity and specificity based on logit transformation [18,19], using the Clopper-Pearson interval to calculate 95% confidence intervals for each study [20]. We used a unified hierarchical summary receiver operating characteristic curve (HSROC) to investigate the relationship between logit-transformed sensitivity and specificity. We calculated the diagnostic odds ratio and used inverse variance weighting for pooling with random effect models [21].

Sensitivity analysis

The logit transformation does not consider the correlation between sensitivity, specificity, and threshold effects; another model is desired to capture this missing part. Barendregt et al. [22] recommend using the Freeman-Tukey double arcsine transformation instead of the logit transformation. Hence, we used the Freeman-Tukey double arcsine transformation as a sensitivity analysis [22] for a random-effects model.

Subgroup analysis

Two subgroup analyses were conducted: 1) three data types (images, tabular, or images and tabular) and 2) different image modalities among image data used in AI. Statistical analysis was performed using R [23], with 'meta' [24] and 'mada' [25] packages. A p-value of < 0.05 was considered statistically significant.

Publication bias

We utilized the contour-enhanced funnel plot [26] to illustrate the assessment of publication bias for each fracture outcome and data type used. Each data point in the contour-enhanced funnel plot represents an individual study, and the plot incorporates contour lines that delineate expected areas of symmetry in the absence of bias. The plot provides insights into potential publication bias, with asymmetry suggesting a deviation from expected publication patterns. We employed the trim-and-fill method to address publication bias [22] further. This statistical approach helps adjust for the potential missing studies due to publication bias by imputing hypothetical "filled" studies and recalculating the effect size accordingly.

Risk of bias and applicability

Two reviewers (JJ and JD) independently evaluated the risk of bias in each study using Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) [27], assessing four domains: patient selection, index test, reference standard, and flow and timing. The risk of applicability was evaluated with the first three domains.

Results

Study selection and characteristics

Our search identified 1,128 studies, yielding 717 unique ones after removing duplicates (**Fig 1**). We screened titles and abstracts and selected 496 studies for full-text review based on our inclusion criteria. We then excluded 254 studies for lacking sensitivity and specificity information (149 studies), not having fracture-related outcomes (75 studies), not using ML models (28 studies), or being survey or review articles (2 studies). We further removed 176 studies because no contingency table could be calculated from the provided information. Ultimately, 66 studies were included in our systematic review and meta-analysis.

The selected studies were published between 2007 and 2022, with 73% (48 studies) published in the last three years (**Table 1**). The studies were conducted in various countries, including Asian countries (26 studies) [6,9,11,28–50], North American countries (19 studies) [14,34,36,51–66], European countries (14 studies) [13,59,67–78], Australia (1 study) [79] and Brazil (2 studies) [10,80] (**Table 1**). Four studies did not provide the country information [81–84].

Fracture identification was performed using imaging-related data in 54 studies, tabular data in nine studies, and imaging and tabular data in three. Of the 57 studies using imaging-related and combined data, 33 analyzed radiograph images [6,7,28–31,35–38,40–42,45,47–49,52–57,59,61,62,66–68,72–74,78], 12 analyzed computed tomography (CT) images [8,9,39,43,50,63,65,69,75,81–83], and the remaining studies analyzed other imaging modalities (**S1 Table, and S2 Table**). The most common fracture outcome was vertebral fracture (20

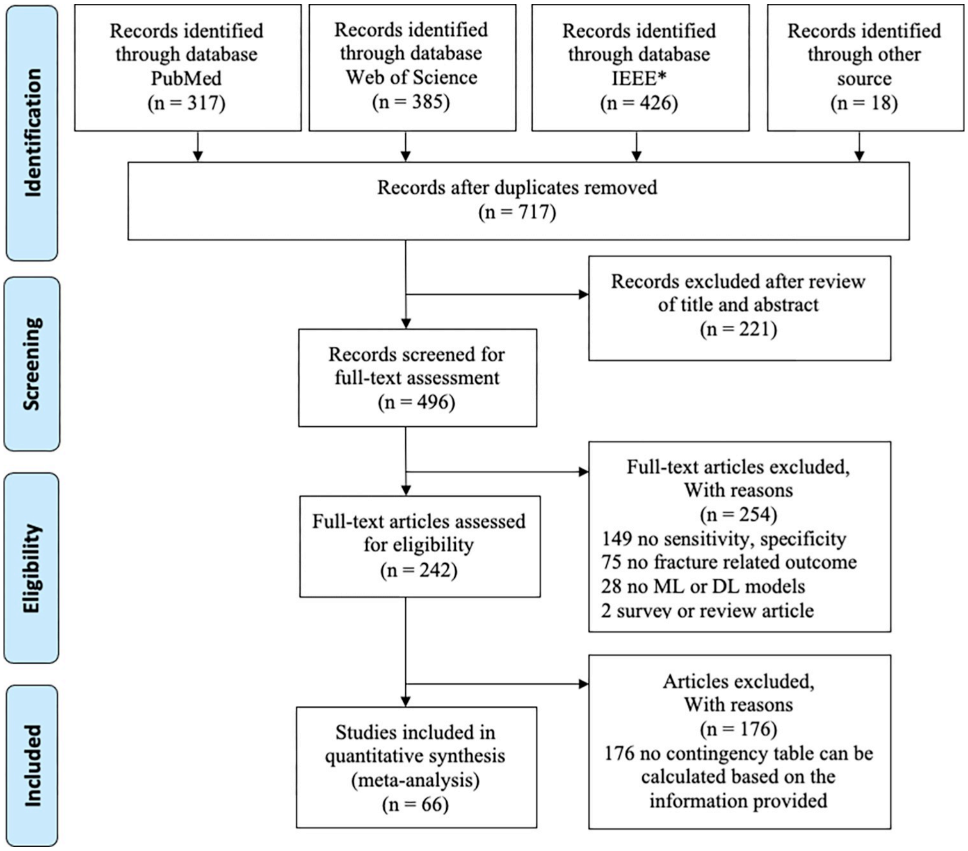
Identification
Screening
Eligibility
Included
Records identified through database PubMed (n = 317)
Records identified through database Web of Science (n = 385)
Records identified through database IEEE* (n = 426)
Records identified through other source (n = 18)
Records after duplicates removed (n = 717)
Records excluded after review of title and abstract (n = 221)
Records screened for full-text assessment (n = 496)
Full-text articles excluded, With reasons (n = 254)
149 no sensitivity, specificity
75 no fracture related outcome
28 no ML or DL models
2 survey or review article
Full-text articles assessed for eligibility (n = 242)
Articles excluded, With reasons (n = 176)
176 no contingency table can be calculated based on the information provided
Studies included in quantitative synthesis (meta-analysis) (n = 66)

Fig 1. Flow chart of the literature selection in PubMed, Web of Science, and Institute of Electrical and Electronics Engineers (search conducted on December 15, 2022). *IEEE: Institute of Electrical and Electronics Engineers.

https://doi.org/10.1371/journal.pdig.0000438.g001

studies) [8,10,11,28,31,34,35,38,44,46,50,51,58,59,65,72,77,80,83,84], followed by hip [6,13,29,32,33,37,39–43,48,53,62,64,66,68,79], and other fracture types (**Table 1**).

AI algorithms summary

Among the 54 studies that utilized imaging-related data, convolutional neural networks (CNN), a deep learning approach, emerged as the predominant choice, followed by instances where transfer learning was adopted. In some cases, the limited availability of labeled image data prompted the utilization of transfer learning [53,69], and certain studies incorporated pre-trained CNNs with non-fracture-related radiological images [6,28,85]. The prevailing preference was for fully connected artificial neural networks within the subset of nine studies involving tabular data. Logistic regression and ensemble learning models were commonly employed, including Random Forest, Gradient Boosting, and XGBoost. Among the three studies that harnessed both image and tabular data, a notable trend was the adoption of the support vector machine with various kernel models [57,68].

Handling imbalanced data and data augmentation

Imbalanced fracture outcomes were reported in 48 studies (**S3 Table**). Only 12 studies addressed the handling of imbalance outcomes during model development, using Synthetic Minority Over-sampling Technique (SMOTE) [86] or undersampling [35]. Data

Table 1. Fracture detection of 66 selected studies using machine learning and deep learning models and general characteristics of the study.

