Int J Burn Trauma 2026;16(3):118-129
www.IJBT.org /ISSN:2160-2026/IJBT0170189

*Review Article*

# Using artificial intelligence for distinguishing benign and malignant vertebral compression fractures by computed tomography: a scoping review

Mobina Mohammadrezaee[1], Rozhin Bakhshi[2], Amirreza Khalaji[3], Imaneh Jarideh[4], Seyed Hamed Tooyserkani[6], Maryam Masjedi Esfahani[4], Mehrangiz Dadkhah[4], Hossein Badiee[5], Seyed Amirhossein Dormiani Tabatabaei[6], Farshad Riahi[4], Shahin Fesharaki[4]

*[1]School of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran; [2]School of Medicine, Islamic Azad University, Najafabad Branch, Isfahan, Iran; [3]Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran; [4]Department of Radiology, Isfahan University of Medical Sciences, Isfahan, Iran; [5]Department of Neurosurgery, Sina Hospital, Tehran University of Medical Sciences, Tehran, Iran; [6]School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran*

Received November 22, 2025; Accepted June 10, 2026; Epub June 25, 2026; Published June 30, 2026

**Abstract:** Objective: The purpose of this scoping review is to compile the evidence regarding the use of computed tomography (CT) to distinguish between benign and malignant vertebral compression fractures (VCFs) in artificial intelligence (AI) applications. Methods: We systematically searched PubMed/MEDLINE, DOAJ, and ScienceDirect through October 2025 for studies reporting AI-based diagnostic approaches (machine learning with radiomics features, deep learning with convolutional neural networks, or hybrid models) applied to CT images in adult patients with VCFs. Eligible studies required diagnostic performance metrics and definitive reference standards. Results: Six retrospective studies from Asia and Europe have been found, involving 1,767 participants with 3,408 vertebrae, including 1,257 benign and 1,442 malignant vertebral fractures. The diagnostic performance of AI algorithms is high, achieving an area under the curve (AUC) score from 0.76 to 0.99. Deep learning approaches achieved performance comparable to experienced radiologists, while radiomics-based methods provided interpretable quantitative features. Hybrid models combining both approaches showed synergistic benefits. Three studies with external validation confirmed reasonable generalizability, though with some performance reduction in independent datasets. Conclusions: AI-assisted evaluation of CT scans also has good potential in differentiating benign from malignant VCFs, which could reduce unnecessary MRI referrals and improve the initial evaluation time. However, the data regarding this are limited because of retrospective studies with variable methods and limited multicenter data, which are still required to implement this in routine clinical practice.

**Keywords:** Vertebral compression fracture, artificial intelligence, deep learning, radiomics, computed tomography

## Introduction

Vertebral compression fractures (VCFs) represent one of the most common skeletal complications, affecting approximately 25% of postmenopausal women and up to 40% of women aged 80 years and older [1, 2]. With the United States having 1 to 1.5 million new cases each year and with the world population aging, the need for accurate and prompt diagnosis of these fractures is becoming more critical for the healthcare system [2]. A single compression fracture elevates the risk of subsequent fractures fivefold, while patients with multiple fractures (two or more) face a more than twelvefold increase in this risk. This progressive cascade of skeletal fragility underscores the critical importance of early detection and timely intervention [3].

The critical clinical question is not simply whether a vertebral fracture is present, but rather whether it is benign or malignant. This distinction directly determines patient management. Multiple guidelines and reviews support conservative management as first-line

https://doi.org/10.62347/ZRQA3915

treatment for stable benign osteoporotic VCFs without neurologic deficit or marked instability. Non-surgical care (analgesia, short bracing, osteoporosis therapy, and physiotherapy) effectively reduces acute pain and usually achieves good functional recovery, with long-term function and quality of life comparable to vertebral augmentation in many studies. Recent network meta-analyses emphasize that conservative regimens should prioritize adequate pain control to enable early mobilization, thereby minimizing complications of prolonged bed rest in elderly patients. Surgery is thus reserved for unstable fractures, neurologic compromise, or refractory pain despite optimized conservative therapy [4-6]. On the other hand, Malignant fractures resulting from metastatic disease or primary spinal tumors require aggressive multimodal therapeutic strategies, including chemotherapy, radiation therapy, and surgical intervention [7]. In these instances, diagnostic misclassification results in significant delays in the commencement of suitable treatment, thereby severely undermining patient outcomes and prognosis [2]. This diagnostic challenge is particularly acute in specific patient populations: elderly patients at risk for both osteoporosis and cancer, patients with a history of malignancy, and patients in whom VCFs are incidentally discovered during imaging for other indication.

Despite all its technical advantages, magnetic resonance imaging (MRI) still represents the gold standard approach to differentiation of benign and malignant VCFs due to superior soft tissue contrast and identification of abnormal changes in the bone marrow [8-10]. Nevertheless, there exist several practical limitations, making routine use of MRI for VCF diagnosis problematic for many individuals. Those issues include long-lasting nature of investigations, considerable costs associated with MRI scanning, and limited availability of MRI in emergency departments and resource-limited clinical environments. There is also an array of individual factors affecting the possibility to perform MRI scanning, including presence of metal implants, extreme claustrophobia, or poor renal function, if necessary to administer contrast agents [11, 12]. All those obstacles result in noticeable gap between ideal algorithm for diagnosis of the disease in question and real life situation, making necessary the development of additional diagnostic approaches.

