SM Journal of Radiology

Short Communication

A CT-Based Deep Learning Model for Automated AOSpine Thoracolumbar Fracture Classification with Osteoporotic Fracture Grading

Mian Huang*

The University of Hong Kong, Hong Kong

Abstract

Background: Accurate thoracolumbar fracture classification is central to treatment planning, but manual interpretation of CT images can be time-consuming and variable.

Objective: To develop a CT-based deep learning workflow for automated vertebral localization, fracture screening, AOSpine thoracolumbar ABC classification, and Osteoporotic Fracture (OF) grading.

Methods: This retrospective study included 845 spinal CT examinations with expert consensus labels. Total Segmentator was used for vertebral segmentation and level identification. Three-dimensional vertebral CT patches were then processed by cascaded 3D ResNet-18 models. The AOSpine model fused CT image features with four automatically extracted bone-void features: total void volume, void-volume ratio, void count, and maximum void volume. The OF model used a hierarchical multi-head structure for OF1–OF5 grading. Performance was assessed using Dice, level-identification accuracy, accuracy, precision, recall, specificity, F1-score, AUC, average precision, and quadratic weighted kappa.

Results: Total Segmentator achieved a mean Dice coefficient of 0.846 and vertebral level-identification accuracy of 92.22%. The fracture screening model achieved an accuracy of 0.988. For AOSpine ABC classification, the CT plus bone-void model achieved an overall accuracy of 62.5% and macro-F1 of 0.512. For OF grading, the hierarchical CT-only model achieved an accuracy of 0.737, macro-F1 of 0.692, quadratic weighted kappa of 0.639, macro-AUC of 0.894, and macro-AP of 0.767.

Conclusion: The proposed CT-based workflow demonstrated feasibility for automated vertebral segmentation, fracture screening, AOSpine ABC classification, and OF grading. Further multicenter validation is required before clinical deployment.

Keywords: AOSpine Classification; Thoracolumbar Fracture; Computed Tomography; Deep Learning; Osteoporotic Fracture; Bone Void.

INTRODUCTION

Thoracolumbar fractures are common spinal injuries, and accurate fracture classification directly affects treatment selection, risk stratification, and follow-up planning. The AOSpine thoracolumbar injury classification system describes injury morphology and stability using A-type compression injuries and more complex B- and C-type injuries. In older patients or patients with impaired bone quality, the Osteoporotic Fracture (OF) classification provides additional clinically relevant grading of fracture severity.

Although CT is the primary imaging modality for thoracolumbar fracture assessment, manual classification requires careful vertebral localization, fracture detection, morphology interpretation, and severity grading. These steps are labor-intensive and can be affected by experience, image quality, and case complexity. Automated CT-based tools may therefore help standardize preliminary assessment and improve workflow efficiency, provided that their limitations are clearly defined [1].

Submitted: 26 May 2026 | **Accepted:** 30 June, 2026 | **Published:** 02 July, 2026

***Corresponding author:** Mian Huang; The University of Hong Kong, Hong Kong

Citation: Huang M, et al. (2026) A CT-Based Deep Learning Model for Automated AOSpine Thoracolumbar Fracture Classification with Osteoporotic Fracture Grading. SM J Radiol 9(1): 1019.

This study developed an automated CT-based workflow for thoracolumbar fracture assessment. The pipeline first performs vertebral segmentation and level identification, then screens for fracture, and finally performs AOSpine ABC classification and OF grading. For AOSpine classification, we incorporated quantitative bone-void features to provide structural information complementary to image features. For OF grading, we evaluated a hierarchical multi-head model designed to reflect the ordered nature of OF1–OF5 severity.

MATERIALS AND METHODS

Study cohort. This retrospective study included 845 spinal CT examinations from Inner Mongolia Second People's Hospital. The cohort contained 845 vertebrae, including 275 normal vertebrae and 570 fractured vertebrae. AOSpine labels and OF labels were determined by consensus annotation from orthopedic specialists [2,3].

Image preprocessing and vertebral localization. CT images were processed using Total Segmentator, an nnU-Net-based segmentation model, to generate vertebral masks and vertebral level labels. Three-dimensional vertebral patches were extracted from the target vertebrae and used as model inputs. Image intensities were normalized to the [0, 1] range during data loading (Table 1).

Bone-void feature extraction. Bone-void features were extracted from the trabecular region of the target vertebra. The vertebral mask was eroded inward by approximately 2 mm to reduce cortical bone influence. CT values were converted to equivalent bone mineral density, and connected low-density regions below 40 mg/cm3 were identified as candidate bone voids. Connected components smaller than 16.5 mm3 were excluded. Four features were retained: total bone-void volume, bone-void volume ratio, number of bone voids, and maximum bone-void volume.