First author (Year published)	Country	Data type	Outcome	Model
Almog et al. (2020) [12]	USA	Tabular	Osteoporotic Fracture	XGBoost, Ensemble
Bae et al. (2021) [7]	Canada	Image	Femoral Neck	CNN, Four Different Convolutional Block Attention Modules
Beyaz et al. (2020) [67]	Turkey	Image	Femoral Neck	CNN
Burns et al. (2017) [8]	USA	Image	Vertebral	SVM
Chen et al. (2021) [28]	Taiwan	Image	Vertebral	CNN
Chen et al. (2022) [46]	China	Image	Vertebral	CNN, Other: Used ResNetSt-50 as the backbone network of the baseline model
Cheng et al. (2019) [6]	Taiwan	Image	Hip	CNN
Cheng et al. (2020) [29]	Taiwan	Image	Hip	CNN
Cheng et al. (2021) [30]	Taiwan	Image	Hip	CNN
Choi et al. (2020) [47]	South Korea	Image	Supracondylar	CNN
Chou et al. (2022) [31]	Taiwan	Image	Vertebral	CNN, Transfer Learning, Ensemble model (ResNet34, DenseNet121, DenseNet201)
Chung et al. (2018) [45]	Korea	Image	Proximal humerus	CNN
Derkatch et al. (2019) [51]	Canada	Image	Vertebral	CNN
Galassi et al. (2020) [68]	Spain	Tabular +Image	Hip	LR, SVM, Decision Trees, Random Forest
Guermazi et al. (2022) [52]	USA	Image	Hip, Wrist, Pelvic, Thoracolumbar, Foot, Ankle, Arm, Shoulder, Rib	Detectron2
Gupta et al. (2020) [53]	USA	Image	Hip	Transfer Learning: used VGG16 architecture with pre-trained weights using the ImageNet
Hayashi et al. (2022) [54]	USA	Image	Hand, Elbow, Shoulder, Foot, Leg	Detectron2
Ho-Le et al. (2017) [79]	Australia	Tabular	Hip	ANN, KNN, SVM
Inoue et al. (2022) [9]	Japan	Image	Pelvic, Spine, Rib	CNN
Kim et al. (2018) [69]	England	Image	Wrist	Inception v3 CNN model (transfer learning, trained in non-fracture images)
Kitamura et al. (2020) [55]	USA	Image	Hip	CNN
Korfiatis et al. (2018) [81]	NA	Image	Trabecular bone	Multilayer Perceptron SVM
Kruse et al. (2017) [13]	Denmark	Tabular	Hip	Twenty-four statistical models were built
Del Lama et al. (2022) [80]	Brazil	Tabular +Image	Vertebral	CNN, Multilayer Perceptron
Lemineur et al. (2007) [70]	France	Tabular	Osteoporotic Fracture	ANN
Lindsey et al. (2018) [56]	USA	Image	Wrist	CNN
Liu et al. (2015) [32]	Taiwan	Tabular	Hip	ANN
Liu et al. (2022) [48]	China	Image	Hip	CNN
Mawatari et al. (2020) [37]	Japan	Image	Hip	CNN
Mehta et al. (2020) [57]	USA	Tabular +Image	Lumbar Spine	SVM with a different kernel
Minonzio et al. (2020) [71]	France	Image	Hip	SVM, LR
Monchka et al. (2021) [58]	Canada	Image	Vertebral	CNN

(*Continued*)

Table 1. (Continued)

First author (Year published)	Country	Data type	Outcome	Model
Monchka et al. (2022) [59]	Switzerland, Canada	Image	Vertebral	CNN, Active Learning
Mu et al. (2021) [49]	China	Image	Femoral Neck	CNN
Murata et al. (2020) [38]	Japan	Image	Vertebral	CNN
Mutasa et al. (2020) [60]	USA	Image	Femoral Neck	CNN
Nguyen et al. (2022) [61]	USA	Image	Foot, Ankle, Knee, Leg, Hand, Wrist, Elbow, Arm, Shoulder, Clavicle	CNN
Nishiyama et al. (2014) [39]	Japan	Image	Hip	SVM
Nissinen et al. (2021) [72]	Finland	Image	Vertebral	CNN
Oakden-Rayner et al. (2022) [62]	USA, Australia	Image	Hip	CNN
Ozkaya et al. (2022) [73]	Turkey	Image	Scaphoid	CNN
Raghavendra et al. (2018) [82]	NA	Image	Thoracolumbar	CNN
Raisuddin et al. (2021) [74]	Finland	Image	Wrist	CNN
Ramos et al. (2022) [10]	Brazil	Image	Vertebral	CNN, SVM, KNN, ExtraTrees, QDA
Regnard et al. (2022) [75]	France	Image	Pelvic, Limbs	CNN
Rosenberg et al. (2022) [76]	Italy	Image	Thoracolumbar	CNN
Salehinejad et al. (2021) [83]	NA	Image	Vertebral	CNN with ResNet-50+BLSTM layer
Sato et al. (2021) [40]	Japan	Image	Hip	CNN with EfficientNet-B4 model (a pre-trained ImageNet model)
Small et al. (2021) [63]	USA	Image	Cervical Spine	CNN
Su et al. (2019) [64]	USA	Tabular	Hip	Classification and regression tree
Tomita et al. (2018) [65]	USA	Image	Vertebral	CNN
Tseng et al. (2013) [33]	Taiwan	Tabular	Hip	LR, Ensemble ANN
Ulivier et al. (2021) [77]	Italy	Tabular	Vertebral	ANN
Urakawa et al. (2019) [41]	Japan	Image	Hip	Transfer learning of CNN (VGG_16 network)
Ureten et al. (2022) [78]	Turkey	Image	Hand	Transfer Learning
Wang et al. (2022) [84]	NA	Image	Vertebral	CNN
Wu et al. (2020) [14]	USA	Tabular	Major Osteoporotic Fractures	LR, GB, RF, ANN
Yabu et al. (2021) [11]	Japan	Image	Vertebral	CNN
Yamada et al. (2020) [42]	Japan	Image	Hip	CNN
Yamamoto et al. (2020) [43]	Japan	Image	Hip	CNN
Yeh et al. (2022) [34]	Taiwan, USA	Image	Vertebral	Transfer Learning
Yi-Chu Li et al. (2021) [35]	Taiwan	Image	Vertebral	Transfer learning, Ensemble model
Yoda et al. (2022) [44]	Japan	Image	Vertebral	CNN, Transfer Learning
Yoon et al. (2021) [36]	Taiwan, USA	Image	Scaphoid	CNN
Yu et al. (2020) [66]	USA	Image	Hip	Transfer Learning

(*Continued*)

Table 1. (Continued)

First author (Year published)	Country	Data type	Outcome	Model
Yuan Li et al. (2021) [50]	China	Image	Vertebral	CNN (ResNet50)

CNN, Convolution Neural Network; SVM, Support Vector Machine; LR, Logistic Regression; RF, Random Forest; ANN, Artificial Neural Network; MLP, Multi Layers Perceptron; KNN, K-Nearest Neighbors; GB, Gradient Boosting; NLP, Natural Language Processing; QDA, Quadratic Discriminant Analysis

https://doi.org/10.1371/journal.pdig.0000438.t001

augmentation was frequently utilized in image studies, including horizontal and vertical rotation [45,50,58,67,69,72], adding Gaussian noise [67], random rescaling and flipping [30,53], mirroring, and lighting and contrast adjustments [56].

Hyperparameter optimization

Thirty-six studies reported the detailed process for optimizing hyperparameters in the final selected models (**S3 Table**). Beyaz et al. utilized genetic algorithms to identify the optimal hyperparameters for their CNN architecture [67]. Liu et al. explored the impact of varying the number of hidden neurons in the output layer [32]. Nissinen et al. [72] employed two approaches for hyperparameter searches: random search [87] and hyperband [88].

Data split and validation in an external data set

Fifty-one studies reported the split sample for model development (training) and validation (testing) (**S3 Table**). No universal rule of data separation was found. A different set of split samples was utilized, e.g., 80% training and 20% testing [10,28,47,57,71], 90% training and 10% testing [32,33,56,81], and 80% training, 10% validation, and 10% testing [40,41,65,69]. Twenty studies reported the cross-validation with 20-folds [66], 10-folds [8,14,33,34,39,45,50,53,57,64,72,76,80,81], 5-folds [13,28,32,38,44,46,48,67,74,78,79], and 7-folds [83]. Thirteen studies performed an out-of-sample external validation [6,7,29–31,35,47,49,56,59,62,72,74]. Choi et al. [47] performed external tests using two types of distinct datasets: temporal data, which was obtained at a different period from the model development, and other geographically separated data, which was collected from a different center. Li et al. [35] utilized a dataset from another medical center that used a different plain radiographic technique.

Meta-analysis

We extracted 66 contingency tables for each selected study (**S4 Table**). The overall pooled sensitivity and specificity, calculated using logit transformation, were 91% (95% CI: 88, 93) and 90% (95% CI: 88, 92), respectively (**Table 2**). The pooled sensitivities for hip and vertebral fractures were found to be 92% (95% CI: 87–96) and 86% (95% CI: 82–89), respectively, while the pooled specificities for these fractures were 90% (95% CI: 85–93) and 86% (95% CI: 81–90), respectively (**Table 2**). The unified hierarchical summary receiver operating characteristic curve for different fracture types is shown in **Fig 2**. The area under the curve (AUC) was highest for femoral neck fractures at 0.98, followed by other fractures (0.97), multiple fractures (0.93), hip fractures (0.91), wrist (0.86), and vertebral (0.84).

Sensitivity analysis

Arcsine transformation yielded similar results with the pooled sensitivity at 89% (95% CI: 87, 91) and specificity at 88% (95% CI: 86, 91). Among data types, studies using only image data

Table 2. Pooled Sensitivities, Specificities, and Diagnostic Odds Ratio for 60 studies in different fractures outcome. Studies with only one selected fracture outcome (cervical spine, hand, lumber spine, proximal humerus, supracondylar, and trabecular bone) were omitted.