Another diagnostic option is represented by computed tomography (CT) investigation. Several features distinguish CT from MRI as the preferable diagnostic modality, including shorter duration of investigations, low costs, wide availability, and minor contraindications [13, 14]. All those characteristics provide a base for preferential use of CT in emergency departments and numerous other settings.

Notwithstanding the above-listed benefits, CT imaging comes with intrinsic limitations related to diagnostic shortcomings associated with distinguishing soft tissues. Traditional CT imaging relies heavily on morphological signs, such as the presence of disrupted cortices, paravertebral soft-tissue masses, altered pedicles, and bone destruction patterns. In practice, however, morphological signs tend to lack specificity. While benign osteoporotic fractures might manifest themselves through an irregular margin and cortical defects, metastasis could appear similarly to benign compression fractures [15]. Such diagnostic uncertainty would pose significant problems in clinical practice, requiring supplementary investigations using MRI, increasing health care expenditures, and delaying treatment.

Artificial intelligence (AI) technologies have gained recognition as game-changers in medical imaging by virtue of their extraordinary ability to solve problems beyond human capability in medical imaging [16]. The use of artificial intelligence in radiology can be categorized into three main categories, which reflect various ways in which AI can facilitate health care delivery. They involve the automation of traditionally routine procedures, assistance in clinical decision making processes, and support in clinical situations where diagnostic uncertainty prevails [17, 18].

In terms of VCF classification on CT imaging, two key approaches of AI have been identified. Radiomics involves the extraction of quantifiable image-derived features that describe tissue characteristics such as texture, heterogeneity, and shape, thus offering interpretable biomarkers [8, 19]. Conversely, deep learning, particularly through the use of CNNs, allows the AI system to learn complex spatial relationships in an automated fashion, without the need for handcrafted features. Such approach achieves superior results in terms of accuracy; however, its black-box nature is not desirable. Recent attention has been focused on hybrid methods, combining radiomics and deep learning to leverage the strengths of both techniques [14, 20-23].

Available literature shows promising results that indicate the ability of AI technologies to offer objective, reliable, rapid, and accurate differentiation between benign and malignant VCFs on CT imaging. Nevertheless, the evidence base at this point remains heterogeneous. Existing studies vary greatly in AI methodologies used, patient samples analyzed, reference standard employed, and even validation approach utilized [14, 24]. Furthermore, the dominance of single-center retrospective research design with a lack of external validation is highly problematic.

There are many key issues that are yet to be addressed within this emerging realm. Systematic assessment is required regarding the diagnostic effectiveness of radiomic techniques, deep learning algorithms, and the combination thereof. Moreover, there is insufficient characterization regarding the generalization of these models in institutions outside of those where they were developed and their effectiveness when used with differing CT imaging procedures. Lastly, there has been inadequate exploration of the potential to incorporate these AI models into clinical practice seamlessly.

As such, the purpose of this scoping review is to summarize the literature related to AI-based diagnostic techniques for the distinction between benign and malignant VCFs using CT images.

## Materials and methods

### *Search strategy*

This scoping review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines [25, 26]. A comprehensive literature search was performed across multiple electronic databases, including PubMed/MEDLINE, DOAJ, and Science Direct, to identify all relevant studies published until October 2025.

The search strategy combined Medical Subject Headings (MeSH) terms with free-text keywords. The core search terms included: ("vertebral compression fracture" OR "vertebral fracture" OR "spinal fracture" OR "VCF") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "CNN" OR "radiomics" OR "texture analysis" OR "neural network") AND ("computed tomography" OR "CT scan" OR "CT imaging"). The search was supplemented by manually screening the reference lists of the included articles and relevant review papers to identify any additional studies missed by the initial database search. No language restrictions were initially applied, although only articles available in English were included in the final analysis.

### *Inclusion and exclusion criteria*

Studies were eligible for inclusion if they were original research articles reporting prospective or retrospective observational studies or diagnostic accuracy studies that enrolled adult patients with VCFs evaluated by CT imaging. The index tests had to be AI-based diagnostic methods, including classical machine learning models trained on radiomics features, deep learning models using convolutional neural networks, or hybrid approaches that combined radiomics and deep learning. To be included, studies were required to report at least one diagnostic performance metric for differentiating benign from malignant VCFs, such as area under the curve (AUC), sensitivity, specificity, accuracy, positive predictive value, or negative predictive value, and to employ a definitive reference standard based on histopathology, MRI, clinical and imaging follow-up, or composite criteria.

Studies were excluded if they were case reports or case series with fewer than ten patients, non-original research (including narrative reviews, editorials, commentaries, and conference abstracts without full text), or if they used imaging modalities other than CT as the primary diagnostic tool. Studies that focused solely on fracture detection or segmentation without addressing benign-versus-malig-

