



Radiological Validation and Explainability of a Deep Learning Model for Vertebral Compression Fracture Detection in Geriatric Spine Imaging

© Anıl ÖZGÜR¹, © Ali Murat BAŞAK²

¹İnegöl State Hospital, Clinic of Orthopedics and Traumatology, Bursa, Türkiye

²University of Health Sciences Türkiye, Gülhane Training and Research Hospital, Clinic of Orthopedics and Traumatology, Ankara, Türkiye

Cite as: Özgür A, Başak AM. Radiological validation and explainability of a deep learning model for vertebral compression fracture detection in geriatric spine imaging. Forbes J Med 2026;7:122-122

ABSTRACT

Objective: Osteoporotic vertebral compression fractures are common in older adults and can significantly impair quality of life. Early and accurate detection of these fractures remains clinically challenging. This study aimed to develop and evaluate an explainable deep learning framework for the automated detection and classification of vertebral compression fractures on lumbar computed tomography (CT) images.

Methods: Eighty anonymized lumbar CT scans from the VerSe2020 and SpineWeb open-access datasets were analyzed. Vertebrae were segmented, and morphological features, including height loss and posterior cortical integrity, were extracted. A dual-stage deep learning pipeline combining YOLOv8 for fracture localization and EfficientNet-B0 for classification was developed to predict both the presence of fractures and their surgical eligibility. Model performance was evaluated using the following metrics: accuracy, F1 score, sensitivity, specificity, and receiver operating characteristic-area under the curve (ROC-AUC).

Results: The model achieved an accuracy of 68%, an F1 score of 63%, a sensitivity of 60%, a specificity of 65%, and an ROC-AUC of 0.66. Additionally, the model's decision-making focus on morphologically relevant zones suggests its potential use for automated screening and prioritization in radiological workflows.

Conclusion: This artificial intelligence-based framework may support early diagnosis and preoperative assessment of vertebral fractures in geriatric patients and has the potential to assist clinical decision-making.

Keywords: Vertebral fracture, artificial intelligence, radiological imaging, deep learning

Received: 17.07.2026

Accepted: 15.09.2026

Publication Date: 02.10.2026

Corresponding Author:

Anıl ÖZGÜR

İnegöl State Hospital, Clinic of
Orthopedics and Traumatology,
Bursa, Türkiye

✉ dr.anilozgur@gmail.com

ORCID ID: 0000-0002-5877-0628

INTRODUCTION

Osteoporotic vertebral compression fractures are common in elderly and postmenopausal patients and are associated with persistent pain, spinal deformity, and reduced quality of life.¹ Early radiological detection is particularly important in elderly patients, as delayed diagnosis may contribute to progressive vertebral collapse and functional deterioration. Percutaneous kyphoplasty is widely used to restore vertebral height and relieve pain in selected patients; however, its indications should be considered

carefully because of variable outcomes and procedure-related complications, including new fractures, cement leakage, and spinal canal compression.^{2,3}

Artificial intelligence (AI)-based approaches in orthopedic imaging have rapidly expanded in recent years, enabling automated fracture detection and classification using computed tomography (CT) and magnetic resonance imaging data.^{4,5} Beyond imaging analysis, large language model-based systems such as ChatGPT have also demonstrated increasing potential in orthopedic decision-



support processes and patient-oriented clinical guidance.⁶ In addition, digital health research has shown that online orthopedic information-seeking behavior demonstrates significant temporal and seasonal variation, highlighting the growing interaction between AI, digital epidemiology, and orthopedic practice.⁷ Deep learning models, including three-dimensional (3D) convolutional neural networks and Mask R-CNN, have shown high accuracy in vertebral segmentation and radiological interpretation.⁸ In addition, AI models have been applied to predict clinical outcomes such as fracture risk, cement leakage, and postoperative complications, supporting more personalized treatment strategies.⁹⁻¹¹

Despite these advances, few studies have combined automated vertebral fracture detection with kyphoplasty-related clinical assessment using AI.¹² Therefore, this study aimed to detect vertebral compression fractures in open-access CT datasets and to evaluate their clinical relevance using a deep learning-based approach.

METHODS

Study Design and Ethical Approval Status

This study is a non-interventional, AI-based analysis that uses anonymized, open-access CT datasets (VerSe 2020 and SpineWeb) to detect vertebral compression fractures and evaluate their clinical relevance. The primary objective was to develop a model capable of identifying fractures and supporting preoperative assessment. No direct patient data, clinical records, or identifiable information were used in the study. All imaging data were obtained from publicly available sources and processed solely for research purposes. The study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines. As the analysis was performed on fully de-identified datasets without involvement of human subjects, approval from the ethics committee was not required. The study design reflects a technical and observational approach aimed at supporting clinical decision-making.