Outcome	Sensitivity (%)$^{1)}$	Specificity (%)$^{1)}$	Sensitivity (%)$^{2)}$	Specificity (%)$^{2)}$	Diagnostic Odds Ratio	No. of Studies included
Overall	0.91 (0.88, 0.93)	0.90 (0.88, 0.92)	0.89 (0.87, 0.91)	0.88 (0.86, 0.91)	81.14 (53.69, 122.63)	66
Vertebral	0.86 (0.82, 0.89)	0.86 (0.81, 0.90)	0.86 (0.82, 0.89)	0.86 (0.81, 0.90)	38.26 (21.36, 68.51)	20
Hip	0.92 (0.87, 0.96)	0.90 (0.85, 0.93)	0.90 (0.85, 0.95)	0.89 (0.85, 0.93)	99.50 (39.37, 251.48)	18
Multiple*	0.90 (0.81, 0.96)	0.92 (0.87, 0.95)	0.88 (0.81, 0.94)	0.91 (0.85, 0.95)	88.71 (33.54, 234.64)	11
Femoral Neck	0.94 (0.87, 0.97)	0.90 (0.64, 0.98)	0.93 (0.86, 0.98)	0.85 (0.68, 0.97)	125.82 (10.96, 1444.74)	4
Wrist	0.90 (0.76, 0.96)	0.93 (0.85, 0.97)	0.89 (0.75, 0.97)	0.93 (0.85, 0.98)	105.68 (56.44, 197.89)	3
Scaphoid	0.92 (0.68, 0.98)	0.81 (0.54, 0.94)	0.89 (0.61, 1.00)	0.80 (0.49, 0.98)	65.27 (44.16, 96.46)	2
Thoracolumbar	0.97 (0.84, 0.99)	0.92 (0.90, 0.95)	0.95 (0.80, 1.00)	0.92 (0.90, 0.95)	278.30 (15.99, 4843.58)	2

Data in parentheses are 95% confidence intervals.

$^{1)}$: the logit transformation was used to calculate the pooled sensitivity and specificity.

$^{2)}$: the arcsine transformation was used to calculate the pooled sensitivity and specificity.

* Multiple fractures outcome studies include hip and pelvic (2), hip and spine (1), major osteoporotic fractures (1), multiple (3), osteoporotic fractures (2), pelvic and limbs (1), pelvic, spine, and rib (1).

https://doi.org/10.1371/journal.pdig.0000438.t002

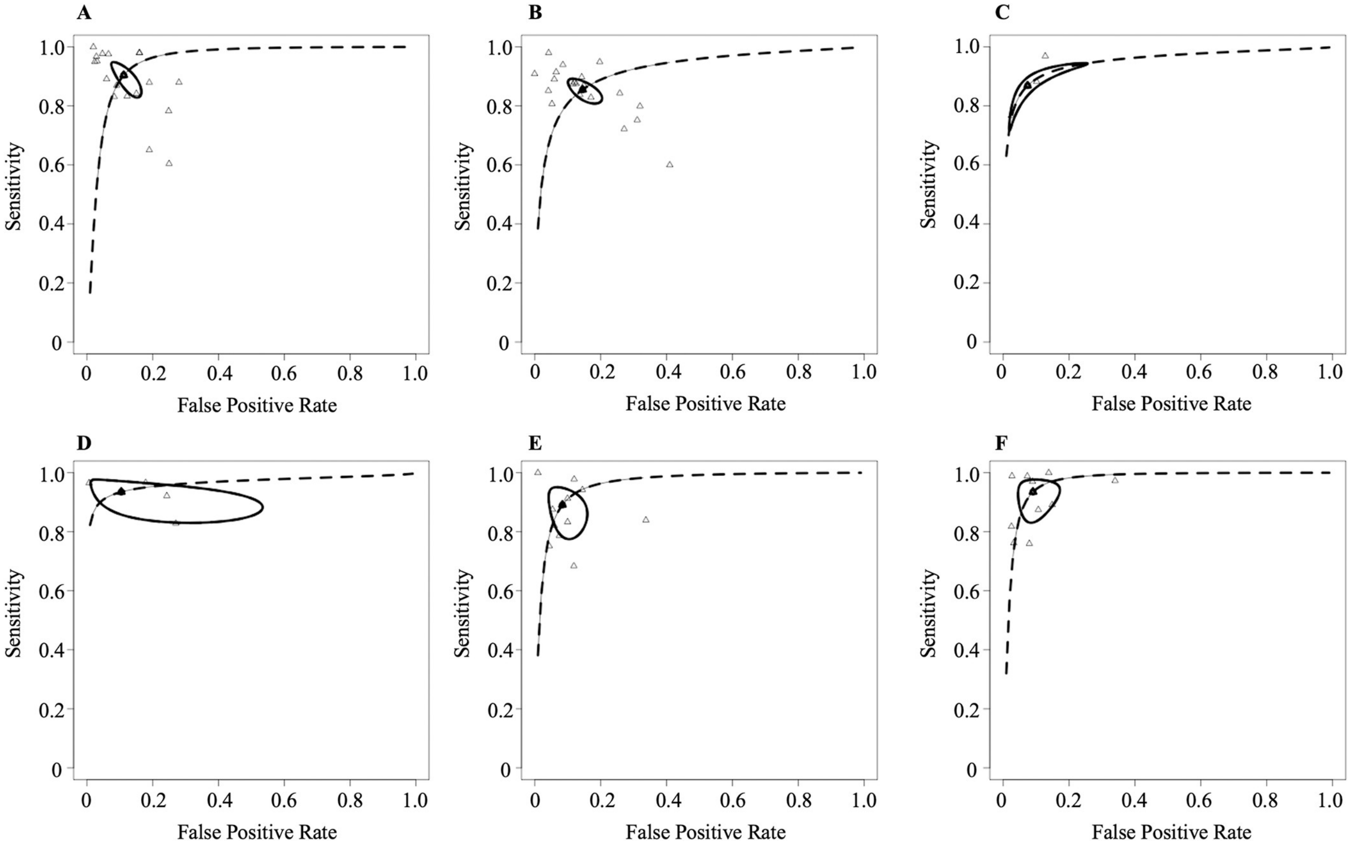

Fig 2. The hierarchical summary receiver operating characteristic curve for different fracture types in the meta-analysis. **A**: Hip (18 studies), **B**: Vertebral (20 studies), **C**: Wrist (3 studies), **D**: Femoral Neck (4 studies), **E**: Multiple (11 studies), and **F**: Others (10 studies).

https://doi.org/10.1371/journal.pdig.0000438.g002

Table 3. Pooled Sensitivities, Specificities, and Diagnostic Odds Ratio for 66 studies in different data type used.

Data Type	Sensitivity (%)[1)]	Specificity (%)[1)]	Sensitivity (%)[2)]	Specificity (%)[2)]	Diagnostic Odds Ratio	No. of Studies included
Tabular	0.81 (0.77, 0.85)	0.83 (0.76, 0.88)	0.81 (0.76, 0.85)	0.82 (0.76, 0.87)	20.06 (12.14, 33.16)	9
Image	0.92 (0.90, 0.94)	0.91 (0.88, 0.93)	0.91 (0.88, 0.93)	0.89 (0.87, 0.91)	104.20 (65.12, 166.72)	54
Tabular + Image	0.84 (0.76, 0.89)	0.95 (0.88, 0.98)	0.84 (0.77, 0.90)	0.96 (0.89, 1.00)	73.15 (27.23, 196.52)	3

Data in parentheses are 95% confidence intervals.

1): the logit transformation was used to calculate the pooled sensitivity and specificity.

2): the arcsine transformation was used to calculate the pooled sensitivity and specificity.

https://doi.org/10.1371/journal.pdig.0000438.t003

exhibited superior diagnostic performance with sensitivity and specificity at 91% (95% CI: 88, 93) and 89% (95% CI: 78, 91) using the arcsine transformation (Table 3). Studies employing radiographs displayed the highest sensitivity (92% [95% CI: 89, 95]) and specificity (90% [95% CI: 87, 93]) using the arcsine transformation (Table 4).

Subgroup analysis

Among data types, studies using only image data exhibited superior diagnostic performance with sensitivity and specificity at 92% (95% CI: 90, 94) and 91% (95% CI: 88, 93), respectively, when using logit transformation (Table 3). Studies employing radiographs displayed the highest sensitivity (94% [95% CI: 90, 96]) and specificity (92% [95% CI: 89, 94]) using logit transformation (Table 4). The AUC for radiograph studies (0.94) was higher than studies using radiograph and CT together (0.89) or MRI alone (0.88). The diagnostic odds ratio (DOR) was highest for hip fractures at 99.50 (95% CI: 39.37, 251.48) compared to vertebral fractures (38.26 [95% CI: 21.36, 68.51]) (Table 2). The AUC for image data studies (0.96) was higher than that for those using tabular and images together (0.83) or tabular data alone (0.81) (Fig 3).

Publication bias

The assessment of publication bias encompassed each fracture outcome and the utilization of distinct data types (**S5 and S6 Tables, S1–S3 Figs**). The Contour-Enhanced Funnel Plot illustrated the study distribution, and its enhanced contour facilitated the identification of potential bias (**S1—S3 Figs**). Notably, asymmetrical distribution was evident in the context of hip and vertebral fracture outcomes, and the studies used image data only (**S1 Fig and S3 Fig**).

Table 4. Pooled sensitivities, specifications, and diagnostic odds ratios for 54 studies (including three from the tabular and image data used) in different image modalities. Studies with only one selected image modality (Radiograph + CT + MRI, Radiograph + MRI, UGWSI) were omitted.

