scientific reports

OPEN

Augmented prediction of vertebral collapse after osteoporotic vertebral compression fractures through parameter-efficient fine-tuning of biomedical foundation models

Sibeen Kim[1,13], Inkyeong Kim[2,3,13], Woon Tak Yuh[4,5,13], Sangmin Han[6], Choonghyo Kim[2,3], Young San Ko[7,8], Wonwoo Cho[9,10,14]✉ & Sung Bae Park[11,12,14]✉

Vertebral collapse (VC) following osteoporotic vertebral compression fracture (OVCF) often requires aggressive treatment, necessitating an accurate prediction for early intervention. This study aimed to develop a predictive model leveraging deep neural networks to predict VC progression after OVCF using magnetic resonance imaging (MRI) and clinical data. Among 245 enrolled patients with acute OVCF, data from 200 patients were used for the development dataset, and data from 45 patients were used for the test dataset. To construct an accurate prediction model, we explored two backbone architectures: convolutional neural networks and vision transformers (ViTs), along with various pre-trained weights and fine-tuning methods. Through extensive experiments, we built our model by performing parameter-efficient fine-tuning of a ViT model pre-trained on a large-scale biomedical dataset. Attention rollouts indicated that the contours and internal features of the compressed vertebral body were critical in predicting VC with this model. To further improve the prediction performance of our model, we applied the augmented prediction strategy, which uses multiple MRI frames and achieves a significantly higher area under the curve (AUC). Our findings suggest that employing a biomedical foundation model fine-tuned using a parameter-efficient method, along with augmented prediction, can significantly enhance medical decisions.

Keywords Compression fracture, Spine, Biomedical foundation model, Vision transformer, Parameter-efficient fine-tuning

[1]School of Biomedical Engineering, Korea University, Seoul, Republic of Korea. [2]Department of Neurosurgery, Kangwon National University Hospital, Chuncheon-si, Gangwon-do, Republic of Korea. [3]Department of Neurosurgery, Kangwon National University College of Medicine, Chuncheon-si, Gangwon-do, Republic of Korea. [4]Department of Neurosurgery, Hallym University College of Medicine, Chuncheon-si, Gangwon-do, Republic of Korea. [5]Department of Neurosurgery, Hallym University Dongtan Sacred Heart Hospital, Hwaseong-si, Gyeonggi-do, Republic of Korea. [6]Department of Intelligence Convergence, Yonsei University, Seoul, Republic of Korea. [7]Department of Neurosurgery, Kyungpook National University Hospital, 130 Dongdeok-ro, Daegu 41944, Republic of Korea. [8]Department of Neurosurgery, School of Medicine, Kyungbook National university, Daegu, Republic of Korea. [9]Kim Jaechul Graduate School of Artificial Intelligence, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. [10]Letsur Inc, 27, Teheran-ro 2-gil, Gangnam-gu, Seoul, Republic of Korea. [11]Department of Medical Device Development, Seoul National University College of Medicine, Seoul, Republic of Korea. [12]Department of Neurosurgery, Seoul National Boramae Medical Center, Seoul, Republic of Korea. [13]Sibeen Kim, Inkyeong Kim and Woon Tak Yuh contributed equally to this work as first authors. [14]Sung Bae Park and Wonwoo Cho contributed equally to this work as corresponding authors. ✉email: wcho@kaist.ac.kr; ddolbae01@naver.com

Vertebral compression fractures frequently occur in osteoporotic spines[1,2]. Osteoporotic vertebral compression fractures (OVCFs) are common in older adults with decreased bone mineral density[3,4]. OVCFs with osteoporosis usually cause back pain and require conservative treatments, such as bed rest, painkillers, bracing, and osteoporosis medication. These treatments often lead to good functional recovery[5,6]. Although most vertebral fractures may heal within eight weeks, vertebral collapse (VC) progresses over time in 7–37% of the patients with vertebral compression fractures[7]. The progression of OVCFs can lead to VC, spinal deformity, chronic back pain, and neurological deficits due to spinal cord compression. Therefore, it is clinically valuable to predict whether OVCFs will progress into VC as early as possible[6,8].

Although recent studies have identified many factors related to the progression of OVCFs, such as bone turnover markers, fracture shape, morphometric measurements, and magnetic resonance imaging (MRI) findings, predicting this progression at the time of diagnosis remains challenging[5,6,9,10]. Recently, machine learning (ML)-based prediction algorithms have been widely employed in medical applications. ML-based image analysis models such as convolutional neural networks (CNNs) have shown promising results in extracting robust and informative features from medical images. For instance, auto-segmentation models of vertebrae and detection of acute and chronic OVCFs using CNNs have been reported on computed tomography (CT) and MRI scans[11,12]. However, to the best of our knowledge, no studies using image analysis models have focused on predicting progressive VC after OVCF based on initial diagnostic MRIs and clinical information. The development of a predictive tool for assessing the progression of VC after OVCF can be used to guide the initial aggressive treatment and improve the functional outcomes for patients with OVCFs.

In the present study, we aimed to develop a predictive support tool and enhance its performance using ML-based image analysis models on a small dataset including initial MRI and clinical information. Based on recent advances in vision foundation models, which have been developed by pre-training vision transformer (ViT)-based models with large-scale data, we constructed our prediction model by fine-tuning a biomedical foundation model in a parameter-efficient manner. Additionally, we further enhanced our model's prediction performance by applying the augmented prediction technique. We assessed the prediction performance and generalizability of ML-based image analysis models by conducting both internal and external evaluations of our model and other CNN and ViT-based baseline models.

Methods

Study population

This retrospective study collected data from patients with OVCFs from five institutions. The study protocol was approved by the Institutional Review Board (IRB) of Seoul National Boramae Medical Center (No 20-2020-200) and conducted in accordance with the Declaration of Helsinki tenets for research involving human subjects. A waiver permission letter was obtained from IRB administrators before the data collection and since the patients with OVCFs were not directly involved in this study (the data were obtained from chart review), informed consent was not required, but the extracted data from the medical records were kept confidentially. The informed consent was waived by IRB. Two hundred forty-five patients (aged ≥ 50 years) with OVCF between January 2010 and December 2020 were enrolled in the study. The inclusion criteria for these patients were: (1) diagnosed with acute OVCF in the thoracic or lumbar spine by MRI, and (2) availability of follow-up X-ray or CT images for over six months after the initial diagnosis of acute OVCF. The exclusion criteria were the detection of spine infection, vertebroplasty, tumor, or spine implants at the time of MRI diagnosis and during the follow-up period. VC was defined as a compressed anterior or central vertebral body height of less than 50% of the posterior height[1]. Patients with VC observed in X-ray or CT during the six-month follow-up period were assigned to the VC group, while others were assigned to the non-VC group. The proportion of VC and the number of included patients varied across institutions (Supplementary Table 1). To balance the proportion of VC in the development dataset while ensuring the test dataset was not too small, we assigned the data from three institutions (Seoul National Boramae Medical Center, Kangwon National University Hospital, Hallym University Dongtan Sacred Heart Hospital) into the development dataset for training and internal validation of the VC prediction models. Data from the remaining two institutions (Keimyung University Dongsan Hospital, Soon Chun Hyang University Hospital Bucheon) were assigned into the test dataset for external validation of the VC prediction models.