Image Sources and Data Set Properties

In this study, open-access CT images from the VerSe 2020 and SpineWeb datasets were used to identify vertebral fractures and assess their clinical relevance. The VerSe dataset includes vertebral segmentation data from thoracolumbar vertebrae, while SpineWeb provides CT images of compression fractures across different age groups. A total of 80 patients (approximately 150-200 sagittal slices per case) were analyzed. Cases were included if sagittal CT images provided adequate visualization of the vertebral bodies and were accompanied by reference annotations that were suitable for model training and

evaluation. Cases with incomplete vertebral visualization, severe image artifacts, or insufficient annotation data were excluded. The 80 subjects, therefore, represented eligible cases meeting these imaging-based criteria rather than a random sample. Images were obtained in DICOM format and preprocessed, and ground-truth segmentation masks were derived from reference annotations. Morphological parameters, including vertebral body height, fracture severity (Genant classification), and posterior cortical integrity, were extracted for analysis.

Artificial Intelligence Model and Training Process

The proposed model was a dual-stage deep-learning framework for vertebral compression fracture detection and radiological validation. First, a YOLOv8-based module was used to localize and segment vertebral regions from sagittal CT images; standard preprocessing and data augmentation were applied to improve generalizability. In the second stage, an EfficientNet-B0 classifier classified the segmented vertebrae as fractured or intact. The model captured key morphological features such as vertebral height loss, cortical irregularity, and posterior wall disruption.

Model training was performed using the PyTorch framework. A total of 5,600 sagittal CT images, obtained from 80 subjects, were included in the image-level classification analysis and constituted the effective sample for model development and evaluation. The images were divided into training, validation, and test sets at a ratio of 70%, 15%, and 15%, respectively. Individual vertebrae were not treated as independent subject-level observations; therefore, the effective sample size was defined by the number of images included in the classification analysis rather than by a separate vertebra-level count. Model performance was evaluated using accuracy, sensitivity, specificity, F1-score, and ROC-AUC metrics. To enhance interpretability, Grad-CAM visualizations were generated and qualitatively assessed by two independent musculoskeletal radiologists for anatomical consistency.

Radiological Criteria for Fracture Severity

Fracture severity was defined using commonly accepted radiological criteria, including >40% anterior vertebral height loss, posterior wall disruption, segmental kyphotic angle >10°, and loss of posterior cortical continuity. These criteria were used to evaluate the model's ability to identify morphologically severe fractures rather than guide treatment decisions. To assess anatomical validity and inter-observer consistency, two independent musculoskeletal radiologists reviewed a 20% subset of the cases as a quality-control sample. This subset assessment was performed to verify the anatomical consistency of

the predefined radiological criteria, rather than to provide a second radiological evaluation of the entire dataset. Substantial inter-observer agreement ($\kappa=0.82$) was observed, supporting the consistency of the radiological assessment within the reviewed subset.

Evaluation and Performance Analysis

Model performance was evaluated on the test dataset using standard classification metrics, including accuracy, sensitivity, specificity, F1-score, and receiver operating characteristic-area under the curve (ROC-AUC). The ROC curve was used to assess the model's ability to discriminate between fractured and non-fractured vertebrae, and an optimal decision threshold was determined. In addition, fracture identification performance was evaluated using the true positive rate, false negative rate, and false positive rate. Segmentation performance was also evaluated using the mean Intersection over Union (mIoU).

RESULTS

The developed deep learning framework was successfully applied to the analysis of vertebral compression fractures using open-access CT datasets. A total of 5,600 sagittal slices obtained from 80 anonymized subjects were processed using the dual-stage model architecture. In the first stage, the YOLOv8-based localization network effectively segmented vertebral boundaries across the entire dataset, reaching a mIoU score of 0.91 ± 0.04 , which confirmed accurate regional detection. Regional detection accuracy was verified by comparing the predicted vertebral regions with reference vertebral segmentation annotations from open-access datasets, which served as the ground truth for IoU calculation. Only a small proportion of slices (1.6%) were excluded due to incomplete vertebral visualization or motion-related artifacts.