Image Modality	Sensitivity (%)[1)]	Specificity (%)[1)]	Sensitivity (%)[2)]	Specificity (%)[2)]	Diagnostic Odds Ratio	No. of Studies included
CT	0.89 (0.80, 0.94)	0.90 (0.85, 0.93)	0.86 (0.79, 0.92)	0.89 (0.84, 0.93)	67.16 (28.34, 159.18)	12
MRI	0.91 (0.83, 0.95)	0.89 (0.84, 0.93)	0.91 (0.84, 0.96)	0.91 (0.84, 0.95)	89.46 (26.41, 302.99)	5
Radiograph	0.94 (0.90, 0.96)	0.92 (0.89, 0.94)	0.92 (0.89, 0.95)	0.90 (0.87, 0.93)	150.92 (76.75, 296.78)	33
Radiograph + CT	0.93 (0.79, 0.98)	0.84 (0.81, 0.87)	0.92 (0.75, 1.00)	0.84 (0.80, 0.88)	66.11 (16.48, 265.26)	2
VFAI	0.87 (0.86, 0.89)	0.88 (0.87, 0.89)	0.87 (0.86, 0.89)	0.88 (0.87, 0.89)	50.64 (42.14, 60.86)	2

Data in parentheses are 95% confidence intervals.

1): the logit transformation was used to calculate the pooled sensitivity and specificity.

2): the arcsine transformation was used to calculate the pooled sensitivity and specificity.

UGWSI: Ultrasonic Guided Wave Spectrum Image, VFAI: Vertebral Fracture Assessment Image

https://doi.org/10.1371/journal.pdig.0000438.t004

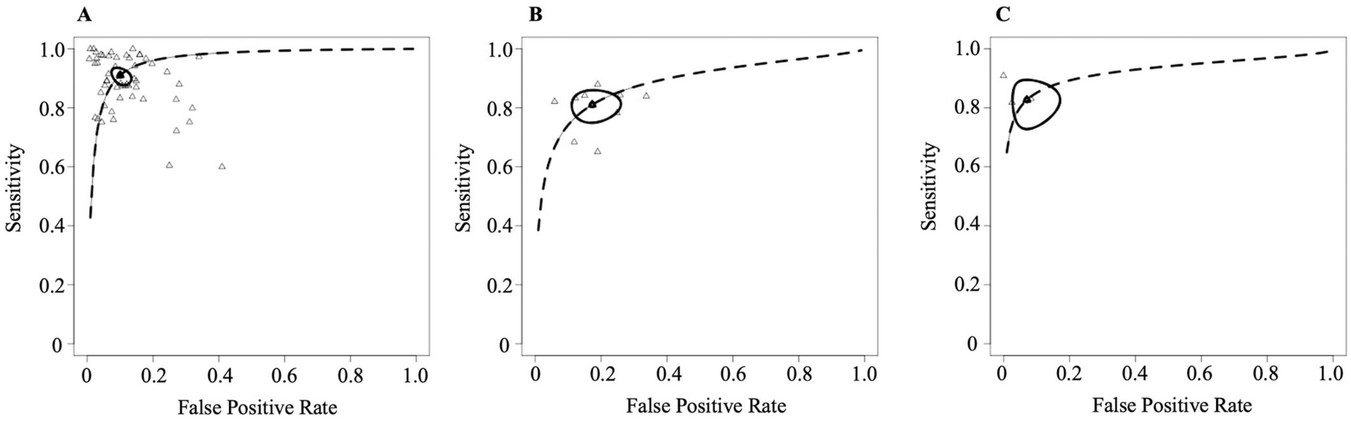

Fig 3. Unified hierarchical summary receiver operating characteristic curve for different data types in the meta-analysis. **A**: image (54 studies), **B**: tabular (9 studies), and **C**: image and tabular (3 studies).

https://doi.org/10.1371/journal.pdig.0000438.g003

This asymmetry implies the presence of possible publication bias, particularly pronounced in studies with smaller sample sizes. However, the trim-and-fill method corrected this asymmetry, rendering the distribution symmetrical (**S2 Fig and S3 Fig**). After using the trim-and-fill method to adjust for publication bias, the diagnostic odds ratio (DOR) has revealed that the effect size remains statistically significant (**S5 and S6 Tables**).

Risk of bias and applicability

The assessment of bias and applicability for 66 studies revealed moderate to low concerns (**Table 5** and **Fig 4**). Patient selection and reference standards were the primary concerns for bias and applicability. Many studies lacked the reporting of sample characteristics such as gender and age, limiting generalizability. Some studies did not report patient selection or reference standard computation methods [62,75,78]. Threshold adjustments in some studies might have led to overfitting, reducing the generalizability of the models [72]. Most studies exhibited applicability concerns and needed to be more easily generalizable to other populations. For example, one study [66] focused on patients visiting the emergency department for acute proximal femoral fracture, limiting generalizability to the general population. Another study included patients with existing vertebral fractures, reducing generalizability to the general population. Data preprocessing often involves the removal of occult fractures, with some studies excluding radiographic occult fractures requiring additional modalities for confirmation [53]. Other studies excluded images with uncertain, traumatic, or pathological fractures or those with insufficient quality or resolution [58]. A few studies did not provide specific locations for fracture types or specify which ones were included [12,70].

Discussion

Our systematic review and meta-analysis offer the most current and comprehensive evaluation of the diagnostic accuracy of Artificial Intelligence (AI) for predicting various osteoporotic fracture outcomes using various imaging modalities and data types. This study represents the first systematic review and quantitative meta-analysis of AI's diagnostic accuracy and comparison using different data types across multiple fracture outcomes. Our analysis reveals four major findings. First, AI provides high classification accuracy for fracture detection when utilizing imaging data, with a pooled sensitivity of 92% (95% CI: 90, 94). Convolutional neural networks with transfer learning exhibit significantly high accuracy when using image data in classifying fractures. Second, our study comprehensively reviews diagnostic accuracy among

Table 5. The result of methodological quality for 66 included studies in the assessment of the risk of bias and applicability.

First author (Year published)	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Almog et al. (2020) [12]	+	+	+	+	+	+	+
Bae et al. (2021) [7]	+	+	+	+	+	+	+
Beyaz et al. (2020) [67]	+	+	+	+	+	+	+
Burns et al. (2017) [8]	–	+	+	+	–	–	+
Chen et al. (2021) [28]	+	+	+	+	+	+	+
Chen et al. (2022) [46]	+	+	+	+	+	+	+
Cheng et al. (2019) [6]	+	+	+	+	+	+	+
Cheng et al. (2020) [29]	+	+	+	+	+	+	+
Cheng et al. (2021) [30]	+	+	+	+	+	+	+
Choi et al. (2020) [47]	+	+	+	+	–	+	+
Chou et al. (2022) [31]	+	+	+	+	+	+	+
Chung et al. (2018) [45]	+	+	–	+	–	+	–
Derkatch et al. (2019) [51]	+	+	+	+	+	+	+
Galassi et al. (2020) [68]	+	+	+	+	+	+	+
Guermazi et al. (2022) [52]	+	+	+	+	+	+	+
Gupta et al. (2020) [53]	+	+	+	+	+	+	+
Hayashi et al. (2022) [54]	+	+	+	+	–	+	+
Ho-Le et al. (2017) [79]	–	+	+	+	+	+	+
Inoue et al. (2022) [9]	+	+	+	+	–	+	+
Kim et al. (2018) [69]	+	+	+	+	+	+	+
Kitamura et al. (2020) [55]	+	+	+	+	+	+	+
Korfiatis et al. (2018) [81]	+	+	+	+	–	+	+
Kruse et al. (2017) [13]	+	+	+	+	+	+	+
Del Lama et al. (2022) [80]	+	+	+	+	+	+	+
Lemineur et al. (2007) [70]	+	+	–	+	–	+	–
Lindsey et al. (2018) [56]	+	+	+	+	+	+	+
Liu et al. (2015) [32]	+	+	+	+	+	+	+
Liu et al. (2022) [48]	+	+	+	+	+	+	+
Mawatari et al. (2020) [37]	+	+	+	+	+	+	+
Mehta et al. (2020) [57]	+	+	+	+	–	+	+
Minonzio et al. (2020) [71]	+	–	+	+	+	–	+
Monchka et al. (2021) [58]	O	+	+	+	–	+	+
Monchka et al. (2022) [59]	+	+	+	+	+	+	+
Mu et al. (2021) [49]	+	+	+	+	+	+	+
Murata et al. (2020) [38]	+	+	+	+	+	+	+
Mutasa et al. (2020) [60]	+	+	+	O	+	+	O
Nguyen et al. (2022) [61]	+	+	+	+	+	+	+

(*Continued*)

Table 5. (Continued)