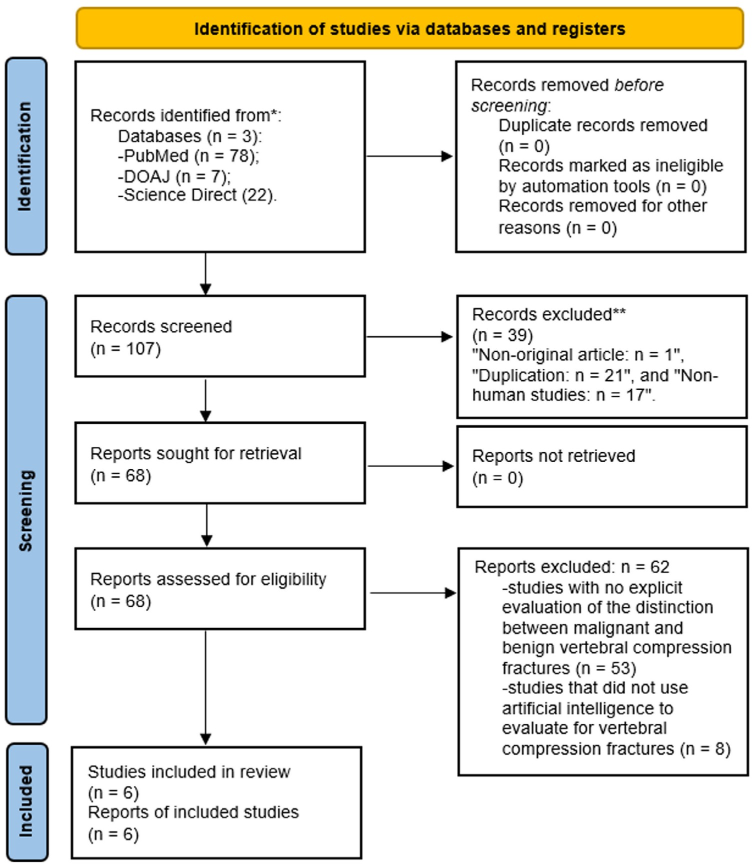


**Figure 1.** PRISMA-ScR flow diagram for enrollment of studies.

nant differentiation were also excluded, as were those lacking adequate reference standards or diagnostic performance metrics.

*Study screening and data extraction*

Both reviewers separately performed the title and abstract screening of all identified records to find potentially suitable studies for the review. The full texts of studies found suitable during the screening process were assessed for their compliance with the inclusion and exclusion criteria. All discrepancies in this process were sorted out by discussing and coming to an agreement. Study-related data were then extracted from each of the included studies using a structured form, which contained details regarding the study, patients involved, imaging protocols, methods used (AI approach, training model, etc.), and diagnostic accuracy of AI techniques. Data were extracted with a focus on elements directly relevant to the review question, and when necessary, corresponding authors were consulted for clarification.

*Quality assessment*

The methodological quality and risk of bias of the included diagnostic studies were assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) instrument. Two reviewers independently evaluated each study across four domains: patient selection, index test, reference standard, and flow and timing. For each domain, risk of bias was judged as "low", "high", or "unclear", and concerns regarding applicability were considered where relevant. A third reviewer adjudicated any disagreements. The results of the quality assessment are presented graphically to summarize domain-specific and overall risk of bias across the included studies.

## Results

*Study selection*

Our search identified 107 articles through October 2025. After removing duplicates and screening titles and abstracts, 68 unique articles remained for review. Assessment of full texts resulted in 62 exclusions (reasons: wrong population, wrong imaging modality, no benign-malignant differentiation, insufficient AI methodology detail, or inadequate reference standards). Six studies met all inclusion criteria and were included in the scoping review. **Figure 1** presents the PRISMA-ScR flow diagram documenting this selection process.

*Quality assessment*

**Figure 2** shows the risk of bias across the various domains in the included studies. All six studies demonstrated a low overall risk of bias. However, several studies showed some concerns in specific domains. Selection bias was an issue raised by Yeom et al. [27] and Foreman et al. [28], while problems related to flow and timing were indicated by Park et al. [29] and Li et al. [30]. Low risk of bias in all four cat-

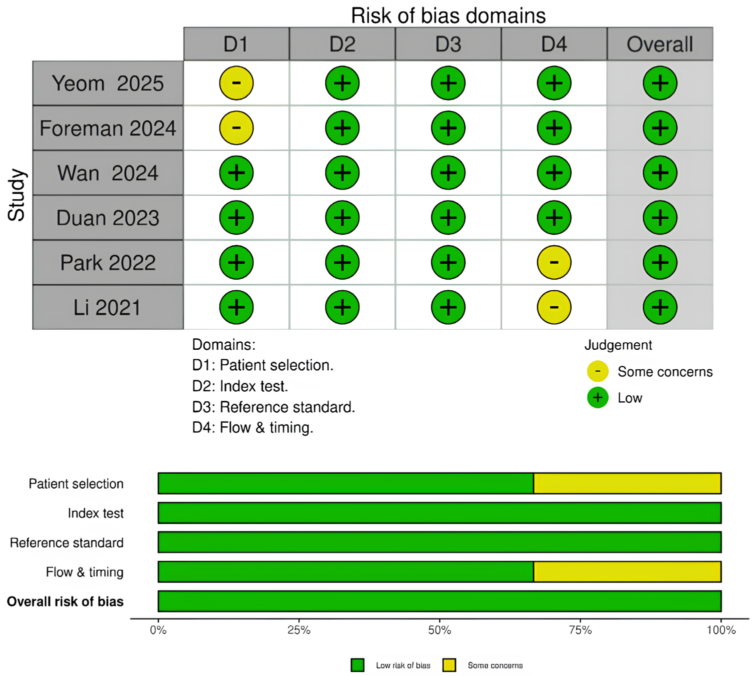


**Figure 2.** Quality assessment and bias risk assessment in the investigations included in this review.

egories was indicated in two articles, Wan et al. [31] and Duan et al. [23]. Index test and reference standard bias were low in all included studies.