Following segmentation, the EfficientNet-B0 classifier was trained to differentiate normal from fractured vertebrae, based on predefined morphological criteria. The model achieved an accuracy of 68% and an F1 score of 0.63, indicating balanced classification performance. Sensitivity and specificity were 60% and 65%, respectively, demonstrating a moderate ability to identify fractured and intact vertebrae. ROC analysis yielded an AUC of 0.66, reflecting moderate discriminatory performance (Table 1).

Based on radiological severity criteria, the model correctly classified 49% of vertebrae with severe morphological features. False-negative and false-positive rates were 22% and 17%, respectively; these were mainly due to subtle fractures and degenerative changes that mimicked compression. Overall agreement between AI predictions and expert radiological assessments was 52%, indicating moderate consistency (Table 2).

To further understand model performance, an error distribution analysis was conducted. The most frequent limitation was the omission of marginal (mild) fractures, accounting for 28% of total errors. Ambiguity in vertebral border segmentation constituted 19%, while undetected posterior wall damage accounted for 14%. Image quality variation among the open datasets accounted for 21% of the observed performance inconsistencies, while model overfitting accounted for 11%.

The ROC curve illustrates the model's discriminative ability to distinguish fractured from non-fractured vertebrae (Figure 1). The curve's deviation from the ideal diagonal indicates that the model has limited but significant discriminative ability, suggesting that performance could be further improved with larger and more diverse datasets. The optimal threshold for binary classification was determined to be 0.53, balancing sensitivity and specificity in future applications.

DISCUSSION

In the present study, the proposed deep learning framework demonstrated moderate diagnostic performance in detecting vertebral compression fractures on open-access CT images. Given the clinical importance of timely fracture recognition, these findings suggest a potential supportive role for AI-assisted radiological assessment and emphasize the limitations of the current model. Importantly, the framework was designed to identify radiologically relevant fracture features rather than to determine treatment indications, which remain dependent on comprehensive clinical and radiological evaluation.^{1,12}

AI-based approaches for vertebral fracture detection have gained increasing attention in recent years. Deep learning models trained on CT data have demonstrated high diagnostic performance for fracture localization and classification under controlled conditions.^{13,14} Some studies have reported ROC-AUC values exceeding 0.90 using 3D convolutional architectures and institution-specific datasets, indicating that automated systems can achieve expert-level accuracy in optimal settings.^{14,15} In comparison, the dual-stage deep learning framework used in the present study, combining YOLOv8 for vertebral localization and EfficientNet-B0 for classification, achieved an ROC-AUC of 0.66, indicating moderate discriminative performance and lower diagnostic accuracy than the >0.90 values reported under more controlled conditions. This finding is more consistent with the moderate performance reported for models based on heterogeneous open-access datasets. In contrast to these high-performing models, approaches relying exclusively on open-access datasets generally report more moderate results due to heterogeneity in image quality, acquisition protocols, and annotation

Metric	Definition	Formula	Model result (%)	Interpretation
Accuracy	Model accuracy shows the proportion of samples correctly classified among all predictions	$(TP + TN)/(TP + TN + FP + FN)$	68	Overall accuracy is at an average level
F1 score	Balance between precision and sensitivity	$2 \times (\text{Precision} \times \text{sensitivity}) / (\text{precision} + \text{sensitivity})$	63	Balance between sensitivity and scope achieved
Sensitivity	The proportion of true positives	$TP/(TP + FN)$	60	The model captures positives at a rate of 60%
Specificity	The proportion of true negatives	$TN/(TN + FP)$	65	Moderate ability to differentiate negatives
ROC-AUC	Ability to differentiate true positive/negative	AUC	66	The model's classification capacity is limited but significant

TP: True positive, TN: True negative, FP: False positive, FN: False negative, ROC-AUC: Receiver operating characteristic-area under the curve

Criterion	Definition	Percentage (%)	Clinical significance	Interpretation
True positive	Radiologically severe fractures correctly identified by the model	49	It has the potential to contribute to treatment	The model correctly identified nearly half of kyphoplasty candidates
False negative	Radiologically severe fractures not identified by the model	22	They may miss the opportunity for treatment.	Lack of sensitivity is a serious limitation.
False positive	Cases incorrectly classified by the model as radiologically severe fractures	17	Increases the risk of unnecessary operations	An important source of error affecting the specificity
Overall agreement	Agreement between model predictions and expert radiological assessment	52	Indicates overall reliability	The decision-making consistency of the model is moderate
Overclassification rate	Cases incorrectly classified as radiologically severe by the model	17	Clinical safety should be questioned	A cautious approach should be taken to clinical decision support

standards.^{4,5} Studies that combine image-based analysis with quantitative morphological features, such as vertebral height loss ratios, have shown improved sensitivity and specificity compared with image-only models.^{15,16} These findings suggest that incorporating structured radiological parameters can substantially enhance fracture detection performance. Importantly, the model presented in this study focused on anatomically relevant regions, specifically cortical disruption and vertebral height loss, thereby further supporting its radiological validity.