Image acquisition

In this study, vertebrae images were acquired using a 3T MRI scanner, which is commonly used in hospitals for high-resolution imaging. Specifically, we focused on T1- and T2-weighted sequence sagittal images, known for their excellent contrast between the different soft tissues. From these sequences, a single key frame image that prominently displayed vertebral fractures was selected by expert spine surgeons. The selection criteria for this image were based on the clarity and visibility of the fracture to ensure accurate annotation and analysis. To ensure reproducibility and to provide context for our image acquisition protocol, the following settings were typically used for our T1-weighted MRI scans: slices per group, 15; distance factor, 10%; position, isocenter; phase encoding direction, head to feet; phase oversampling, 50%; field of view, 200 · 200 mm; slice thickness, 3.0 mm; repetition time (TR), 480.0 ms; echo time (TE) 7.10 ms; flip angle, 125°; average, 2; and concatenation, 2.

Model development and additional techniques for accurate prediction

To develop ML-based image analysis models, we conducted image pre-processing to enhance consistency across the MRI scans. Initially, we applied N4 bias field correction[13] using the SimpleITK[14] library to correct non-uniformities in the MRI intensities. Expert spine surgeons then identified landmarks for the most important vertebra within the key frame for analysis. The tight bounding box defined by the landmarks was expanded by

150% horizontally and vertically to ensure comprehensive coverage of the vertebral body and the surrounding structures, while minimizing unnecessary background regions. After cropping the image within the expanded bounding box, we applied quantile clipping between the 5th and 95th percentiles to handle outliers and performed min-max normalization to standardize the image intensities.

Using these pre-processed MRI scans, we explored various backbone architectures, pre-trained weights, and fine-tuning methods. For backbone architectures, we employed ResNet-18[15] and ViT-B/16[16]. ResNet-18 is a CNN architecture conventionally used in ML-based image analysis, while ViT-B/16 is a vision transformer model specialized in capturing intricate and wide-range dependencies across image features. For ResNet-18, we considered two initialization strategies: random initialization (scratch) and ImageNet pre-trained weights. For ViT-B/16, in addition to the scratch and ImageNet pre-trained settings, we used BiomedCLIP[17] weights pre-trained on PMC-15 M, which consists of 15 million biomedical image-text pairs.

When using the CNN and ViT backbones initialized with the pre-trained weights, we primarily employed full-parameter fine-tuning, which involves updating all weights in the model during the training process. However, for ViT-B/16, which has a large number of parameters, we also considered a parameter-efficient fine-tuning method called Low-Rank Adaptation (LoRA)[18]. LoRA injects trainable low-rank matrices into weight matrices, allowing for efficient adaptation with fewer parameters and reducing the computational requirements while preventing overfitting.

After designing our ML-based image analysis model, we explored two additional techniques to enhance its robustness. First, we used the augmented prediction approach that incorporates multiple frames from each MRI scan. Specifically, we utilized not only the key frame selected by experts but also its two adjacent frames during both the training and inference phases. By training our model with the original key frames and their adjacent frames, the model assesses the risk of VC progression for each patient by evaluating the three frames and then averaging their prediction probabilities during inference. We expect that this augmented prediction strategy would improve robustness, especially when trained with small-scale data, resulting in more consistent and accurate predictions. Second, we provided clinical features as additional information to our image analysis model. These features included multiple variables: age, bone mineral density (BMD), gender, pre-fracture medication for osteoporosis, and post-fracture medication for osteoporosis. To effectively incorporate these features into the image model, we extracted deep features using a multi-layer perceptron (MLP) after standardization. The MLP features were then concatenated with the image features extracted from the image model.

In summary, we conducted MRI pre-processing, explored various backbone architectures with pre-trained weights, and applied a parameter-efficient fine-tuning method for model development. Additionally, we implemented the augmented prediction approach and incorporated clinical features to enhance prediction robustness. Our structured workflow is depicted in Fig. 1.

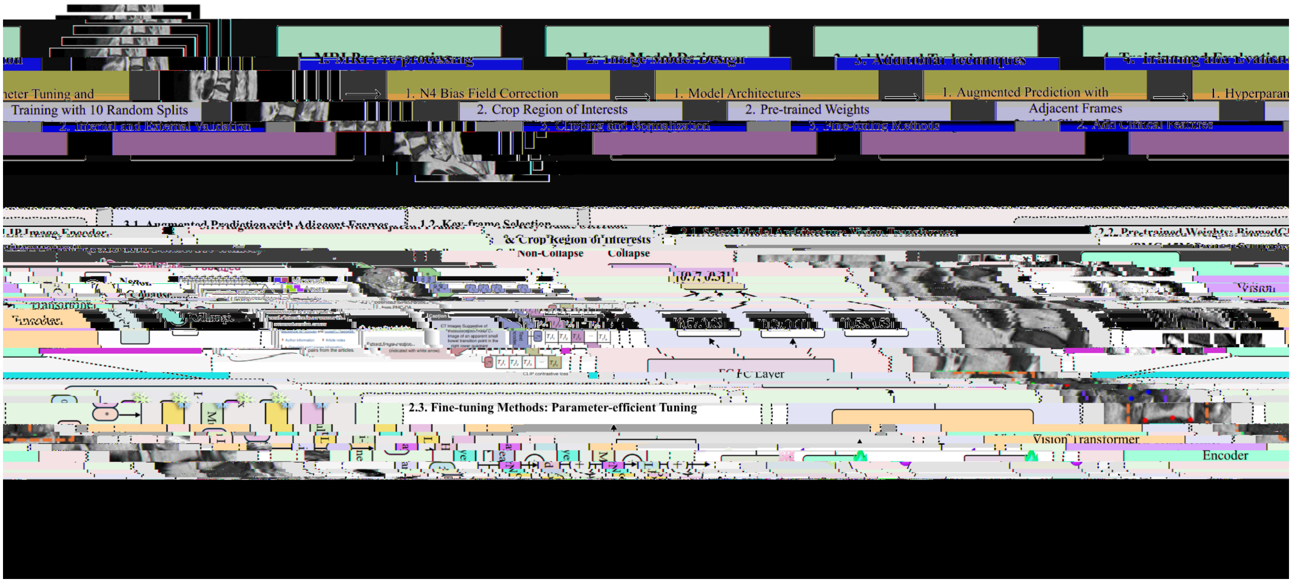
1. MRI Pre-processing
2. Image Model Design
3. Additional Techniques
4. Training and Evaluation
1. N4 Bias Field Correction
2. Crop Region of Interests
3. Clipping and Normalization
1. Model Architectures
2. Pre-trained Weights
3. Fine-tuning Methods
1. Augmented Prediction with Adjacent Frames
2. Add Clinical Features
1. Hyperparameter Tuning and Training with 10 Random Splits
2. Internal and External Validation
3.1. Augmented Prediction with Adjacent Frames
1.2. Key-frame Selection & Crop Region of Interests
Non-Collapse
Collapse
2.1. Select Model Architecture: Vision Transformer
2.2. Pre-trained Weights: BiomedCLIP
FC Layer
2.3. Fine-tuning Methods: Parameter-efficient Tuning
Vision Transformer
Encoder

