Osteoporosis International
https://doi.org/10.1007/s00198-025-07727-3

ORIGINAL ARTICLE

Explainable opportunistic osteoporosis screening from chest X-rays: a retrospective comparison of foundation models

Jaewon Kim[1] · Sangmin Kwak[2] · Hyeokjong Lee[1] · Jooyoung Chang[3] · Sang Min Park[1,3]

Received: 25 May 2025 / Accepted: 9 October 2025

Abstract

Summary We evaluated foundation models for opportunistic osteoporosis screening from chest X-rays using a novel explainability framework. DINOv2 with low-rank adaptation achieved the best performance (AUC 0.93) while demonstrating clear clinical reasoning. Our findings highlight that explainability should be prioritized alongside accuracy in medical AI, enhancing trust in clinical deployment.

Purpose Deep learning models show promise for opportunistic osteoporosis screening from chest X-rays but have traditionally relied on convolutional neural networks with limited explainability. This study introduces a quantitative framework for explainability evaluation and systematically compares diverse foundation models to identify an optimal balance between performance and explainability.

Methods In this retrospective study, a retrospective dataset comprising 21,031 chest X-rays paired with bone mineral density scores from 14,502 female patients at Seoul National University Hospital was used. Twelve foundation model variants—combinations of natural- and medical-domain models fine-tuned using various strategies—were trained to classify osteoporosis status (normal, osteopenia, or osteoporosis). Foundation models were evaluated based on predictive performance (AUC, accuracy, sensitivity, and specificity) and explainability, assessed through occlusion analysis (AUC change after bone perturbation, Δ_{bone}) and saliency-map analysis (overlap between bone regions and saliency maps, IoU_{bone}).

Results DINOv2, fine-tuned with low-rank adaptation, achieved the highest predictive performance (AUC of 0.93; 95% CI, 0.92–0.94) and demonstrated robust explainability by focusing on clinically relevant bone structures, such as the spine and ribs. In osteoporosis screening from chest X-rays, statistical analysis showed that medical foundation models did not consistently outperform natural-domain models, and higher performance did not always correlate with better explainability.

Conclusion Our findings underscore the necessity of incorporating explainability as a key criterion when selecting deep learning models for opportunistic osteoporosis screening. Furthermore, the proposed framework can be readily extended to other medical tasks, fostering the development of more trustworthy and interpretable AI-assisted screening tools.

Keywords Chest X-rays · Explainability evaluation · Explainable AI · Foundation models · Osteoporosis

Introduction

With the global increase in life expectancy, osteoporosis—a systemic disease characterized by low bone mineral density (BMD) and microstructural deterioration of bone—has become a significant public health challenge due to its association with an increased risk of bone fracture [1]. Early management, such as lifestyle interventions, accurate diagnosis, and pharmacologic treatments, can mitigate fracture risk and mortality [2, 3]. While dual-energy X-ray absorptiometry (DXA) remains the gold standard for BMD measurement [4], its limited utilization often delays diagnosis and treatment, despite the existence of well-established clinical guidelines for DXA referral (e.g., USPSTF recommendations for women aged ≥ 65 years or younger women with risk factors [5]). Systemic reviews and survey studies

✉ Sang Min Park
fmpark1@snu.ac.kr

1 Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 03080, Republic of Korea

2 Department of Medicine, Seoul National University, Seoul 03080, Republic of Korea

3 XAIMED Co., Inc., Seoul 03187, Republic of Korea

have identified key barriers including limited access to DXA scanners, lack of awareness, clinical time constraints, and concerns about insurance coverage or testing costs [6, 7].