Beyond fracture detection, machine learning methods have been increasingly applied to predict clinical outcomes related to vertebral augmentation procedures, including postoperative pain, cement leakage, and fracture recurrence.^{11,17,18} While these predictive models typically incorporate demographic and clinical variables in addition to imaging data, they highlight the broader potential of AI to support orthopedic decision-making across multiple stages of patient management.

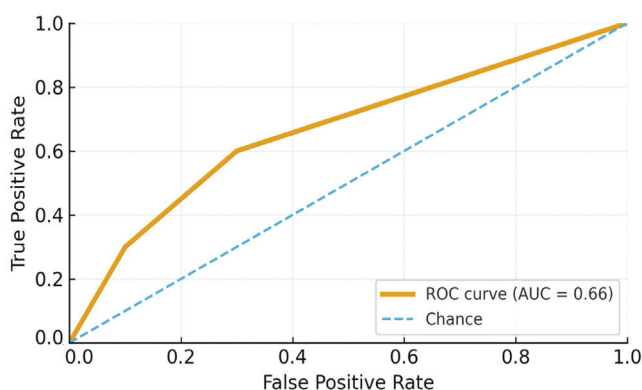


Figure 1. Receiver operating characteristic curve illustrating the model's ability to discriminate between fractured and non-fractured vertebrae (AUC = 0.66)

ROC: Receiver operating characteristic, AUC: Area under the curve

Model interpretability is essential for clinical adoption of AI systems. Visualization techniques, such as Grad-CAM, improve transparency by highlighting regions that influence model decisions.^{19,20} In this study, Grad-CAM analysis demonstrated that the model focused on morphologically relevant vertebral regions, supporting the reliability of its predictions.

Despite its potential, the present study has several limitations. The model was trained using a limited number of open-access imaging datasets, and heterogeneity in image quality, acquisition protocols, and annotation standards may restrict its generalizability across different patient populations. In addition, the absence of demographic and clinical variables, such as age, sex, and osteoporosis severity, limits the clinical applicability of the current framework. The error analysis also demonstrated reduced performance in detecting subtle fractures and posterior cortical disruption and revealed segmentation-related uncertainties. These findings indicate that the current model is not sufficient to replace expert radiological interpretation. Future studies should therefore focus on larger and more diverse datasets, integration of relevant clinical parameters, and refinement of the model architecture, particularly for assessing posterior wall involvement.

The proposed framework may be used as a pre-screening tool within radiological workflows, helping to identify suspicious vertebral regions and to reduce diagnostic delays in geriatric patients.^{5,21} Such systems should be considered supportive tools rather than replacements for expert interpretation.²² Overall, these findings demonstrate the feasibility of explainable AI models for vertebral fracture assessment and their potential role in orthopedic imaging.

CONCLUSION

The proposed dual-stage deep learning framework demonstrated potential as a supportive tool for radiological detection of vertebral compression fractures in CT images. Its ability to identify morphologically relevant changes in vertebrae may facilitate early detection and prioritization within radiological workflows. Further prospective, multicenter validation is required before clinical implementation.

Ethics

Ethics Committee Approval: As the analysis was performed on fully de-identified datasets without involvement of human subjects, approval from the ethics committee was not required.

Informed Consent: No direct patient data, clinical records, or personally identifiable information were used in this study. All imaging data were obtained from publicly available sources and processed solely for research purposes. Therefore, informed consent was not required.

Footnotes

Authorship Contributions

Surgical and Medical Practices: A.Ö., A.M.B., Concept: A.Ö., A.M.B., Design: A.Ö., A.M.B., Data Collection or Processing: A.Ö., A.M.B., Analysis or Interpretation: A.Ö., A.M.B., Literature Search: A.Ö., A.M.B., Writing: A.Ö., A.M.B.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Ma Y, Lu Q, Yuan F, Chen H. Comparison of the effectiveness of different machine learning algorithms in predicting new fractures after PKP for osteoporotic vertebral compression fractures. *J Orthop Surg Res.* 2023;18:62.
2. Chettrit D, Meir T, Lebel H, et al. 3D convolutional sequence to sequence model for vertebral compression fractures identification in CT. 2020.
3. Nicolaes J, Raeymaeckers S, Robben D, et al. Detection of vertebral fractures in CT using 3D convolutional neural networks. 2020.
4. Paik S, Park J, Hong JY, Han SW. Deep learning application of vertebral compression fracture detection using mask R-CNN. *Sci Rep.* 2024;14:16308.
5. Kim YR, Yoon YS, Cha JG. Opportunistic screening for acute vertebral fractures on a routine abdominal or chest computed tomography scans using an automated deep learning model. *Diagnostics.* 2024;14:781.
6. Karadamar ÖL, Aydıle A. Evaluation of surgical recommendation algorithms in hallux valgus cases using ChatGPT models: an artificial intelligence approach based on 50 simulated scenarios. *Acta Orthop Traumatol Turc.* 2026;60:e25580.