Fig. 1. Workflow of VC prediction model development. MRI pre-processing includes N4 bias field correction, cropping to the region of interest, and intensity normalization. Image model design highlights the Vision Transformer (ViT) architecture with BiomedCLIP pre-trained weights and parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). Additional techniques explored to further enhance the model's performance include augmented prediction with adjacent MRI frames and addition of clinical features via a multi-layer perceptron (MLP). Model training and evaluation involve hyperparameter tuning, internal validation using 10 random splits, and external validation.

Training details

We experimented with using T1-weighted MR images (T1WI), T2WI, and both T1 and T2WIs. From our analysis, it was found that using T1WI only led to the best performance and was simpler compared to using both T1 and T2WIs. Thus, we used T1WI only as the image input. The pre-processed T1WI frames were resized to 224 · 224 pixels, and the grayscale frames were duplicated across three channels.

Data augmentation techniques included random shifts (–20% to + 20%), scaling (0.8· to 1.2·), and rotations (–50° to + 50°), each applied with a 0.5 probability. Additionally, random brightness and contrast adjustments and random gamma adjustments were applied, each with a probability of 0.5. The images were normalized based on the mean and standard deviation values according to the pre-trained model specifications. We used the binary cross entropy[19] loss function for training.

Hyperparameters for each model were determined through grid search, optimizing for the area under the curve (AUC). Detailed information on the hyperparameters can be found in Supplementary Table S2. We explored additional techniques, including augmented prediction with adjacent frames and the incorporation of clinical features, with the best-performing model.

All experiments were conducted using PyTorch 2.2.0[20] and four NVIDIA Tesla V100 GPUs with 32 GB of memory each.

Model evaluation

We used specific notations to indicate models developed with different backbones, pre-training datasets, and fine-tuned methods. The backbone architectures were CNNs and ViTs. The pre-training datasets included scratch, ImageNet, and PMC. The full fine-tuning or LoRA methods were used. We denoted each model by combining these terms, such as CNN-scratch-full to indicate a CNN model trained from scratch.

We compared six image models to identify the best-performing model: CNN-scratch-full, CNN-ImageNet-full, ViT-scratch-full, ViT-ImageNet-full, ViT-PMC-full, and ViT-PMC-LoRA. We calculated the mean and standard deviation for AUC, specificity, and sensitivity. Optimal cutoff values for receiver operating characteristic analysis were determined from Youden's J statistic[21]. Paired t-tests were employed to compare the AUC of the best-performing model with that of the other models, with the Bonferroni correction applied to account for multiple comparisons[22].

After identifying the best-performing image model, we visualized gradient-weighted class activation mappings (Grad-CAMs[23]) and attention rollouts[24] to gain insights into its decision-making process. Grad-CAM highlights the regions of the input image that are most important for making our model's predictions. Attention rollout visualizes how our model distributes attention across different parts of the input image. We categorized both the Grad-CAMs and attention rollouts into true positive, true negative, false positive, and false negative for our post-hoc analysis.

To further enhance the prediction performance of our best-performing model, we investigated the efficacy of augmented prediction and incorporation of clinical features. We compared four configurations: without augmented prediction or clinical features, with augmented prediction only, with clinical features only, and with both augmented prediction and clinical features. The evaluation metrics and statistical analysis method were identical to those used for the comparison of image models.

Results

Patient characteristics

In this study, the patient characteristics between the non-VC group ($n = 125$, 51.0%) and the VC group ($n = 120$, 49.0%) showed no significant differences. Detailed information on these characteristics can be found in Supplementary Table S3. The development dataset comprised 200 patients (81.6%) sourced from three institutions, with 109, 55, and 36 patients, respectively, and was split into 10 distinct subsets for training and internal validation (80:20). The test dataset included 45 patients (18.4%), with 30 and 15 patients from two additional institutions, used for external validation. The proportion of VC was 51.0% in the development dataset and 40.0% in the test dataset, with no significant difference ($p = 0.243$). Aside from the T-score of the BMD being lower in the test dataset compared to the development dataset (–3.52 ± 1.16 vs. − 3.08 ± 1.04, $p = 0.013$), no significant differences were found in patient characteristics between these datasets. Further details on the patient characteristics in both datasets are summarized in Table 1.

Performance comparison of image models

In internal validation, the mean AUCs from the 10 distinct dataset splits were 0.7830, 0.7760, 0.8149, 0.8185, 0.8269, and 0.8404 for CNN-scratch-full, CNN-ImageNet-full, ViT-scratch-full, ViT-ImageNet-full, ViT-PMC-full, and ViT-PMC-LoRA, respectively (Table 2). The mean AUC of ViT-PMC-LoRA was significantly higher than that of CNN-ImageNet-full ($P < 0.001$). In external validation, the mean AUCs were 0.7097, 0.7772, 0.7784, 0.7825, 0.8051, and 0.8113 for CNN-scratch-full, CNN-ImageNet-full, ViT-scratch-full, ViT-ImageNet-full, ViT-PMC-full, and ViT-PMC-LoRA, respectively. The mean AUC of ViT-PMC-LoRA was significantly higher than that of CNN-scratch-full ($P = 0.004$). Notably, ViT-PMC-LoRA demonstrated the highest mean AUC and sensitivity among all models. Furthermore, ViT-PMC-LoRA shows consistently higher true positive rates at most false positive rates compared to CNN-ImageNet-full (Fig. 2). To address potential concerns about institution-specific biases, we also conducted two separate leave-one-institution-out validations. In both cases, ViT-PMC-LoRA achieved the highest mean AUC among all models in both internal and external validations (Supplementary Results). Thus, we chose ViT-PMC-LoRA as our finalized image model.