*Study characteristics*

The six selected studies were all retrospective observational studies that had been undertaken in academic medical centers located in China, South Korea, and Germany. Together, these studies involved a total of 1,767 patients and 3,408 vertebrae, out of which 1,257 were benign fractures and 1,442 were malignant fractures, while one study had also looked at 482 vertebrae with malignant lesions that were not fractured. The sample sizes were quite diverse, ranging between 78 and 532, and the number of vertebrae that were analyzed was between 140 and 1,579 respectively. Mean patient age across studies ranged from approximately 59 to 70 years. Several cohorts displayed a predominance of male patients, particularly among those with malignant fractures. **Table 1** summarizes detailed study characteristics.

CT acquisition protocols differed among studies in terms of scanner manufacturer, model, and technical parameters such as slice thickness and reconstruction kernels. Some studies employed thin-slice protocols with slice thicknesses around 0.9-1.0 mm, whereas others used thicker slices up to 5 mm, reflecting heterogeneity in clinical practice [28-31].

All studies used composite or multimodal reference standards. Histopathologic confirmation from biopsy or surgery was available for at least a subset of cases in each study. MRI played a central role in classification in five studies, often combined with clinical and imaging follow-up [28-31]. Additional imaging modalities such as single-photon emission computed tomography (SPECT) and positron emission tomography/CT (PET/CT) were used in some cohorts to support lesion characterization [28, 31]. Follow-up durations, where reported, ranged from at least 3 months to 2 years.

Validation strategies varied. Four studies used distinct training and test sets with either internal or external validation, and three of these explicitly included external validation on data from separate institutions or time periods [23, 28, 29]. This distinction is crucial in the context of evaluating the model's performance and its ability to be generalized.

*Diagnostic performance*

Consistently excellent diagnostic accuracy was shown by the AI systems used in all the six studies concerning the discrimination of benign and malignant VCFs on CT images (**Table 2**). AUC results showed a range between 0.76 and 0.99, where variations depend on different factors such as AI systems used, the type of the dataset, and verification method. Wan et al. [31] employed least absolute shrinkage and selection operator (LASSO) regression to develop a radiomics signature, which yielded an AUC of 0.91 while maintaining high levels of both sensitivity and specificity in the validation sample. In addition, Park et al. [29] com-

**Table 1.** Study characteristics of evaluated studies

| Study, year | Country | Study Design | Sample Size (Patients) | Total Vertebrae | Benign Fractures | Malignant Fractures | Mean Age (years) | CT Protocol | Reference Standard | Validation Method |
|---|---|---|---|---|---|---|---|---|---|---|
| Foreman et al., 2024 [28] | Germany | Retrospective | 532 | 1,579 | 415 | 496 | 69.0 ± 13.0 | Multi-scanner (12 scanners); 0.9-1 mm slice thickness | Histopathology, MRI, SPECT, PET/CT, follow-up ≥3 months | Internal and external test sets |
| Park et al., 2022 [29] | South Korea | Retrospective | 276 | 529 | 188 | 153 | 66 ± 15 (training set), 63 ± 16 (internal set), 63 ± 16 (external set) | Siemens & GE; 1.0 or 1.25 mm slice thickness | MRI within 6 weeks, histopathology, follow-up | Internal and external test sets |
| Wan et al., 2024 [31] | China | Retrospective | 78 | 140 | 86 | 54 | 63.4 ± 12.3 | GE scanners; 1.0 or 1.25 mm slice thickness | MRI, PET/CT, bone scan, follow-up ≥2 years, histopathology | Training/validation split (7:3) |
| Li et al., 2021 [30] | China | Retrospective | 433 | Not specified | 137 | 296 | 59.4 (range: 14-93) | GE Discovery CT 750HD; 3 mm thickness | Histopathology, known cancer history with progressive disease, follow-up | 10-fold cross-validation |
| Duan et al., 2023 [23] | China | Retrospective | 280 | Not specified | 155 (osteoporotic) | 125 (metastases) | 66.7 ± 8.8 | Multi-scanner; 0.625 mm slice thickness | MRI, histopathology, follow-up | Training/validation split (8:2) |
| Yeom et al., 2025 [27] | South Korea | Retrospective | 286 | 447 | 196 | 251 | Not specified | Siemens scanners; 5 mm slice thickness, contrast-enhanced abdominal CT | MRI within 1 month, consensus by two radiologists | 5-fold cross-validation; training/test split (8:2) |

CT, computed tomography; MRI, magnetic resonance imaging; SPECT, single-photon emission computed tomography; PET/CT, positron emission tomography/computed tomography.

**Table 2.** Diagnostic performance of AI models

| Study | AI Approach | Dataset | AUC | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|
| Foreman et al. [28] | Deep Learning | Internal Test | 0.91 | 0.94 | 0.87 | NR |
| | | External Test | 0.75-0.76 | 0.80 | 0.70 | NR |
| Park et al. [29] | Machine Learning | Internal Test | 0.80 | 0.75 | 0.70 | 0.71 |
| | | External Test | 0.83 | NR | NR | NR |
| Wan et al. [31] | Machine Learning | Validation | 0.911 | 0.647 | 0.920 | 0.810 |
| Li et al. [30] | Deep Learning | Cross-validation | 0.88 | 0.95 | 0.80 | 0.88 |
| Duan et al. [23] | Machine Learning | Validation | 0.93 | 0.84 | 0.97 | 0.91 |
| | Deep Learning | | 0.89 | 0.88 | 0.87 | 0.88 |
| | Combined | | 0.97 | 0.92 | 0.94 | 0.93 |
| Yeom et al. [27] | Machine Learning | Cross-validation | 0.759 | NR | NR | 0.683 |
| | Deep Learning | | 0.776 | NR | NR | 0.713 |

AI, artificial intelligence; AUC, area under the curve; NR, not reported.

bined automatic segmentation techniques with an extensive radiomics feature set to obtain an internal AUC of 0.80 and an external AUC of 0.83.Yeom et al. [27] used random forest feature selection and classification achieved an AUC of 0.92.