First author (Year published)	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Nishiyama et al. (2014) [39]	O	+	+	+	+	+	+
Nissinen et al. (2021) [72]	+	+	O	+	O	+	O
Oakden-Rayner et al. (2022) [62]	+	+	–	+	O	+	–
Ozkaya et al. (2022) [73]	+	+	+	+	+	+	+
Raghavendra et al. (2018) [82]	O	+	+	+	O	+	+
Raisuddin et al. (2021) [74]	+	+	+	+	+	+	+
Ramos et al. (2022) [10]	+	+	+	+	+	+	+
Regnard et al. (2022) [75]	–	+	+	+	–	+	+
Rosenberg et al. (2022) [76]	+	+	+	+	+	+	+
Salehinejad et al. (2021) [83]	+	+	+	+	+	+	+
Sato et al. (2021) [40]	+	+	+	+	+	+	+
Small et al. (2021) [63]	+	+	+	+	+	+	+
Su et al. (2019) [64]	+	+	+	+	+	+	+
Tomita et al. (2018) [65]	+	+	+	+	+	+	+
Tseng et al. (2013) [33]	+	+	+	+	+	+	+
Ulivier et al. (2021) [77]	+	+	+	+	+	+	+
Urakawa et al. (2019) [41]	+	+	+	+	+	+	+
Ureten et al. (2022) [78]	–	+	+	+	–	+	+
Wang et al. (2022) [84]	+	+	+	+	+	+	+
Wu et al. (2020) [14]	+	+	+	+	+	+	+
Yabu et al. (2021) [11]	+	+	+	+	+	+	+
Yamada et al. (2020) [42]	+	+	+	+	+	+	+
Yamamoto et al. (2020) [43]	+	+	+	+	+	+	+
Yeh et al. (2022) [34]	+	–	+	+	+	+	+
Yi-Chu Li et al. (2021) [35]	+	+	+	+	+	+	+
Yoda et al. (2022) [44]	+	+	+	+	+	+	+
Yoon et al. (2021) [36]	+	+	+	+	+	+	+
Yu et al. (2020) [66]	+	+	+	+	+	+	+
Yuan Li et al. (2021) [50]	+	+	+	+	+	+	+

+: Low risk of bias/no concerns regarding applicability, –: High risk of bias/concerns regarding applicability, O: Unclear risk of bias/unclear whether there are concerns regarding applicability.

https://doi.org/10.1371/journal.pdig.0000438.t005

different image modalities with AI. While all image modalities provide comparable results, AI with radiograph images yields the highest results with a pooled sensitivity of 94% (95% CI: 90, 96). Third, our sensitivity analysis, employing the arcsine transformation, which was complemented by the primary analysis utilizing the logit transformation, provides the robustness of our findings. Both methodologies yielded similar results regarding pooled sensitivity and

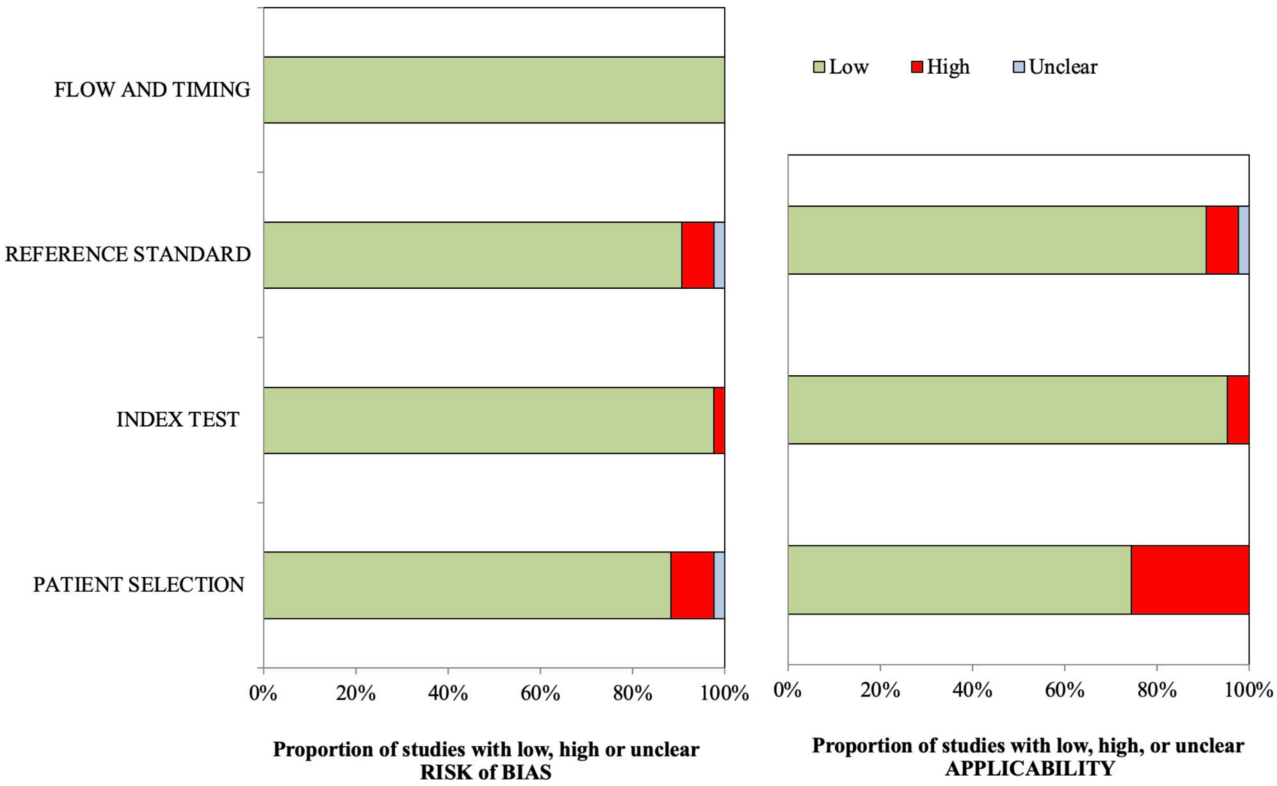

Fig 4. Summary of the Quality Assessment of Diagnostic Accuracy Studies for the risk of bias and applicability in the included 66 studies. The risk of bias was measured in four domains: patient selection, index test, reference standard, and flow and timing. The risk of applicability was evaluated with three domains: patient selection, index test, and reference.

https://doi.org/10.1371/journal.pdig.0000438.g004

specificity, which underscores the reliability and consistency of our findings. Fourth, significant flaws were observed in the study design and reporting of AI for real-world applicability. For example, only a few studies described the patient characteristics of data, and only half (n = 33) reported the hyperparameter selection process.

Our findings align with other systematic reviews and meta-analyses [15,16], showing that AI demonstrates considerably higher pooled sensitivity and specificity. However, inconsistent results have been observed when comparing different image modalities in fracture detection. External validation enables a more robust demonstration of clinical utility versus simple internal train/test cross-validation. Our study shows that only thirteen studies (20%) out of sixty-six performed external validation. The limitation of validating in an external dataset is the lack of availability of large, labeled datasets due to resistance to sharing data across institutions because of patient privacy issues and the necessity of experts for labeling the datasets. Although external validation enhances the robustness of AI systems, it could potentially attenuate their impact on the system. Consequently, it's crucial to acknowledge that external validation might not always be advisable due to the potential impact of factors like sample size and the diversity of the training set. Two systematic reviews [89,90] provide valuable insights into the current limitations of AI studies. A broad discussion of possible solutions is necessary because methodological challenges, risk of bias, and applicability concerns can arise in AI during all stages of development, including data curation, model selection, implementation, and validation. Both reviews recommend that researchers follow standardized reporting guidelines to determine the risk of bias and improve methodological quality assessment.

Our study has limitations; the major one is that only a few studies that employed tabular data or combined tabular and image data are eligible. Second, we excluded non-English-language articles, which may have overlooked some studies published in a different language. Third, many of these included studies had study design flaws. They were classified as having great concern for bias and applicability, limiting the conclusions that could be drawn from the meta-analysis because studies with a high risk of bias and applicability overestimated algorithm performance.

This systematic review and meta-analysis have important implications for clinical practice. Given the high diagnostic performance of AI, these techniques could be integrated into existing fracture risk assessment tools to enhance the identification of patients at risk and facilitate early intervention. Healthcare professionals should be trained in interpreting and applying these methods in clinical practice.

This study observed superior prediction performance with single radiograph input data over multimodal imaging, which can be attributed to the radiographs' consistent and standardized anatomical view, reducing noise and variability inherent in multimodal inputs [91]. Radiographs precisely capture fracture-relevant features, while added modalities like CT and MRI can diversify and possibly weaken these key features [92]. Multimodal inputs can also elevate overfitting risks, particularly with limited datasets [93]. Radiographs, being more accessible and cost-effective than CT or MRI, allow for larger, representative datasets enhancing model performance. The decision between single radiographs and multimodal inputs should be rooted in the research context, data availability, and prediction objectives. Despite the evident advantages of radiographs, specific scenarios may warrant multimodal integration for improved predictions. We also observed that solely relying on image data produced better AUC values than combining it with tabular data. Image data's richness and direct relevance to fracture detection offer clear diagnostic advantages [94]. Convolutional neural networks (CNNs), identified in our study, are adept at processing this data, emphasizing subtle fracture-related visual nuances [95]. In contrast, tabular data could infuse noise and inconsistencies. Sole image data ensures focus on vital visual features and offers a more standardized data format than diverse tabular inputs.