	Overall (*n* = 245)		
	Development dataset (*n* = 200, 81.6%)	Test dataset (*n* = 45, 18.4%)	*p*-value
VC, *n* (%)	102 (51.0)	18 (40.0)	0.243
Age, mean ± SD (years)	72.6 ± 9.56	73.7 ± 8.56	0.460
Female, *n* (%)	166 (83.0)	31 (68.9)	0.052
T-score of BMD, mean ± SD (Lumbar)	-3.08 ± 1.04	-3.52 ± 1.16	0.013*
Lumbar Fracture, *n* (%)	134 (67.0)	29 (64.4)	0.878
History of medication for osteoporosis, *n* (%)			
Before OVCF Dx	42 (21.0)	8 (17.8)	0.780
After OVCF Dx	153 (76.5)	36 (80.0)	0.758

Table 1. Patient characteristics in the development and test datasets. p-values less than 0.05 are considered statistically significant. *VC* Vertebral collapse, *SD* standard deviation, *BMD* bone mineral density, *OVCF* osteoporotic vertebral compression fracture, *dx* diagnosis *Indicates a statistically significant difference.

	Params	AUC mean (SD)	Specificity mean (SD)	Sensitivity mean (SD)	*p*-value
Internal validation using the development dataset					
CNN-scratch-full	11 M	0.7830 (0.0862)	0.7844 (0.1332)	0.6699 (0.1698)	0.045
CNN-ImageNet-full	11 M	0.7760 (0.0397)	0.7910 (0.1415)	0.6650 (0.1321)	< 0.001*
ViT-scratch-full	86 M	0.8149 (0.0460)	0.8585 (0.0856)	0.6595 (0.1097)	0.04
ViT-ImageNet-full	86 M	0.8185 (0.0308)	**0.8642** (0.0967)	0.6517 (0.1243)	0.07
ViT-PMC-full	86 M	0.8269 (0.0353)	0.8057 (0.1104)	**0.7306** (0.0761)	0.397
ViT-PMC-LoRA	**0.89 M**	**0.8404** (0.0312)	0.8557 (0.0953)	0.7012 (0.1154)	-
External validation using the test dataset					
CNN-scratch-full	11 M	0.7097 (0.0623)	0.6963 (0.1770)	0.6611 (0.1469)	0.004*
CNN-ImageNet-full	11 M	0.7772 (0.0247)	0.6778 (0.0941)	0.7889 (0.0820)	0.062
ViT-scratch-full	86 M	0.7784 (0.0211)	**0.7519** (0.0973)	0.6667 (0.1080)	0.107
ViT-ImageNet-full	86 M	0.7825 (0.0248)	0.7333 (0.0969)	0.7333 (0.1041)	0.195
ViT-PMC-full	86 M	0.8051 (0.0347)	0.7296 (0.1004)	0.7556 (0.1148)	0.768
ViT-PMC-LoRA	**0.89 M**	**0.8113** (0.0519)	0.6963 (0.1155)	**0.8111** (0.0915)	-

Table 2. Vertebral collapse (VC) prediction performances of image models with various backbone architectures, pre-train datasets, and fine-tune methods, along with the number of trainable parameters. Each model is compared against the best-performing model, ViT-PMC-LoRA. Bonferroni correction for multiple comparisons across 5 tests was applied for internal and external validation, respectively. Thus, p-values for area under the curve (AUC) are considered statistically significant if less than 0.010. *Params* number of trainable parameters, *SD* standard deviation, *CNN* convolutional neural network, *ViT* vision transformer, *LoRA* Low-Rank Adaptation. *Indicates a statistically significant difference. The best value for each column is marked in bold.

Model interpretation

Grad-CAM demonstrated that the model highlights the cortex of the vertebrae for decision-making, with less emphasis on the trabecular areas. Attention rollout indicates that the model considers both the cortex and trabecular areas during inference (Fig. 3). In cases where predictions were correct, both Grad-CAM and attention rollout show that the model consistently focused on the cortex for decision-making. On the other hand, misclassified cases were typically associated with vertebral fractures exhibiting highly irregular shapes. For such cases, Grad-CAM continued to focus on the cortex of the vertebra, and attention rollout appeared to lose focus, spreading its attention across the entire vertebral body rather than concentrating on specific regions.

Impact of additional techniques on model performance

In internal validation, the mean AUCs from the 10 distinct dataset splits were 0.8307, 0.8404, 0.8502, and 0.8539 for ViT-PMC-LoRA with clinical features only, without augmented prediction or clinical features, with both augmented prediction and clinical features, and with augmented prediction only, respectively (Table 3). In external validation, the mean AUCs were 0.8103, 0.8113, 0.8566, and 0.8656 for ViT-PMC-LoRA with clinical features only, without augmented prediction or clinical features, with both augmented prediction and clinical features, and with augmented prediction only, respectively. Only the mean AUC of ViT-PMC-LoRA with augmented prediction was significantly higher than that of with clinical features only ($P < 0.001$) and without augmented prediction or clinical features ($P = 0.011$). However, there was no significant difference compared to