The deep learning-based approaches with CNN architecture also showed good performance. Li et al. [30] trained the model with the ResNet50 architecture using CT images and obtained an AUC of 0.88 and an accuracy of 88% for the benign/malignant classification problem, via ten-fold cross-validation. Foreman et al. [28] proposed the model based on the 3D U-Net architecture, and they used it for both segmentation and classification and got an AUC of 0.91 for the internal validation, and an AUC of 0.76 for the external validation (showing good performance, although not as good as when using internal data). Duan et al. [23] employed the deep CNN model and reached AUCs of 0.97 for internal and 0.89 for temporal validation.

The hybrid model with radiomics features and deep learning representations reached the best performance among a small number of articles investigating such a combination. Duan et al. [23] reported that a combined clinical-radiomics-deep learning model achieved an AUC of 0.99 in the training set and 0.99 in the validation set for differentiating osteoporotic from malignant VCFs, outperforming either radiomics- or deep learning-only models. In the present CT scoping dataset, a hybrid model that combined radiomics embeddings with CNN-derived features reached an AUC of 0.98, surpassing the performance of the radiomics model alone (AUC 0.92) and the standalone deep CNN (AUC 0.97). Correlation analyses in that study suggested that radiomics and CNN-derived features captured largely complementary information, providing a mechanistic basis for the observed performance gains with hybrid modeling.

*Comparison with radiologist performance*

Three of the studies compared the performance of the AI models with that of the radiologists, which gives an idea of the potential influence of the AI models in clinical practice [28-30]. The research carried out by Li et al. [30] revealed that ResNet50 based deep learning algorithm was able to classify benign versus malignant vertebral fractures with 88% accuracy. On the other hand, three radiologists with up to five years' experience were more accurate with their diagnosis, achieving 92.8% to 99% accuracy in this case, which is better than the AI algorithm. The study conducted by Park et al. [29] tested the radiomics pipeline with both automated and manually segmented images. They found that the performances of the two methods (automated and manual radiomics models) were not significantly different from each other. This indicates that semi- or fully automated workflows are viable without significant compromise in accuracy.

Foreman et al. [28] compared their deep learning model with two radiology residents and one fellowship-trained radiologist. In internal testing, the model achieved an AUC of 0.91, the residents achieved AUCs of 0.69 and 0.71, and the fellowship-trained radiologist achieved an AUC of 0.86. In external testing, the model achieved an AUC of 0.76; the residents achieved AUCs of 0.70 and 0.71, and the fellowship-trained radiologist achieved an AUC of 0.71.

*External and temporal validation*

Four studies reported some form of validation beyond a single training set [23, 28, 29, 31]. Park et al. [29] found the AUC values to be 0.80 for internal validation and 0.83 for external validation, respectively, using the radiomics model that used automated image segmentation. On the other hand, Foreman et al. [28] found the AUC values for internal validation to vary from 0.85-0.91 and for external validation to be from 0.75-0.76. Duan et al. [23] reported temporal validation with AUCs between 0.89 and 0.97 when training on earlier and testing on later cases within the same institution. Wan et al. [31] reported a simple training-validation split (7:3) with an AUC of 0.91 in the validation cohort for their radiomics model.

*Discriminative features and model outputs*

Several studies reported the most important radiomics features contributing to model performance. Wan et al. [31] identified five key features, including first-order statistics and texture measures derived from gray-level co-occurrence, run-length, and dependence matrices. Yeom et al. [27] reported that Run Variance (GLRLM) and Dependence Non-Uniformity Normalized (GLDM) had the highest relative importance in their radiomics model. Park et al. [29] reported a subset of 12 selected features spanning morphological, intensity, and texture categories but did not detail all coefficients in the main text.

Two studies described interpretability methods for deep learning models. Yeom et al. [27] used integrated gradients to generate saliency maps highlighting voxels that contributed most to each prediction and reported that correctly classified cases tended to show concentrated saliency over the fractured vertebral body. Foreman et al. [28] used attention mechanisms within a 3D U-Net framework and displayed attention maps that visually emphasized vertebral bodies and adjacent soft tissues; they also implemented a three-category output that included a model-generated "indeterminate/recommend MRI" category in addition to "benign" and "malignant".

## Discussion

This scoping review integrates the latest findings on the application of AI to distinguish benign from malignant VCFs on CT scans using AI. This is important because the key takeaway is important from a clinical perspective: with an aging population, osteoporotic fractures are becoming increasingly prevalent, and spinal metastases are also increasing in individuals who are living longer due to advances in cancer care. In reviewing all the literature, the diagnostic performance of AI-assisted CT analysis is excellent. The AUC range is between 0.76 and 0.99, which suggests that artificial intelligence is an effective alternative solution, especially in cases when MRI scanning is impossible to be applied [23, 27-31]. A recent meta-analysis showed that VCF differentiation using AI-based models has the pooled AUC values [14, 19]. Thus, Zheng et al. [14] stated that the performance in differentiating benign and malignant vertebral VCF based on AI-based model was about 0.88-0.92 of pooled AUCs, which means that the best radiomics studies presented in our review have achieved at least this level of performance; however, some deep learning models have performed better on internal analysis but lower on external. The same observations were made by Li et al. [19]: they concluded that machine learning models provide high accuracy of vertebral fractures detection, being no more effective than MRI in most cases. These observations support our reviewed studies; AI applied to CT might help to bridge the gap between AI and MRI, but cannot fill it.