Table 1: Cohort and label summary

Item	Value
Patients/vertebrae included	845 patients; 845 vertebrae
Age	50–92 years; mean 67.1 ± 8.3 years
Sex	217 men and 628 women
Vertebral status	275 normal vertebrae; 570 fractured vertebrae
AOSpine distribution	A0/A1/A2/A3/A4/B1/B2/B3/C = 28/254/37/99/68/9/57/17/1
Merged ABC classes	A0/A1/A2/A3/A4/BC; B1, B2, B3 and C were merged as BC
OF distribution	OF1/OF2/OF3/OF4/OF5 = 28/248/102/100/92

Model architecture. The classification workflow used a cascaded strategy. A binary fracture-screening model first distinguished fractured from non-fractured vertebrae. Fractured vertebrae were then passed to AOSpine ABC classification and OF grading. All classification models used 3D ResNet-18 as the image backbone [4]. The AOSpine model concatenated a 512-dimensional CT image feature vector with a 32-dimensional encoded bone-void feature vector and predicted six classes: A0, A1, A2, A3, A4, and BC, where B1, B2, B3, and C were merged because of limited sample size. The OF model used a hierarchical multi-head design with a coarse branch for OF1–OF3 versus OF4–OF5, a mild branch for OF1–OF3, and a severe branch for OF4–OF5.

Training and evaluation. A stratified 10% independent test set was held out for final assessment, and the remaining 90% development set was used for five-fold cross-validation. Models were trained with AdamW. The fracture-screening task used binary cross-entropy, the AOSpine task used cross-entropy with label smoothing, and the hierarchical OF task used focal loss for each branch. Segmentation was evaluated using Dice, IoU, HD95, and vertebral level-identification accuracy. Classification was evaluated using accuracy, precision, recall/sensitivity, specificity, F1-score, one-vs-rest AUC, average precision, and quadratic weighted kappa where applicable.

RESULTS

Cohort characteristics. The study included 845 patients, including 217 men and 628 women, aged 50–92 years (mean 67.1 ± 8.3 years). The AOSpine distribution was A0/A1/A2/A3/A4/B1/B2/B3/C = 28/254/37/99/68/9/57/17/1. The OF distribution was OF1/OF2/OF3/OF4/OF5 = 28/248/102/100/92 (Table 2 and Table 3).

Vertebral segmentation and fracture screening. Total Segmentator achieved a mean Dice coefficient of 0.846, mean IoU of 0.778, and overall vertebral level-identification accuracy of 92.22%. Thoracic and lumbar level-identification accuracies were 91.46% and 92.70%, respectively. Stratified analysis showed similar segmentation performance in fractured and non-fractured vertebrae, with median HD95 of 1.40 mm in both groups. After excluding obvious level-mismatch cases, mean Dice increased to 0.909 in fractured vertebrae and 0.918 in normal vertebrae, suggesting that low Dice outliers were mainly related to level mismatch rather than complete segmentation failure. The fracture-screening model achieved an accuracy of 0.988, precision of 0.988, recall of 0.982, and F1-score of 0.991 on the test set.

Table 2: Test-set performance for AOSpine ABC classification using CT and bone-void feature fusion.

Class	Prec.	Recall	Spec.	F1	ACC	N	AUC	AP
A0	0.0000	0.0000	1.0000	0.0000	0.9464	3	0.8679	0.2130
A1	0.8095	0.6800	0.8710	0.7391	0.7857	25	0.8761	0.8511
A2	0.6667	0.6667	0.9811	0.6667	0.9643	3	0.9811	0.6389
A3	0.4091	0.9000	0.7174	0.5625	0.7500	10	0.8413	0.4984
A4	0.8333	0.7143	0.9796	0.7692	0.9464	7	0.9650	0.7984
BC	0.5000	0.2500	0.9583	0.3333	0.8571	8	0.8438	0.4515

Table 3: Test-set class-wise performance for hierarchical OF grading

Class	Prec.	Recall	F1	ACC	N	AUC	AP
OF1	1.000	0.667	0.800	0.9825	3	0.951	0.758
OF2	0.714	1.000	0.833	0.8246	25	0.930	0.913
OF3	0.667	0.600	0.632	0.8772	10	0.851	0.696
OF4	0.833	0.500	0.625	0.8947	10	0.962	0.862
OF5	0.800	0.444	0.571	0.8947	9	0.778	0.606

AOSpine ABC classification. The CT plus bone-void fusion model achieved an overall accuracy of 62.5% and macro-F1 of 0.512 for six-class AOSpine classification. A1 and A4 showed the strongest class-wise performance, with F1-scores of 0.739 and 0.769, respectively. A2 and A3 achieved F1-scores of 0.667 and 0.563. BC remained challenging, with an F1-score of 0.333, and A0 was not correctly recalled in the test set. These results indicate that bone-void features can be integrated with CT image features for ABC classification, but minority classes and heterogeneous complex injuries remain difficult [5].