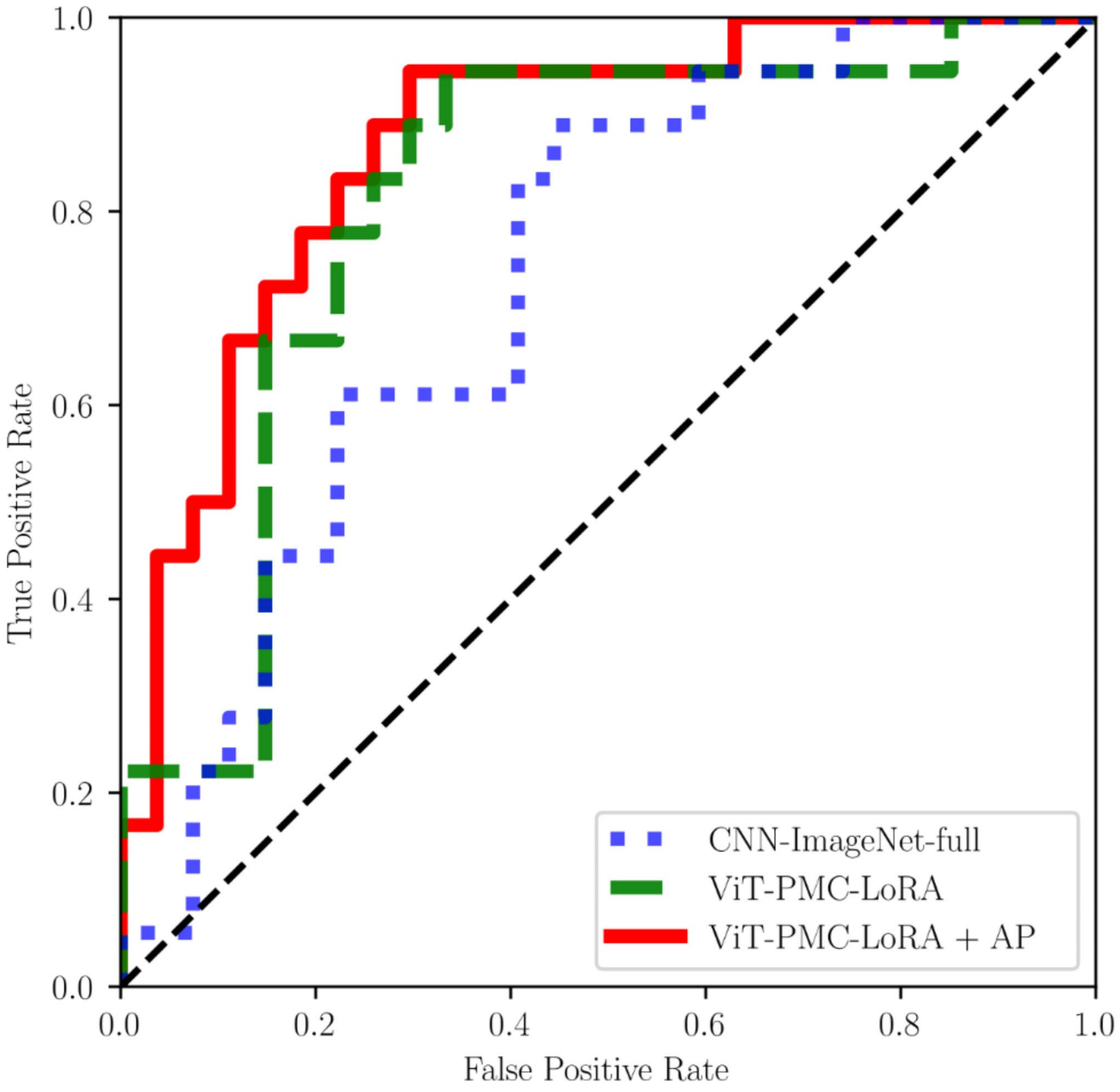

Fig. 2. Performances of vertebral collapse (VC) prediction models in the test dataset. ViT-PMC-LoRA consistently outperforms CNN-ImageNet-full. Using augmented prediction (AP) further enhances the performance of ViT-PMC-LoRA.

the mean AUC with both augmented prediction and clinical features ($P = 0.172$). Notably, ViT-PMC-LoRA with augmented prediction only demonstrated the highest mean AUC, while ViT-PMC-LoRA with both augmented prediction and clinical features showed a comparable mean AUC and the highest sensitivity among all models. Furthermore, using augmented prediction consistently enhances the true positive rate of ViT-PMC-LoRA at most false positive rates (Fig. 2).

Discussion

The primary objective of this study was to develop an accurate predictive model for the progression of VC in OVCFs by extracting overall risk factors from a small dataset of MRI and clinical data. Our findings demonstrate that ViT-PMC-LoRA, the ViT model pre-trained on PMC-15 M and fine-tuned with LoRA, achieved the highest performance, surpassing other models with various backbone architectures, pre-trained weights, and fine-tuning methods. Furthermore, the use of the augmented prediction technique substantially improved the model's prediction performance.

Deep neural networks have been extensively applied in vertebral fracture analysis, with CNNs becoming the standard approach across various tasks. These tasks range from segmenting vertebral structures to detecting