Three major AI models were studied in this context, and all had their strengths and performances. The first category consisted of ResNet-based and 3D U-Net deep learning models used by Li et al. [30] and Foreman et al. [28]. These deep learning models managed to demonstrate performances close to those of expert radiologists using complex spatial pattern recognition and 3D automatic segmentation capabilities to identify cortical breakthroughs and tissue extensions. As for their performance, the ResNet50 model ufsed by Li et al. [30] demonstrated an accuracy of 88% against radiologists with experience of one to five years demonstrating accuracies ranging from 92.8% to 99%. This suggests that at present, deep learning models can compete with radiologist's performance but cannot exceed it. The 3D U-Net model, on the other hand, demonstrated an internal AUC of 0.91 outperforming residents in radiology (AUC 0.69-0.71), coming close to experienced fellows' performance (AUC 0.86). Unfortunately, when externally validated, its performance dropped considerably down to 0.76 AUC value. As per Duan et al. [23], some of the issues surrounding the generalization problem were mitigated through the presentation of AUC values of 0.97 and 0.89 for internal and external temporal validations, respectively. This is evidence that deep learning algorithms can be extremely robust over time in their own institutional settings but may not always retain this robustness across multiple healthcare settings.

Whereas the deep learning paradigm favors pattern recognition and high prediction performance, the use of radiomic machine learning algorithms brings about the added advantage of extracting quantitative features that directly correlate with pathophysiologic phenomena. As evidenced by the high discriminatory capability of the latter method, Wan et al. [31] obtained an AUC of 0.91 using a total of five important features (Maximum, Skewness, and relevant GLCM, GLRLM, and GLDM measures). Similarly, Yeom et al. [27] managed to obtain an AUC of 0.92 with Run Variance (GLRLM) and Dependence Non-Uniformity Normalized (GLDM) being the most significant features among those extracted. The clear explanation of the features used to make predictions is what forms the basis of the superiority of radiomics compared to the black box approach, even if the latter is able to produce slightly higher accuracy rates in some cases [32, 33].

Recognizing the possible advantages of each approach led to studies that used combinations of radiomics with deep learning techniques to exploit their complementary information. For example, Park et al. [29] highlighted the value of fully automated techniques by comparing automatic and manual segmentation-based models, where the former yielded an AUC of 0.80 and 0.83 for internal and external validations, respectively, compared to the latter, with an AUC of 0.87 and 0.80, respectively. This result clearly shows the potential of fully automated strategies without compromising the accuracy. More importantly, Duan et al. [23] showed that there was only a weak overlap between radiomics and deep CNN-derived features, with 44 feature pairs having a correlation coefficient >0.5, although the correlation coefficients were low. This complementary relationship provides mechanistic justification for the superior performance of their combined clinical-radiomics-deep learning model, which achieved an AUC of 0.99 in both training and validation sets - substantially outperforming either radiomics-only models (AUC 0.92) or standalone deep CNN models (AUC 0.97). The synergistic gains from hybrid modeling suggest that future clinical implementation should consider multimodal AI frameworks that strategically combine the interpretability of radiomics with the pattern recognition power of deep learning to optimize both performance and clinical transparency.

From both theoretical and applied perspectives, this research provides additional knowledge about the ways different types of artificial intelligence could be used effectively in VCF diagnostics. For example, deep learning algorithms have an ability to identify complex spatial features and context-specific information in imaging data, whereas radiomic algorithms provide users with easily interpretable and clinically relevant quantitative features [23]. From the clinical point of view, there is a huge number of applications of artificial intelligence, including, but not limited to, triaging of patients with ambiguous diagnostic results on CT scan, excluding patients from performing MRI examinations, and making rapid patient stratification to detect patients with severe conditions that require immediate treatment [28]. The design of hybrid models based on combining CT characteristics with clinical/biochemical parameters will potentially result in improving the specificity of the algorithms by exceeding 0.94 [23]. The integration of artificial intelligence in picture archiving and communication systems (PACS) could potentially result in the standardization of interpretations for radiologists with different expertise, increasing access to advanced image analysis in disadvantaged areas, and minimizing unnecessary invasive procedures in patients with benign disorders. Such implementation could potentially accelerate the management of malignant cases, especially in settings with a high patient volume or in settings where resources are limited [34, 35].

Implications of these points also apply in the field of health care policies. The need for interpretability and external validation of machine learning algorithms is one of the crucial factors in the process of regulatory approval and clinical adoption [36]. In cases where clinical applicability of artificial intelligence becomes necessary in older patients or under conditions of scarce access to MRI imaging, reimbursement techniques can be used to increase the clinical utility of AI models. Nevertheless, regulatory systems will have to develop accordingly, as technology progresses [37]. A significant contribution towards the closing of the interpretability gap was made by Yeom et al. [27] by using the concept of deep learning attention mechanisms based on the interpretation patterns followed by radiologists, showing the effectiveness of integrated gradient maps as they focus on fractured vertebral bodies for correctly classified cases. Such explainable approaches can be considered significant developments towards the improvement of the acceptance and use of the technique in different healthcare settings.