Further research is needed to address the limitations identified in the included studies and to explore the performance of specific ML and DL algorithms. Researchers should provide more detailed information about their study populations and methods, including patient selection, fracture type location, and the reference standard used. Future studies should also investigate the impact of factors such as training dataset size, model architecture, and the inclusion of clinical and demographic variables on the diagnostic performance of AI. Future research will help develop more accurate and generalizable models for predicting osteoporotic fractures and inform evidence-based clinical practice. Several novel diagnostic meta-analysis methodologies have recently been introduced [96–98]. Nevertheless, due to the limited sample sizes within selected studies focusing on fractures beyond vertebral and hip injuries and studies involving tabular and tabular and image data types, incorporating these methodologies into our present study was unfeasible. While we acknowledge their potential applicability, the current study's unique characteristics led us to refrain from their implementation. We will implement these methodologies in our forthcoming investigations, particularly as more comprehensive studies become available. In aid of future researchers, we provide an array of crucial challenges and their potential resolutions pertinent to applying machine learning or deep learning for fracture diagnosis (**S7 Table**).

In conclusion, our meta-analysis highlights the high diagnostic accuracy of AI in various fracture outcomes. As AI demonstrates reliable results in fracture detection, it holds the potential to streamline fracture diagnosis in healthcare systems. However, transparent reporting of study methods and designs for AI development and validation is essential to ensure their real-world applicability. By addressing the current research landscape's limitations and promoting standardized guidelines, we can facilitate the integration of AI technologies into clinical practice and enhance the prediction of osteoporotic fractures, ultimately leading to improved patient care.

Supporting information

S1 PRISMA Checklist. PRISMA DTA Checklist.
(DOCX)

S1 Text. The search term used for each engine: 1) PubMed, 2) Web of Science, and 3) IEEE.
(DOCX)

S1 Table. A characteristic of 57 selected studies for Image modality, Image Data Type, and Data Source.
(DOCX)

S2 Table. The data source of 9 selected studies used tabular data, and 3 studies (in bold) used both tabular and image data.
(DOCX)

S3 Table. A characteristic of 66 selected studies for the unbalanced outcome, a technique used for an unbalanced outcome, data preprocessing, hyperparameters optimization, and performance measurement used.
(DOCX)

S4 Table. A summary of the contingency table for 66 selected studies.
(DOCX)

S5 Table. Summary of Publication Bias Assessment across different fracture outcomes. TF: Trim and Fill method, DOR: Diagnostic Odds Ratio, CI: Confidence Interval.
(DOCX)

S6 Table. Summary of Publication Bias Assessment across different data types. TF: Trim and Fill method, DOR: Diagnostic Odds Ratio, CI: Confidence Interval.
(DOCX)

S7 Table. Overview of Key Challenges and Potential Resolutions in the Utilization of Machine Learning or Deep Learning for Fracture Diagnosis.
(DOCX)

S1 Fig. Contour-Enhanced Funnel Plot for Publication Bias Assessment across Different Fracture Outcomes.
(DOCX)

S2 Fig. Contour-Enhanced Funnel Plot for Publication Bias Assessment across Different Fracture Outcomes after Employing the Trim & Fill Method. The open circle represents the "filled" studies from the Trim & Fill Method in each fracture outcome plot.
(DOCX)

S3 Fig. Contour-Enhanced Funnel Plot: Evaluating Publication Bias Across Various Data Types. The top row illustrates the funnel plot encompassing all studies. The second row shows the Contour-Enhanced Funnel Plot for Publication Bias Assessment after employing the Trim & Fill Method. The open circle designates the studies "filled" through the Trim & Fill Method

within each contour-enhanced funnel plot in the second row.
(DOCX)

Acknowledgments

This research was partially conducted under the affiliation of the Nevada Institute of Personalized Medicine, College of Sciences (QW, JJ, and JD), Department of Epidemiology and Biostatistics, School of Public Health (QW and JJ), Department of Mathematical Sciences, College of Sciences (BL), the University of Nevada, Las Vegas.

Author Contributions

Conceptualization: Jongyun Jung, Jingyuan Dai, Qing Wu.

Data curation: Jongyun Jung, Jingyuan Dai.

Formal analysis: Jongyun Jung.

Funding acquisition: Qing Wu.

Investigation: Qing Wu.

Methodology: Jongyun Jung, Qing Wu.

Resources: Qing Wu.

Software: Jongyun Jung.

Validation: Jongyun Jung.

Visualization: Jongyun Jung.

Writing – original draft: Jongyun Jung, Qing Wu.

Writing – review & editing: Jongyun Jung, Jingyuan Dai, Bowen Liu, Qing Wu.

References

1. Court-Brown CM, Caesar B. Epidemiology of adult fractures: A review. Injury. 2006; 37: 691–697. https://doi.org/10.1016/j.injury.2006.04.130 PMID: 16814787
2. Wu A-M, Bisignano C, James SL, Abady GG, Abedi A, Abu-Gharbieh E, et al. Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet Healthy Longev. 2021; 2: e580–e592. https://doi.org/10.1016/S2666-7568(21)00172-0 PMID: 34723233
3. Pike C, Birnbaum HG, Schiller M, Sharma H, Burge R, Edgell ET. Direct and Indirect Costs of Non-Vertebral Fracture Patients with Osteoporosis in the US. PharmacoEconomics. 2010; 28: 395–409. https://doi.org/10.2165/11531040-000000000-00000 PMID: 20402541
4. Borgström F, Karlsson L, Ortsäter G, Norton N, Halbout P, Cooper C, et al. Fragility fractures in Europe: burden, management and opportunities. Arch Osteoporos. 2020; 15: 59. https://doi.org/10.1007/s11657-020-0706-y PMID: 32306163
5. Williamson S, Landeiro F, McConnell T, Fulford-Smith L, Javaid MK, Judge A, et al. Costs of fragility hip fractures globally: a systematic review and meta-regression analysis. Osteoporos Int. 2017; 28: 2791–2800. https://doi.org/10.1007/s00198-017-4153-6 PMID: 28748387
6. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol. 2019; 29: 5469–5477. https://doi.org/10.1007/s00330-019-06167-y PMID: 30937588
7. Bae J, Yu S, Oh J, Kim TH, Chung JH, Byun H, et al. External Validation of Deep Learning Algorithm for Detecting and Visualizing Femoral Neck Fracture Including Displaced and Non-displaced Fracture on Plain X-ray. J Digit Imaging. 2021; 34: 1099–1109. https://doi.org/10.1007/s10278-021-00499-2 PMID: 34379216

8. Burns JE, Yao J, Summers RM. Vertebral body compression fractures and bone density: Automated detection and classification on CT Images. Radiology. 2017; 284: 788–797. https://doi.org/10.1148/radiol.2017162100 PMID: 28301777

9. Inoue T, Maki S, Furuya T, Mikami Y, Mizutani M, Takada I, et al. Automated fracture screening using an object detection algorithm on whole-body trauma computed tomography. Sci Rep. 2022; 12: 16549. https://doi.org/10.1038/s41598-022-20996-w PMID: 36192521

10. Ramos J. S., de Aguiar E. J., Belizario I. V., Costa M. V. L., Maciel J. G., Cazzolato M. T., et al. Analysis of vertebrae without fracture on spine MRI to assess bone fragility: A Comparison of Traditional Machine Learning and Deep Learning. 2022; 78–83. https://doi.org/10.1109/CBMS55023.2022.00021

11. Yabu A, Hoshino M, Tabuchi H, Takahashi S, Masumoto H, Akada M, et al. Using artificial intelligence to diagnose fresh osteoporotic vertebral fractures on magnetic resonance images. Spine J. 2021; 000: 1–7. https://doi.org/10.1016/j.spinee.2021.03.006 PMID: 33722728

12. Almog YA, Rai A, Zhang P, Moulaison A, Powell R, Mishra A, et al. Deep Learning with Electronic Health Records for Short-Term Fracture Risk Identification: Crystal Bone Algorithm Development and Validation. J Med Internet Res. 2020;22. https://doi.org/10.2196/22550 PMID: 32956069

13. Kruse C, Eiken P, Vestergaard P. Machine Learning Principles Can Improve Hip Fracture Prediction. Calcif Tissue Int. 2017; 100: 348–360. https://doi.org/10.1007/s00223-017-0238-7 PMID: 28197643

14. Wu Q, Nasoz F, Jung J, Bhattarai B, Han MV. Machine Learning Approaches for Fracture Risk Assessment: A Comparative Analysis of Genomic and Phenotypic Data in 5130 Older Men. Calcif Tissue Int. 2020; 1–9. https://doi.org/10.1007/s00223-020-00734-y PMID: 32728911

15. Kuo Rachel Y L, Harrison Conrad, Curran Terry-Ann, Jones Benjamin, Freethy Alexander, Cussons David, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. 2022; 304: 50–62. https://doi.org/10.1148/radiol.211785 PMID: 35348381

16. Yang S, Yin B, Cao W, Feng C, Fan G, He S. Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. Clin Radiol. 2020; 75: 713.e17–713.e28. https://doi.org/10.1016/j.crad.2020.05.021 PMID: 32591230

17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021; 372: n71. https://doi.org/10.1136/bmj.n71 PMID: 33782057

18. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: A single indicator of test performance. J Clin Epidemiol. 2003; 56: 1129–1135. https://doi.org/10.1016/s0895-4356(03)00177-x PMID: 14615004

19. Sterne JAC, Gavaghan D, Egger M. Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000; 53: 1119–1129. https://doi.org/10.1016/s0895-4356(00)00242-0 PMID: 11106885

20. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika. 1934; 26: 404. https://doi.org/10.2307/2331986

21. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005; 58: 882–893. https://doi.org/10.1016/j.jclinepi.2005.01.016 PMID: 16085191

22. Barendregt JJ, Doi SA, Lee YY, Norman RE, Vos T. Meta-analysis of prevalence. J Epidemiol Community Health. 2013; 67: 974–978. https://doi.org/10.1136/jech-2013-203104 PMID: 23963506

23. Team RC. R: A language and environment for statistical computing. R Found Stat Comput Vienna Austria. 2019;3. Available: https://www.r-project.org/

24. Schwarzer G. meta: An R Package for Meta-Analysis. R News. 2007. Available: http://cran.r-project.org/doc/Rnews/

25. Doebler P, Holling H. Meta-Analysis of Diagnostic Accuracy with mada. Compr R Arch Netw. 2012; 1–15.

26. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol. 2008; 61: 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010 PMID: 18538991

27. Whiting PF, Rutjes AWW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011; 155: 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009 PMID: 22007046

28. Chen HY, Hsu BWY, Yin YK, Lin FH, Yang TH, Yang RS, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. PLoS ONE. 2021; 16: 1–10. https://doi.org/10.1371/journal.pone.0245992 PMID: 33507982

29. Cheng CT, Chen CC, Cheng FJ, Chen HW, Su YS, Yeh CN, et al. A human-algorithm integration system for hip fracture detection on plain radiography: System development and validation study. JMIR Med Inform. 2020; 8: 1–13. https://doi.org/10.2196/19416 PMID: 33245279

30. Cheng CT, Wang Y, Chen HW, Hsiao PM, Yeh CN, Hsieh CH, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat Commun. 2021;12. https://doi.org/10.1038/s41467-021-21311-3 PMID: 33594071

31. Chou PH, Jou TH, Wu HH, Yao YC, Lin HH, Chang MC, et al. Ground truth generalizability affects performance of the artificial intelligence model in automated vertebral fracture detection on plain lateral radiographs of the spine. Spine J. 2022; 22: 511–523. https://doi.org/10.1016/j.spinee.2021.10.020 PMID: 34737066

32. Liu Q, Cui X, Chou YC, Abbod MF, Lin J, Shieh JS. Ensemble artificial neural networks applied to predict the key risk factors of hip bone fracture for elders. Biomed Signal Process Control. 2015; 21: 146–156. https://doi.org/10.1016/j.bspc.2015.06.002

33. Tseng WJ, Hung LW, Shieh JS, Abbod MF, Lin J. Hip fracture risk assessment: Artificial neural network outperforms conditional logistic regression in an age- and sex-matched case control study. BMC Musculoskelet Disord. 2013;14. https://doi.org/10.1186/1471-2474-14-207 PMID: 23855555

34. Yeh LR, Zhang Y, Chen JH, Liu YL, Wang AC, Yang JY, et al. A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: retrospective training and validation of ResNet. Eur Spine J. 2022; 31: 2022–2030. https://doi.org/10.1007/s00586-022-07121-1 PMID: 35089420

35. Li YC, Chen HH, Horng-Shing Lu H, Hondar Wu HT, Chang MC, Chou PH. Can a Deep-learning Model for the Automated Detection of Vertebral Fractures Approach the Performance Level of Human Subspecialists? Clin Orthop Relat Res. 2021; 479: 1598–1612. https://doi.org/10.1097/CORR.0000000000001685 PMID: 33651768

36. Yoon AP, Lee Y-L, Kane RL, Kuo C-F, Lin C, Chung KC. Development and Validation of a Deep Learning Model Using Convolutional Neural Networks to Identify Scaphoid Fractures in Radiographs. JAMA Netw Open. 2021; 4: e216096. https://doi.org/10.1001/jamanetworkopen.2021.6096 PMID: 33956133

37. Mawatari T, Hayashida Y, Katsuragawa S, Yoshimatsu Y, Hamamura T, Anai K, et al. The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. Eur J Radiol. 2020; 130: 109188. https://doi.org/10.1016/j.ejrad.2020.109188 PMID: 32721827

38. Murata K, Endo K, Aihara T, Suzuki H, Sawaji Y, Matsuoka Y, et al. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. Sci Rep. 2020; 10: 1–8. https://doi.org/10.1038/s41598-020-76866-w PMID: 33208824

39. Nishiyama KK, Ito M, Harada A, Boyd SK. Classification of women with and without hip fracture based on quantitative computed tomography and finite element analysis. Osteoporos Int. 2014; 25: 619–626. https://doi.org/10.1007/s00198-013-2459-6 PMID: 23948875

40. Sato Y, Takegami Y, Asamoto T, Ono Y, Hidetoshi T, Goto R. Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. BMC Musculoskelet Disord. 2021; 1–10.

41. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol. 2019; 48: 239–244. https://doi.org/10.1007/s00256-018-3016-3 PMID: 29955910

42. Yamada Y, Maki S, Kishida S, Nagai H, Arima J, Yamakawa N, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. Acta Orthop. 2020; 91: 699–704. https://doi.org/10.1080/17453674.2020.1803664 PMID: 32783544

43. Yamamoto N, Rahman R, Yagi N, Hayashi K, Maruo A, Muratsu H, et al. An automated fracture detection from pelvic CT images with 3-D convolutional neural networks. 2020 Int Symp Community-Centric Syst CcS 2020. 2020; 3–8. https://doi.org/10.1109/CcS49175.2020.9231453

44. Yoda T, Maki S, Furuya T, Yokota H, Matsumoto K, Takaoka H, et al. Automated Differentiation Between Osteoporotic Vertebral Fracture and Malignant Vertebral Fracture on MRI Using a Deep Convolutional Neural Network. Spine. 2022; 47: E347–E352. https://doi.org/10.1097/BRS.0000000000004307 PMID: 34919075

45. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. 2018; 89: 468–473. https://doi.org/10.1080/17453674.2018.1453714 PMID: 29577791

46. Chen W, Liu X, Li K, Luo Y, Bai S, Wu J, et al. A deep-learning model for identifying fresh vertebral compression fractures on digital radiography. Eur Radiol. 2022; 32: 1496–1505. https://doi.org/10.1007/s00330-021-08247-4 PMID: 34553256

47. Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, et al. Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. Invest Radiol. 2020; 55: 101–110. https://doi.org/10.1097/RLI.0000000000000615 PMID: 31725064

48. Liu P, Lu L, Chen Y, Huo T, Xue M, Wang H, et al. Artificial intelligence to detect the femoral intertrochanteric fracture: The arrival of the intelligent-medicine era. Front Bioeng Biotechnol. 2022;10. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2022.927926 PMID: 36147533

49. Mu L, Qu T, Dong D, Li X, Pei Y, Wang Y, et al. Fine-Tuned Deep Convolutional Networks for the Detection of Femoral Neck Fractures on Pelvic Radiographs: A Multicenter Dataset Validation. IEEE Access. 2021; 9: 78495–78503. https://doi.org/10.1109/ACCESS.2021.3082952

50. Li Y, Zhang Y, Zhang E, Chen Y, Wang Q, Liu K, et al. Differential diagnosis of benign and malignant vertebral fracture on CT using deep learning. Eur Radiol. 2021. https://doi.org/10.1007/s00330-021-08014-5 PMID: 33993335

51. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Michael Davidson J, Leslie WD. 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: 404–411. https://doi.org/10.1148/radiol.2019190201 PMID: 31526255

52. Guermazi A, Tannoury C, Kompel AJ, Murakami AM, Ducarouge A, Gillibert A, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. Radiology. 2022; 302: 627–636. https://doi.org/10.1148/radiol.210937 PMID: 34931859

53. Gupta V, Demirer M, Bigelow M, Yu SM, Yu JS, Prevedello LM, et al. Using Transfer Learning and Class Activation Maps Supporting Detection and Localization of Femoral Fractures on Anteroposterior Radiographs. Proc—Int Symp Biomed Imaging. 2020;2020-April: 1526–1529. https://doi.org/10.1109/ISBI45749.2020.9098436

54. Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard NE, et al. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skelet Radiol. 2022; 51: 2129–2139. https://doi.org/10.1007/s00256-022-04070-0 PMID: 35522332

55. Kitamura G. Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection. Eur J Radiol. 2020; 130: 109139. https://doi.org/10.1016/j.ejrad.2020.109139 PMID: 32623269

56. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018; 115: 11591–11596. https://doi.org/10.1073/pnas.1806905115 PMID: 30348771

57. Mehta SD, Sebro R. Computer-Aided Detection of Incidental Lumbar Spine Fractures from Routine Dual-Energy X-Ray Absorptiometry (DEXA) Studies Using a Support Vector Machine (SVM) Classifier. J Digit Imaging. 2020; 33: 204–210. https://doi.org/10.1007/s10278-019-00224-0 PMID: 31062114

58. Monchka BA, Kimelman D, Lix LM, Leslie WD. Feasibility of a generalized convolutional neural network for automated identification of vertebral compression fractures: The Manitoba Bone Mineral Density Registry. Bone. 2021; 150: 116017. https://doi.org/10.1016/j.bone.2021.116017 PMID: 34020078