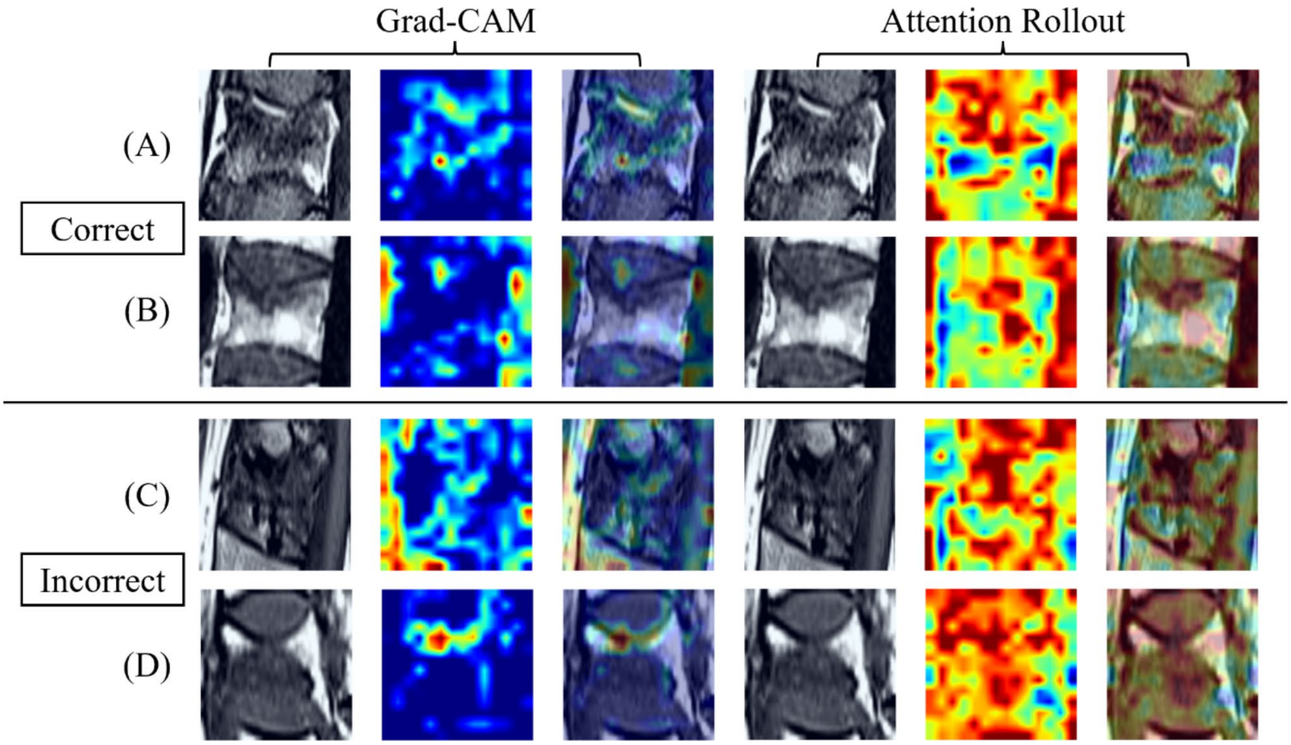

Fig. 3. Grad-CAMs and attention rollouts for representative cases predicted by ViT-PMC-LoRA. (**A**, **B**) are cases that were correctly predicted, and (**C**, **D**) are cases that were incorrectly predicted. (**A**) True Positive, (**B**) True Negative, (**C**) False Negative, (**D**) False Positive.

	AUC mean (SD)	**Specificity mean (SD)**	**Sensitivity mean (SD)**	***p*-value**
Internal validation using the development dataset				
ViT-PMC-LoRA + CF	0.8307 (0.0372)	0.8016 (0.1033)	0.7418 (0.1229)	0.179
ViT-PMC-LoRA	0.8404 (0.0312)	**0.8557** (0.0953)	0.7012 (0.1154)	0.442
ViT-PMC-LoRA + AP + CF	0.8502 (0.0297)	0.8234 (0.0720)	**0.7773** (0.0594)	0.775
ViT-PMC-LoRA + AP	**0.8539** (0.0445)	0.8230 (0.0893)	0.7739 (0.0796)	-
External validation using the test dataset				
ViT-PMC-LoRA + CF	0.8103 (0.0169)	0.6741 (0.0969)	**0.8611** (0.0878)	< 0.001*
ViT-PMC-LoRA	0.8113 (0.0519)	0.6963 (0.1155)	0.8111 (0.0915)	0.011*
ViT-PMC-LoRA + AP + CF	0.8566 (0.0246)	0.7630 (0.0804)	0.8167 (0.0695)	0.172
ViT-PMC-LoRA + AP	**0.8656** (0.0137)	**0.8111** (0.1010)	0.7611 (0.1173)	-

Table 3. Impact of augmented prediction and incorporation of clinical features on vertebral collapse (VC) prediction performance of ViT-PMC-LoRA. Each configuration is compared against the best-performing configuration, ViT-PMC-LoRA + AP. Bonferroni correction for multiple comparisons across 3 tests was applied for internal and external validation, respectively. Thus, p-values for area under the curve (AUC) are considered statistically significant if less than 0.017. *AUC* Area under the curve, *SD* standard deviation, *CNN* convolutional neural network, *ViT* vision transformer, *LoRA* Low-Rank Adaptation, *AP*, augmented prediction, *CF*, incorporation of clinical features. *Indicates a statistically significant difference. The best value for each column is marked in bold.

and classifying fractures in medical images. For instance, CNN-based models have been developed to enhance spine fracture segmentation using CT[25–27] and MRI[28] data, employed for predicting fracture risk with CT[29], and detecting vertebral fractures with CT[12,30] and MRI[11,31]. While CNNs have been the go-to approach, there has been a growing trend in medical image analysis to leverage large, pre-trained models, which are fine-tuned for downstream tasks, especially in scenarios with limited data. This shift towards parameter-efficient fine-tuning (PEFT) has shown significant potential in enhancing performance for small datasets, as highlighted in recent studies. PEFT has been proven effective in low-data scenarios, improving the transferability to discriminative medical tasks[32]. It has even been suggested that PEFT can outperform full fine-tuning in some medical applications[33], demonstrating its suitability for situations where available data is sparse. Comparative studies between CNNs and ViTs in medical AI research have also emerged, showing that ViTs can outperform CNNs in some tasks. For example, ViTs have shown superior performance in coronary plaque diagnosis using computed tomography angiography[34] and osteoporosis detection from X-ray images[35]. However, in the context of vertebral fractures, the use of large models like ViTs remains underexplored. Given this gap, our study aimed to develop and compare models using both CNN and ViT backbones, with a particular focus on MRI analysis and limited data, addressing the challenge of predicting vertebral collapse. Unlike detection tasks, which often involve signals for identifying current conditions, prediction tasks such as ours must capture weaker signals to foresee future disease progression.

In comparison with previous studies that primarily utilized CNNs for OVCF tasks, our approach using ViTs represents a novel and effective advancement in vertebral collapse prediction. The superior performance of ViT models, particularly those fine-tuned with domain-specific PMC-15 M pre-trained weights, demonstrates the importance of leveraging large, specialized datasets in medical AI applications[36]. A key challenge in vertebral collapse prediction lies in the weak signal present in MRI data and the limited size of the dataset. By utilizing a model pre-trained on large biomedical data, we were able to mitigate some of these challenges. The LoRA fine-tuning method further enhanced the model's capability by efficiently adapting the extensive parameters of the ViT model without overfitting, even with a relatively small dataset. This result aligns with previous research that highlighted the effectiveness of parameter-efficient fine-tuning approaches in small medical dataset scenarios[32]. Moreover, ViT-PMC-LoRA achieved the highest AUC and the highest sensitivity among all models, which is critical for accurately identifying patients at risk for VC. This high sensitivity, particularly in external validation, suggests that the model can aid in early clinical diagnosis and proactive treatment planning before vertebral collapse occurs. Thus, our study reinforces the hypothesis that sophisticated neural networks, when fine-tuned with parameter-efficient methods, can offer more accurate predictions in medical contexts, even when the available dataset is small.