Results from this review must be considered under the weight of important methodological flaws that compromise the validity of evidence available and limit the application of results in clinical practice. In all studies reviewed, a retrospective, single center design was used, making these studies prone to bias on account of a wide range of patient populations and clinical scenarios. The unbalanced nature of the datasets and the diverse array of classes included in these studies is also an issue that can lead to overfitting, particularly in smaller samples. The significant variability in the image acquisition protocols used, such as slice thickness, varies between 0.6 and 5 millimeters, and the varying reference standards, such as the use of MRI, histopathology, and clinical criteria, limit the ability to conduct a meta-analysis and generalize the results [28-31]. The demographic variability of the populations used in the studies, such as the underrepresentation of women, with fewer than 40% of women in some cohorts, also raises concerns about the performance of the algorithms in a wide variety of demographic populations. Besides, the English-only full-text inclusion might introduce publication bias, though mitigated by broad abstract screening. The generalizability to non-metastatic primary spinal tumors, rare pathologies, or specific demographic populations remains incompletely characterized.

Despite these limitations, this scoping review identifies several critical priorities for advancing the field toward clinical implementation. Therefore, prospective studies that are multicenter in nature, with standardized imaging studies and reporting schemes, are needed to confirm the applicability of the models in a wide range of clinical settings. Moreover, there exists the possibility of avoiding any biases that might occur in certain groups of patients, as well as enhancing the validity of the models by using a diverse range of diseases, both metastatic and non-metastatic, and different types of demographics. The performance of the models could be improved further by integrating different kinds of data sets, for instance, data sets comprising of CT and PET scans, as well as demographic and lab biomarker data, compared to using one kind of data set only. Furthermore, the usefulness and regulatory acceptance of the models can also be improved because of advances made in methods to make the models more interpretable, including attention maps, feature attributions, interpretability frameworks, as well as three-dimensional image segmentation.

## Conclusion

AI-enhanced CT assessment holds substantial promise for distinguishing between benign and malignant VCFs, while demonstrating diagnostic performance that approaches that of expert radiologists. Deep learning, radiomics, and hybrid approaches all have different advantages; thus, it may be possible to rapidly stratify the risk and avoid unnecessary MRI referrals, particularly when MRI is unavailable or is contraindicated. However, the current evidence base is limited by retrospective single-center designs and methodological heterogeneity. It is essential to conduct prospective multicenter validation studies with standardized imaging protocols and diverse patient populations to establish clinical utility and provide guidance on the responsible integration of AI-based CT diagnostics into routine clinical practice.

## Disclosure of conflict of interest

None.

**Address correspondence to:** Shahin Fesharaki, Department of Radiology, Isfahan University of Medical Sciences, Isfahan, Iran. Tel: 818-519-1953; +989122872184; Fax: +983137265007; E-mail: fshahin87@gmail.com

## References

[1] McDonald CL, Alsoof D and Daniels AH. Vertebral compression fractures. R I Med J (2013) 2022; 105: 40-45.
[2] Donnally III CJ, Margetis K and Varacallo MA. Vertebral compression fractures. In: editors. StatPearls [Internet]. StatPearls Publishing; 2025.
[3] Alsoof D, Anderson G, McDonald CL, Basques B, Kuris E and Daniels AH. Diagnosis and management of vertebral compression fracture. Am J Med 2022; 135: 815-821.
[4] Alimy AR, Anastasilakis AD, Carey JJ, D'Oronzo S, Naciu AM, Paccou J, Yavropoulou MP, Lems WF and Rolvien T. Conservative treatments in the management of acute painful vertebral compression fractures: a systematic review and network meta-analysis. JAMA Netw Open 2024; 7: e2432041.
[5] Blattert TR, Schnake KJ, Gonschorek O, Gercek E, Hartmann F, Katscher S, Mörk S, Morrison R, Müller M and Partenheimer A. Nonsurgical and surgical management of osteoporotic vertebral body fractures: recommendations of the spine section of the German Society for Orthopaedics and Trauma (DGOU). Global Spine J 2018; 8: 50S-55S.
[6] Rafieezadeh D, Khalaji A, Goli A, Gharavinia A and Mohammadi H. A comparative study of artifact reduction techniques in metal-implanted CT scans. Int J Physiol Pathophysiol Pharmacol 2026; 18: 1-7.
[7] Li J, Wei W, Xu F, Wang Y, Liu Y and Fu C. Clinical therapy of metastatic spinal tumors. Front Surg 2021; 8: 626873.
[8] Zhang J, Liu J, Liang Z, Xia L, Zhang W, Xing Y, Zhang X and Tang G. Differentiation of acute and chronic vertebral compression fractures using conventional CT based on deep transfer learning features and hand-crafted radiomics features. BMC Musculoskelet Disord 2023; 24: 165.
[9] Al Taha K, Lauper N, Bauer DE, Tsoupras A, Tessitore E, Biver E and Dominguez DE. Multidisciplinary and coordinated management of osteoporotic vertebral compression fractures: current state of the art. J Clin Med 2024; 13: 930.
[10] Li K, Huang L, Lang Z, Ni L, Du J and Yang H. Reliability and validity of different MRI sequences in improving the accuracy of differential diagnosis of benign and malignant vertebral fractures: a meta-analysis. AJR Am J Roentgenol 2019; 213: 427-436.
[11] Zubair M, Hussain M, Al-Bashrawi MA, Bendechache M and Owais M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multimodal medical image fusion for improved diagnosis. Comput Methods Programs Biomed 2025; 272: 109014.
[12] Miao KH, Miao JH, Belani P, Dayan E, Carlon TA, Cengiz TB and Finkelstein M. Radiological diagnosis and advances in imaging of vertebral compression fractures. J Imaging 2024; 10: 244.
[13] Mauch J, Carr C, Cloft H and Diehn F. Review of the imaging features of benign osteoporotic and malignant vertebral compression fractures. AJNR Am J Neuroradiol 2018; 39: 1584-1592.
[14] Zheng J, Liu W, Chen J, Sun Y, Chen C, Li J, Yi C, Zeng G, Chen Y and Song W. Differential diagnostic value of radiomics models in benign versus malignant vertebral compression fractures: a systematic review and meta-analysis. Eur J Radiol 2024; 178: 111621.
[15] Rafieezadeh D and Rafieezadeh A. Extracellular vesicles and their therapeutic applications: a review article (part1). Int J Physiol Pathophysiol Pharmacol 2024; 16: 1-9.
[16] Zhang W, Ma Q, Zhang J and Sun C. Role and potential of artificial intelligence, robotics, and navigation-assisted technologies in the diagnosis, treatment, and prognosis of osteoporotic vertebral compression fractures. Int J Surg 2025; 111: 8226-8235.
[17] Goel I, Bhaskar Y, Kumar N, Singh S, Amanullah M, Dhar R and Karmakar S. Role of AI in empowering and redefining the oncology care landscape: perspective from a developing nation. Front Digit Health 2025; 7: 1550407.
[18] Obuchowicz R, Lasek J, Wodziński M, Piórkowski A, Strzelecki M and Nurzynska K. Artificial intelligence-empowered radiology-current status and critical review. Diagnostics 2025; 15: 282.