7. Karadamar ÖL, Kara M, Özdin T, Aydılek A, Özgür A. Search trends of orthopedic terms in Turkey: a five-year time series analysis based on Google Trends. *Journal of Multidisciplinary Orthopaedic Surgery*. 2025;1:22-6.
8. Lee J, Kim M, Park H, et al. Enhanced detection performance of acute vertebral compression fractures using a hybrid deep learning and traditional quantitative measurement approach: beyond the limitations of genant classification. *Bioengineering (Basel)*. 2025;12:64.
9. Namireddy SR, Gill SS, Peerbhai A, et al. Artificial intelligence in risk prediction and diagnosis of vertebral fractures. *Sci Rep*. 2024;14:30560.
10. Seo JW, Lim SH, Jeong JG, Kim YJ, Kim KG, Jeon JY. A deep learning algorithm for automated measurement of vertebral body compression from X-ray images. *Sci Rep*. 2021;11:13732.
11. Hu YL, Wang PY, Xie ZY, et al. Interpretable machine learning model to predict bone cement leakage in percutaneous vertebral augmentation for osteoporotic vertebral compression fracture based on SHapley additive exPlanations. *Global Spine J*. 2025;15:689-701.
12. Zhao Y, Bo L, Chen X, et al. Evaluation and analysis of risk factors for adverse events of the fractured vertebra post-percutaneous kyphoplasty: a retrospective cohort study using multiple machine learning models. *J Orthop Surg Res*. 2024;19:575.
13. Choi E, Park D, Son G, et al. Weakly supervised deep learning for diagnosis of multiple vertebral compression fractures in CT. *Eur Radiol*. 2024;34:3750-60.
14. Chen J, Liu S, Li Y, et al. Deep learning model for automated detection of fresh and old vertebral fractures on thoracolumbar CT. *Eur Spine J*. 2025;34:1177-86.
15. Lee J, Park H, Yang Z, Woo OH, Kang WY, Kim JH. Improved detection accuracy of chronic vertebral compression fractures by integrating height loss ratio and deep learning approaches. *Diagnostics (Basel)*. 2024;14:2477.
16. Biamonte E, Levi R, Carrone F, et al. Artificial intelligence-based radiomics on computed tomography of lumbar spine in subjects with fragility vertebral fractures. *J Endocrinol Invest*. 2022;45:2007-17.
17. Dong ST, Zhu J, Yang H, Huang G, Zhao C, Yuan B. Development and internal validation of supervised machine learning algorithm for predicting the risk of recollapse following minimally invasive kyphoplasty in osteoporotic vertebral compression fractures. *Front Public Health*. 2022;10:874672.
18. Wu H, Li C, Song J, Zhou J. Developing predictive models for residual back pain after percutaneous vertebral augmentation treatment for osteoporotic thoracolumbar compression fractures based on machine learning technique. *J Orthop Surg Res*. 2024;19:803.
19. Singh M, Tripathi U, Patel KK, Mohit K, Pathak S. An efficient deep learning based approach for automated identification of cervical vertebrae fracture as a clinical support aid. *Sci Rep*. 2025;15:25651.
20. Inigo B, Shen Y, Killeen BD, et al. An intrinsically explainable approach to detecting vertebral compression fractures in CT scans via neurosymbolic modeling. *Proc SPIE Int Soc Opt Eng*. 2025;13406:134062R.
21. Kazerouni A, Aghdam EK, Heidari M, et al. Diffusion models in medical imaging: a comprehensive survey. *Med Image Anal*. 2023;88:102846.
22. Tahir A, Saadia A, Khan K, Gul A, Qahmash A, Akram RN. Enhancing diagnosis: ensemble deep-learning model for fracture detection using X-ray images. *Clin Radiol*. 2024;79:e1394-402.