The augmented prediction technique notably enhanced the model performance, reflecting the benefit of incorporating multiple frames from MRI scans to mitigate noise and anomalies. This approach resulted in more robust and consistent predictions, which is critical in clinical settings where precision is paramount. Interestingly, while the integration of clinical features did not significantly improve the model's AUC, it did increase sensitivity. This suggests that the inclusion of clinical data, such as age, bone mineral density (BMD), and osteoporosis-related medication, helped to detect more positive cases, complementing the image-based model. However, this also indicates that the imaging model alone may be sufficiently powerful, suggesting that MRI-derived features can capture the essential information needed for accurate prediction. Additionally, with more optimized representation and integration of medical domain knowledge, the performance of the model combining both augmented prediction and clinical features could potentially be improved[37].

The clinical implications of our predictive tool are profound. The early prediction of VC can significantly affect treatment decisions, allowing for timely and aggressive interventions that may improve patient outcomes. Given that our model was developed using a small dataset, it shows promise for use in medical fields where data-sharing is challenging[38]. Integrating this tool into clinical workflows could streamline decision-making processes and enhance the management of patients with OVCF, ultimately reducing the incidence of severe complications, such as chronic pain and neurological deficits. The broader application of AI in medical imaging and diagnostics is further supported by our study, highlighting that advanced AI models can augment traditional diagnostic methods and provide critical insights into disease progression. In future work, employing our model as an initial framework in a federated learning approach could help mitigate data insufficiency, potentially enhancing its performance and applicability[39].

Model interpretability remains a crucial aspect of deploying AI in clinical practice. Our findings suggest that attention rollouts provide better interpretability by considering both cortical and trabecular regions compared to Grad-CAM, which focuses primarily on the cortex. However, both methods faced difficulties with certain misclassified cases, particularly in instances where vertebral fractures exhibited highly irregular shapes. In these cases, Grad-CAM tended to remain focused on the cortex, potentially overlooking significant details in other regions, while attention rollout dispersed its focus too broadly across the vertebral body, failing to concentrate on critical areas. This may be due to insufficient learning from the small dataset, indicating the need for further refinement and training, particularly when dealing with complex fracture morphologies. Future research should aim to improve the evaluation of trabecular areas, potentially by balancing the focus between cortical and trabecular regions, to enhance model performance in difficult scenarios. This approach could lead to more comprehensive and accurate interpretations, ultimately benefiting clinical decision-making.

While our study presents significant advancements, it also has limitations that need to be addressed. First, the retrospective design, while useful for initial model development, can introduce biases, such as selection bias, that may affect the generalizability of the findings[40]. Second, the relatively small dataset poses challenges to the robustness of the model[41]. We applied parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA) to mitigate overfitting while maintaining model performance. We also incorporated data augmentation, including random slice selection, and weight regularization. Nonetheless, these strategies cannot completely eliminate the risk of overfitting. Furthermore, we acknowledge that the sample size used for external validation is limited for robust generalization. This is largely due to the clinical practice in OVCF cases, where patients frequently undergo surgical or invasive procedures, making it rare to find cases with more than six months of non-interventional follow-up. Despite gathering data from multiple institutions, the number of eligible cases remained low. Thus, future studies should aim to incorporate larger datasets to enhance model training and validation. Third, the reliance on multi-institution data enhances the model's applicability across different settings but also adds variability in imaging protocols and quality, which could affect the model's performance. A key challenge in dividing the data into development and test sets was ensuring a balanced proportion of VC in the development set while avoiding an overly small test set. Class imbalance in the development set could result in model bias toward the majority class during training, potentially diminishing performance. Moreover, a small test set constrains the ability to reliably evaluate the model's generalization capacity. To address this, we allocated data from three institutions to the development set, ensuring a balanced VC proportion, while assigning the remaining two institutions to the test set, which, although not large, represented a substantial portion of the total available data. In multi-institutional studies like ours, leave-one-institution-out cross-validation is crucial since it involves training the model on data from multiple institutions while leaving out one institution's data for testing, helping to identify any institution-specific biases and ensuring the model performs well across diverse clinical settings[42]. However, due to disparities in VC proportions and the number of included patients across institutions, implementing this as the primary validation strategy was infeasible. To overcome this limitation, we transferred a broad, adaptable feature space to our task by using a ViT model pre-trained on a large-scale, diverse biomedical dataset. This approach helps the model generalize better by learning from a variety of sources, reducing the likelihood of the model becoming biased toward specific data from any single institution. Furthermore, to address potential institution-specific biases, we performed two separate leave-one-institution-out validations when finalizing our image model. Despite these efforts, the higher AUC for models using augmented prediction in external validation compared to internal validation, an unusual result, suggests potential variability across institutions. Therefore, in future research, we aim to enroll a larger cohort of multi-institutional cases. This will allow us to further mitigate institutional biases and strengthen the robustness of the model through comprehensive leave-one-institution-out cross-validation. Ultimately, these efforts will support the establishment of clinical guidelines using our predictive model. Finally, our study used binary classification to predict VC based on a 50% collapse criterion, where VC was defined as a compressed anterior or central vertebral body height measuring less than 50% of the posterior height[1]. This threshold may limit the model's capacity to capture more subtle variations in vertebral collapse. Future work could focus on developing continuous or multi-class predictions that account for varying degrees of collapse, providing a more detailed understanding of patient outcomes and enabling improved risk stratification to support clinical decision-making and personalized treatment strategies.

In conclusion, our study highlights the effectiveness of using a ViT model pre-trained on a domain-specific dataset, PMC-15 M, and fine-tuned with a parameter-efficient method, LoRA, in predicting the progression of VC in OVCFs. By employing the augmented prediction strategy, we further improved the prediction performance of our model.

Data availability

The data that support the plots within this paper and other findings including the code for the models are available from the corresponding author upon reasonable request.