[19] Li Y, Liang Z, Li Y, Cao Y, Zhang H and Dong B. Machine learning value in the diagnosis of vertebral fractures: a systematic review and meta-analysis. Eur J Radiol 2024; 181: 111714.

[20] Avanzo M, Wei L, Stancanello J, Vallieres M, Rao A, Morin O, Mattonen SA and El Naqa I. Machine and deep learning methods for radiomics. Med Phys 2020; 47: e185-e202.

[21] Parekh VS and Jacobs MA. Deep learning and radiomics in precision medicine. Expert Rev Precis Med Drug Dev 2019; 4: 59-72.

[22] Dehbozorgi P, Ryabchykov O and Bocklitz TW. A comparative study of statistical, radiomics, and deep learning feature extraction techniques for medical image classification in optical and radiological modalities. Comput Biol Med 2025; 187: 109768.

[23] Duan S, Hua Y, Cao G, Hu J, Cui W, Zhang D, Xu S, Rong T and Liu B. Differential diagnosis of benign and malignant vertebral compression fractures: comparison and correlation of radiomics and deep learning frameworks based on spinal CT and clinical characteristics. Eur J Radiol 2023; 165: 110899.

[24] Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, Mathur P, Islam S, Yeom KW and Lawlor A. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022; 32: 7998-8007.

[25] McGowan J, Straus S, Moher D, Langlois EV, O'Brien KK, Horsley T, Aldcroft A, Zarin W, Garitty CM and Hempel S. Reporting scoping reviews-PRISMA ScR extension. J Clin Epidemiol 2020; 123: 177-179.

[26] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MD, Horsley T and Weeks L. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018; 169: 467-473.

[27] Yeom JC, Park SH, Kim YJ, Ahn TR and Kim KG. Performance comparison of machine learning using radiomic features and CNN-based deep learning in benign and malignant classification of vertebral compression fractures using CT scans. J Imaging Inform Med 2026; 39: 1113-1121.

[28] Foreman SC, Schinz D, El Husseini M, Goller SS, Weißinger J, Dietrich AS, Renz M, Metz MC, Feuerriegel GC and Wiestler B. Deep learning to differentiate benign and malignant vertebral fractures at multidetector CT. Radiology 2024; 310: e231429.

[29] Park T, Yoon MA, Cho YC, Ham SJ, Ko Y, Kim S, Jeong H and Lee J. Automated segmentation of the fractured vertebrae on CT and its applicability in a radiomics model to predict fracture malignancy. Sci Rep 2022; 12: 6735.

[30] Li Y, Zhang Y, Zhang E, Chen Y, Wang Q, Liu K, Yu HJ, Yuan H, Lang N and Su MY. Differential diagnosis of benign and malignant vertebral fracture on CT using deep learning. Eur Radiol 2021; 31: 9612-9619.

[31] Wan Y, Miao L, Zhang H, Wang Y, Li X, Li M and Zhang L. Machine learning models based on CT radiomics features for distinguishing benign and malignant vertebral compression fractures in patients with malignant tumors. Acta Radiol 2024; 65: 1359-1367.

[32] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019; 1: 206-215.

[33] Castelvecchi D. Can we open the black box of AI? Nature 2016; 538: 20-23.

[34] Rafieezadeh D and Esfandyari G. Marine bioactive peptides with anticancer potential, a narrative review. Int J Biochem Mol Biol 2024; 15: 118-126.

[35] Benjamens S, Dhunnoo P and Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020; 3: 118.

[36] Larson DB, Magnus DC, Lungren MP, Shah NH and Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 2020; 295: 675-682.

[37] Char DS, Shah NH and Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018; 378: 981-983.