Osteoporotic fracture grading. The hierarchical CT-only OF model achieved an accuracy of 0.737, macro-F1 of 0.692, weighted-F1 of 0.718, quadratic weighted kappa of 0.639, mean absolute error of 0.456, macro-AUC of 0.894, and macro-AP of 0.767. Compared with the single-head baseline reported in the source manuscript, the hierarchical model improved accuracy from 0.702 to 0.737, macro-F1 from 0.618 to 0.692, quadratic weighted kappa from 0.406 to 0.639, and reduced mean absolute error from 0.649 to 0.456. OF5 recall improved from 0.222 to 0.444, and OF5 F1 improved from 0.267 to 0.571. However, OF4 F1 decreased from 0.762 to 0.625, suggesting possible error propagation from the coarse branch (Table 4).

Table 4: Overall performance of the hierarchical OF model.

Metric	Value
Accuracy	0.737
Macro F1	0.692
Weighted F1	0.718
Quadratic weighted kappa	0.639
MAE	0.456
Macro AUC	0.894
Macro AP	0.767

DISCUSSION

This study presents a compact CT-based workflow for automated thoracolumbar fracture assessment. The pipeline combines vertebral segmentation, fracture screening, AOSpine ABC classification, and OF grading. This design follows a clinically intuitive cascade: the system first localizes the vertebra and screens for fracture, then performs detailed classification only when a target vertebra is considered fractured.

Reliable vertebral segmentation and level identification are essential prerequisites for downstream fracture classification. Total Segmentator achieved stable segmentation performance across fractured and non-fractured vertebrae. The increase in Dice after excluding level-mismatch cases indicates that most outliers were related to anatomical level assignment, especially near adjacent thoracolumbar levels, rather than failure of vertebral mask generation. Future implementations should therefore include robust level verification or a lightweight manual correction step [6].

For AOSpine ABC classification, the model incorporated bone-void features in addition to CT image features. Bone voids reflect connected low-density regions within trabecular bone and may capture structural fragility that is not explicitly represented by standard convolutional image features. The fusion model performed reasonably for A1 and A4 but remained weak for A0 and BC. This limitation is expected because A0 had very limited test support and the merged BC category combined biomechanically heterogeneous B- and C-type injuries. These class-wise estimates should therefore be interpreted cautiously.

The OF grading experiment suggests that hierarchical modeling may be useful when target labels have an ordered severity structure. By decomposing OF1–OF5 into coarse severity grouping and within-group classification, the model improved macro-F1, quadratic weighted kappa, and OF5 detection compared with the single-head baseline [7]. However, the decline in OF4 performance highlights a known risk of hierarchical systems: errors at an upstream branch can constrain downstream decisions. More robust coarse-branch calibration and larger samples for severe categories are needed.

This study has several limitations. First, it was retrospective and apparently single-center, which limits generalizability. Second, several categories had small support, particularly A0, A2, and the B/C subclasses, making class-wise estimates unstable. Third, B and C injuries were merged as BC, which improves feasibility but reduces anatomical specificity. Fourth, external validation and prospective workflow testing were not performed. Multicenter validation, improved minority-class modeling, and standardized reporting of failure cases are required before clinical deployment.

CONCLUSION

The proposed CT-based deep learning workflow demonstrated feasibility for automated vertebral segmentation, fracture screening, AOSpine ABC classification, and OF grading. Bone-void features provided structurally interpretable information for ABC classification, while hierarchical modeling improved several OF grading metrics, particularly for severe OF categories. Further multicenter external validation is required to determine robustness and clinical utility.

DECLARATIONS

Ethics approval, informed consent, funding, conflicts of interest, and data availability statements will be provided or updated before formal editorial processing.

REFERENCES

1. Zhang J, Liu F, Xu J, Zhao Q, Huang C, Yu Y, et al. Automated detection and classification of acute vertebral body fractures using a convolutional neural network on computed tomography. Front Endocrinol (Lausanne). 2023; 14: 1132725.
2. Liawrungrueang W, Cho ST, Kotheeranurak V, Jitpakdee K, Kim P, Sarasombath P. Osteoporotic Vertebral Compression Fracture (OVCF) detection using artificial neural networks model based on the AO spine-DGOU osteoporotic fracture classification system. N Am Spine Soc J. 2024; 19: 100515.
3. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.
4. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proc IEEE Conf Comput Vis Pattern Recognit. 2016: 2818-2826.
5. Lin TY, Goyal P, Girshick P, He K, Dollar P. Focal loss for dense object detection. Proc IEEE Int Conf Comput Vis. 2017:2980-2988.
6. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101.
7. Lin J, Liu Z, Fu G, Zhang H, Chen C, Qi H, et al. Distribution of bone voids in the thoracolumbar spine in Chinese adults with and without osteoporosis: A cross-sectional multi-center study based on 464 vertebrae. Bone. 2023; 172: 116749.