Received: 16 July 2024; Accepted: 10 December 2024
Published online: 30 December 2024

References

1. Sugita, M., Watanabe, N., Mikami, Y., Hase, H. & Kubo, T. Classification of vertebral compression fractures in the osteoporotic spine. *J. Spinal Disord Tech.* **18**, 376–381 (2005).
2. Johnell, O. & Kanis, J. A. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos. Int.* **17**, 1726–1733 (2006).
3. Parreira, P. C. S., Maher, C. G., Megale, R. Z., March, L. & Ferreira, M. L. An overview of clinical guidelines for the management of vertebral compression fracture: A systematic review. *Spine J.* **17**, 1932–1938 (2017).
4. Rzewuska, M., Ferreira, M., McLachlan, A. J., Machado, G. C. & Maher, C. G. The efficacy of conservative treatment of osteoporotic compression fractures on acute pain relief: a systematic review with meta-analysis. *Eur. Spine J.* **24**, 702–714 (2015).
5. Muratore, M., Ferrera, A., Masse, A. & Bistolfi, A. Osteoporotic vertebral fractures: predictive factors for conservative treatment failure. A systematic review. *Eur. Spine J.* **27**, 2565–2576 (2018).
6. Luo, J., Dolan, P., Adams, M. A., Annesley-Williams, D. J. & Wang, Y. A predictive model for creep deformation following vertebral compression fractures. *Bone* **141**, 115595 (2020).
7. Lee, S. H., Kim, E. S. & Eoh, W. Cement augmented anterior reconstruction with short posterior instrumentation: a less invasive surgical option for Kummell's disease with cord compression. *J. Clin. Neurosci.* **18**, 509–514 (2011).
8. Ito, Y., Hasegawa, Y., Toda, K. & Nakahara, S. Pathogenesis and diagnosis of delayed vertebral collapse resulting from osteoporotic spinal fracture. *Spine J.* **2**, 101–106 (2002).
9. Han, M. S., Lee, G. J., Lee, S. K., Lee, J. K. & Moon, B. J. Clinical application of bone turnover markers in treating osteoporotic vertebral compression fractures and their role in predicting fracture progression. *Medicine* **101**, e29983 (2022).
10. Jeon, I., Kim, S. W. & Yu, D. Paraspinal muscle fatty degeneration as a predictor of progressive vertebral collapse in osteoporotic vertebral compression fractures. *Spine J.* **22**, 313–320 (2022).
11. Yabu, A. et al. Using artificial intelligence to diagnose fresh osteoporotic vertebral fractures on magnetic resonance images. *Spine J.* **21**, 1652–1658 (2021).
12. Tomita, N., Cheung, Y. Y. & Hassanpour, S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput. Biol. Med.* **98**, 8–15 (2018).
13. Tustison, N. J. et al. N4ITK: improved N3 Bias correction. *IEEE Trans. Med. Imaging*. **29**, 1310–1320 (2010).
14. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of SimpleITK. *Front. Neuroinform.* **7**, 45 (2013).
15. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
16. Dosovitskiy, A. et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. in *International Conference on Learning Representations* (2021).
17. Zhang, S. et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. Preprint at (2024). https://arxiv.org/abs/2303.00915
18. Hu, E. J. et al. LoRA: Low-rank adaptation of large language models. Preprint at (2021). https://arxiv.org/abs/2106.09685
19. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
20. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32, (2019).

21. Schisterman, E. F., Perkins, N. J., Liu, A. & Bondell, H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* **16**, 73 (2005).
22. Dunn, O. J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).
23. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. in *Proceedings of the IEEE International Conference on Computer Vision* 618–626 (2017).
24. Abnar, S. & Zuidema, W. Quantifying attention flow in transformers. Preprint at (2020). https://arxiv.org/abs/2005.00928
25. Saeed, M. U., Bin, W., Sheng, J. & Mobarak Albarakati, H. An automated multi-scale feature fusion network for spine fracture segmentation using computed tomography images. *J. Digit. Imaging Inf. med.* https://doi.org/10.1007/s10278-024-01091-0 (2024).
26. Saeed, M. U. et al. An automated deep learning approach for spine segmentation and vertebrae recognition using computed tomography images. *Diagnostics* **13**, 2658 (2023).
27. Saeed, M. U., Bin, W., Sheng, J., Ali, G. & Dastgir, A. 3D MRU-Net: a novel mobile residual U-Net deep learning model for spine segmentation using computed tomography images. *Biomed. Sign. Process. Control.* **86**, 105153 (2023).
28. Saeed, M. U., Bin, W., Sheng, J., Albarakati, H. M. & Dastgir, A. MSFF: an automated multi-scale feature fusion deep learning model for spine fracture segmentation using MRI. *Biomed. Sign. Process. Control.* **91**, 105943 (2024).
29. Zhang, J. et al. Development and validation of a predictive model for vertebral fracture risk in osteoporosis patients. *Eur. Spine J.* **33**, 3242–3260 (2024).
30. Tian, J. et al. Development of a deep learning model for detecting lumbar vertebral fractures on CT images: An external validation. *Eur. J. Radiol.* **180**, 111685 (2024).
31. Wang, Y. N. et al. A deep-learning model for diagnosing fresh vertebral fractures on magnetic resonance images. *World Neurosurg.* **183**, e818–e824 (2024).
32. Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S. A. & Hospedales, T. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. in *Medical Imaging with Deep Learning* (2024).
33. Lian, C., Zhou, H. Y., Yu, Y. & Wang, L. Less could be better: Parameter-efficient fine-tuning advances medical vision foundation models. Preprint at (2024). https://doi.org/10.48550/arXiv.2401.12215
34. Park, S. et al. A novel deep learning model for a computed tomography diagnosis of coronary plaque erosion. *Sci. Rep.* **13**, 22992 (2023).
35. Sarmadi, A., Razavi, Z. S., van Wijnen, A. J. & Soltani, M. Comparative analysis of vision transformers and convolutional neural networks in osteoporosis detection from X-ray images. *Sci. Rep.* **14**, 18007 (2024).
36. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **6**, 1346–1352 (2022).
37. Xie, X. et al. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Med. Image. Anal.* **69**, 101985 (2021).
38. van Panhuis, W. G. et al. A systematic review of barriers to data sharing in public health. *BMC Public. Health.* **14**, 1144 (2014).
39. Sheller, M. J. et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
40. Sica, G. T. Bias in research studies. *Radiology* **238**, 780–789 (2006).
41. Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E. & Moons, K. G. M. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *J. Clin. Epidemiol.* **56**, 441–447 (2003).
42. Soda, P. et al. AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study. *Med. Image. Anal.* **74**, 102216 (2021).

Acknowledgements

This work was supported by the New Faculty Startup Fund from Seoul National University (Number 800-20220279).

Author contributions

Study protocol design: S.H., W.C., S.B.P.; Study supervision: W.C., S.B.P., W.T.Y.; Data collection: S.B.P., I.K., W.T.Y., C.K., Y.S.K.; Algorithm implementation: S.K., S.H., W.C.; Experiment conduct: S.K., S.H., W.C.; Result analysis: S.K., W.C.; Manuscript writing: S.K., W.C., S.B.P.; Manuscript reviewing: S.K., W.C., S.B.P., W.T.Y., Y.S.K.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-82902-w.

Correspondence and requests for materials should be addressed to W.C. or S.B.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024