59. Monchka BA, Schousboe JT, Davidson MJ, Kimelman D, Hans D, Raina P, et al. Development of a manufacturer-independent convolutional neural network for the automated identification of vertebral compression fractures in vertebral fracture assessment images using active learning. Bone. 2022; 161: 116427. https://doi.org/10.1016/j.bone.2022.116427 PMID: 35489707

60. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ. Advanced Deep Learning Techniques Applied to Automated Femoral Neck Fracture Detection and Classification. J Digit Imaging. 2020;csvgrdgnfmb mfs 33: 1209–1217. https://doi.org/10.1007/s10278-020-00364-8 PMID: 32583277

61. Nguyen T, Maarek R, Hermann AL, Kammoun A, Marchi A, Khelifi-Touhami MR, et al. Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. Pediatr Radiol. 2022; 52: 2215–2226. https://doi.org/10.1007/s00247-022-05496-3 PMID: 36169667

62. Oakden-rayner L, Gale W, Bonham TA, Lungren MP, Carneiro G, Bradley AP, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. The Lancet. 2022; 7500: 4–8. https://doi.org/10.1016/S2589-7500(22)00004-8 PMID: 35396184

63. JE S, Osler P, Paul AB, Kunst M. CT Cervical Spine Fracture Detection Using a Convolutional Neural Network. AJNR Am J Neuroradiol. 2021; 42: 1341–1347. https://doi.org/10.3174/ajnr.A7094 PMID: 34255730

64. Su Y, Kwok TCY, Cummings SR, Yip BHK, Cawthon PM. Can Classification and Regression Tree Analysis Help Identify Clinically Meaningful Risk Groups for Hip Fracture Prediction in Older American Men (The MrOS Cohort Study)? JBMR Plus. 2019; 3: 1–6. https://doi.org/10.1002/jbm4.10207 PMID: 31687643

65. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med. 2018; 98: 8–15. https://doi.org/10.1016/j.compbiomed.2018.05.011 PMID: 29758455

66. Yu JS, Yu SM, Erdal BS, Demirer M, Gupta V, Bigelow M, et al. Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. Clin Radiol. 2020; 75: 237.e1-237.e9. https://doi.org/10.1016/j.crad.2019.10.022 PMID: 31787211

67. Beyaz S, Açici K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg. 2020; 31: 175–183. https://doi.org/10.5606/ehc.2020.72163 PMID: 32584712

68. Galassi A, Martín-Guerrero JD, Villamor E, Monserrat C, Rupérez MJ. Risk Assessment of Hip Fracture Based on Machine Learning. Appl Bionics Biomech. 2020. https://doi.org/10.1155/2020/8880786 PMID: 33425008

69. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018; 73: 439–445. https://doi.org/10.1016/j.crad.2017.11.015 PMID: 29269036

70. Lemineur G, Harba R, Kilic N, Ucan ON, Osman O, Benhamou L. Efficient estimation of osteoporosis using Artificial Neural Networks. IECON Proc Ind Electron Conf. 2007; 3039–3044. https://doi.org/10.1109/IECON.2007.4460070

71. Minonzio JG, Cataldo B, Olivares R, Ramiandrisoa D, Soto R, Crawford B, et al. Automatic classifying of patients with non-traumatic fractures based on ultrasonic guided wave spectrum image using a dynamic support vector machine. IEEE Access. 2020; 8: 194752–194764. https://doi.org/10.1109/ACCESS.2020.3033480

72. Nissinen T, Suoranta S, Saavalainen T, Sund R, Hurskainen O. Detecting pathological features and predicting fracture risk from dual-energy X-ray absorptiometry images using deep learning. Bone Rep. 2021;14. https://doi.org/10.1016/j.bonr.2021.101070 PMID: 33997147

73. Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, Karakaya Z. Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. Eur J Trauma Emerg Surg. 2022; 48: 585–592. https://doi.org/10.1007/s00068-020-01468-0 PMID: 32862314

74. Raisuddin AM, Vaattovaara E, Nevalainen M, Nikki M, Järvenpää E, Makkonen K, et al. Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep. 2021; 11: 6006. https://doi.org/10.1038/s41598-021-85570-2 PMID: 33727668

75. Regnard NE, Lanseur B, Ventre J, Ducarouge A, Clovis L, Lassalle L, et al. Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays. Eur J Radiol. 2022; 154: 110447. https://doi.org/10.1016/j.ejrad.2022.110447 PMID: 35921795

76. Rosenberg GS, Cina A, Schiró GR, Giorgi PD, Gueorguiev B, Alini M, et al. Artificial Intelligence Accurately Detects Traumatic Thoracolumbar Fractures on Sagittal Radiographs. Medicina (Mex). 2022; 58: 998. https://doi.org/10.3390/medicina58080998 PMID: 35893113

77. Ulivieri FM, Rinaudo L, Piodi LP, Messina C, Sconfienza LM, Sardanelli F, et al. Bone strain index as a predictor of further vertebral fracture in osteoporotic women: An artificial intelligence-based analysis. PLoS ONE. 2021; 16: 1–13. https://doi.org/10.1371/journal.pone.0245967 PMID: 33556061

78. Üreten K, Sevinç HF, İğdeli U, Onay A, Maraş Y. Use of deep learning methods for hand fracture detection from plain hand radiographs. Ulus Travma Acil Cerrahi Derg. 2022; 28: 196–201. https://doi.org/10.14744/tjtes.2020.06944 PMID: 35099027

79. Ho-Le TP, Center JR, Eisman JA, Nguyen TV, Nguyen HT. Prediction of hip fracture in post-menopausal women using artificial neural network approach. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS. 2017; 4207–4210. https://doi.org/10.1109/EMBC.2017.8037784 PMID: 29060825

80. Del Lama RS, Candido RM, Chiari-Correia NS, Nogueira-Barbosa MH, de Azevedo-Marques PM, Tinós R. Computer-Aided Diagnosis of Vertebral Compression Fractures Using Convolutional Neural Networks and Radiomics. J Digit Imaging. 2022; 35: 446–458. https://doi.org/10.1007/s10278-022-00586-y PMID: 35132524

81. Korfiatis VC, Tassani S, Matsopoulos GK. A New Ensemble Classification System For Fracture Zone Prediction Using Imbalanced Micro-CT Bone Morphometrical Data. IEEE J Biomed Health Inform. 2018; 22: 1189–1196. https://doi.org/10.1109/JBHI.2017.2723463 PMID: 28692998

82. Raghavendra U, Bhat NS, Gudigar A, Acharya UR. Automated system for the detection of thoracolumbar fractures using a CNN architecture. Future Gener Comput Syst. 2018; 85: 184–189. https://doi.org/10.1016/j.future.2018.03.023

83. Salehinejad H, Ho E, Lin H, Crivellaro P, Samorodova O, Arciniegas MT, et al. Deep Sequential Learning For Cervical Spine Fracture Detection On Computed Tomography Imaging. IEEE 18th Int Symp Biomed Imaging. 2021; 1911–1914.

84. Yuzhao W, Tian B, Tong L, Lang H. Osteoporotic Vertebral Fracture Classification in X-rays Based on a Multi-modal Semantic Consistency Network. J BIONIC Eng. 2022; 19: 1816–1829. https://doi.org/10.1007/s42235-022-00234-9

85. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-Decem: 2818–2826. https://doi.org/10.1109/CVPR.2016.308

86. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002; 16: 321–357. https://doi.org/10.1613/jair.953

87. Bergstra J, Bengio Y. Random Search For Hyper-Parameter Optimization. J Mach Learn Res. 2012; 13: 281–305.

88. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. J Mach Learn Res. 2018; 18: 1–52.

89. Zhou Q, Chen Z, Cao Y, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. Npj Digit Med. 2021; 4: 1–12. https://doi.org/10.1038/s41746-021-00524-2 PMID: 34711955

90. Navarro CLA, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021; 375: n2281. https://doi.org/10.1136/bmj.n2281 PMID: 34670780

91. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018; 18: 500–510. https://doi.org/10.1038/s41568-018-0016-5 PMID: 29777175

92. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging —"how-to" guide and critical reflection. Insights Imaging. 2020; 11: 91. https://doi.org/10.1186/s13244-020-00887-2 PMID: 32785796

93. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022; 22: 114–126. https://doi.org/10.1038/s41568-021-00408-3 PMID: 34663944

94. Krupinski EA. Current perspectives in medical image perception. Atten Percept Psychophys. 2010; 72: 1205–1217. https://doi.org/10.3758/APP.72.5.1205 PMID: 20601701

95. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. J Big Data. 2019; 6: 60. https://doi.org/10.1186/s40537-019-0197-0

96. Preisser JS, Inan G, Powers JM, Chu H. A population-averaged approach to diagnostic test meta-analysis. Biom J. 2019; 61: 126–137. https://doi.org/10.1002/bimj.201700187 PMID: 30370548

97. Xiaoye Ma Chu YL, Chen Yong, Stijnen Theo, Haitao. Meta-Analysis of Diagnostic Tests. Handbook of Meta-Analysis. Chapman and Hall/CRC; 2020.

98. Liu Z, Al Amer FM, Xiao M, Xu C, Furuya-Kanamori L, Hong H, et al. The normality assumption on between-study random effects was questionable in a considerable number of Cochrane meta-analyses. BMC Med. 2023; 21: 112. https://doi.org/10.1186/s12916-023-02823-9 PMID: 36978059