Opportunistic osteoporosis screening using deep learning models has demonstrated promising performance in predicting low BMD or osteoporosis from chest radiographs [8–17]. Chest radiographs often include visible bones such as ribs, clavicle, and vertebrae, and their morphology carries clinically relevant information about bone fragility. Experimental studies demonstrate that rib cortical thickness is strongly correlated with fracture load, suggesting its role as a structural predictor of fragility fractures [18]. Population-based radiogrammetry reports show that clavicle cortical thickness measured on chest X-rays correlates significantly with central DXA BMD [19]. Vertebral bodies remain the most common site of osteoporotic fractures and are directly evaluated by DXA [20]. Together, these findings support the plausibility that deep learning models focusing on these skeletal structures in chest X-rays may capture meaningful signals of osteoporosis risk.

Task-agnostic foundation models have become the standard in natural language processing [21, 22] and computer vision [23–25]. Transfer learning has already proven effective in medical imaging [26, 27], and studies have evaluated the generalization of natural foundation models in segmentation [28] and classification [29]. Meanwhile, medical foundation models have been developed specifically for modalities such as retinal fundus images [30], endoscopy videos [31], and chest radiographs [32, 33]. Despite these advancements, the application of foundation models to osteoporosis diagnosis remains largely unexplored. As deep learning models grow increasingly complex, their decision-making processes become less transparent, emphasizing the critical need for explainability. Thus, applying foundation models to osteoporosis diagnosis necessitates not only robust performance evaluation but also a comprehensive assessment of their explainability.

In this retrospective study, we introduce a comprehensive evaluation framework to identify the optimal foundation model for opportunistic osteoporosis screening from chest X-rays, prioritizing both predictive performance and explainability (Supplemental Fig. 1). The study leverages paired chest X-rays and BMD measurements obtained via DXA at the Health Promotion Center of Seoul National University Hospital. We evaluate both natural- and medical-domain foundation models, adapting diverse parameter-efficient fine-tuning methods. While predictive performance is typically evaluated using metrics such as AUC and accuracy, explainability is more difficult to quantify. To address this challenge, novel approaches are proposed to quantify explainability in the context of osteoporosis screening.

Methods

Study population

The dataset used for the study was collected at the Health Promotion Center of Seoul National University Hospital (HPC-SNUH). The study population included females aged 15 or older who had completed medical examinations, including chest X-rays and DXA scans, between January 2004 and December 2019. Bone mineral density (BMD) (g/cm^2) was measured using Lunar Prodigy Advance DXA (GE Lunar, Madison, WI, USA). The participants were classified into osteoporosis, osteopenia, and normal by the lowest T-score at lumbar spine 1–4 and femur neck according to the World Health Organization criteria [34]. Osteoporosis was defined by the lowest T-score of ≤ -2.5, osteopenia by $-2.5 <$ the lowest T-score < -1.0, and normal as the lowest T-score ≥ -1.0. Participants with chest X-rays and DXA scans were used to develop the deep learning model for the osteoporosis screening ($N = 14502$). We randomly split the datasets into training, validation, and test sets by proportions of 70%, 10%, and 20% at the individual level, ensuring that X-rays from the same patient remained within the same dataset. Figure 1 shows the flowchart of the configurations of the three datasets with training dataset ($N = 14824$), validation dataset ($N = 2044$), and test dataset ($N = 4163$).

Foundation models

A foundation model is a large deep learning model, typically built on transformer architectures, pre-trained on massive amounts of unlabeled data. This process produces a versatile model capable of being adapted to a wide range of downstream tasks [35]. From a self-supervised learning perspective, foundation models can be broadly categorized into text-guided self-supervised learning [24] and image-only self-supervised learning [25]. Additionally, based on the domain of the pre-training data, foundation models can be categorized as follows: natural foundation models [23–25], which are pre-trained on large-scale natural image datasets without medical-domain adaptation, and medical foundation models [30, 32, 33], which are pre-trained on medical imaging datasets to capture domain-specific representations. In this study, we comprehensively evaluated four representative foundation models spanning these categories: OpenCLIP [24], DINOv2 [25], CheXagent [32], and RAD-DINO [33], as detailed in Supplemental Table 1.

Fine-tuning methods

Supplemental Fig. 2 illustrates three fine-tuning methods used in this study to optimize the performance of the foundation

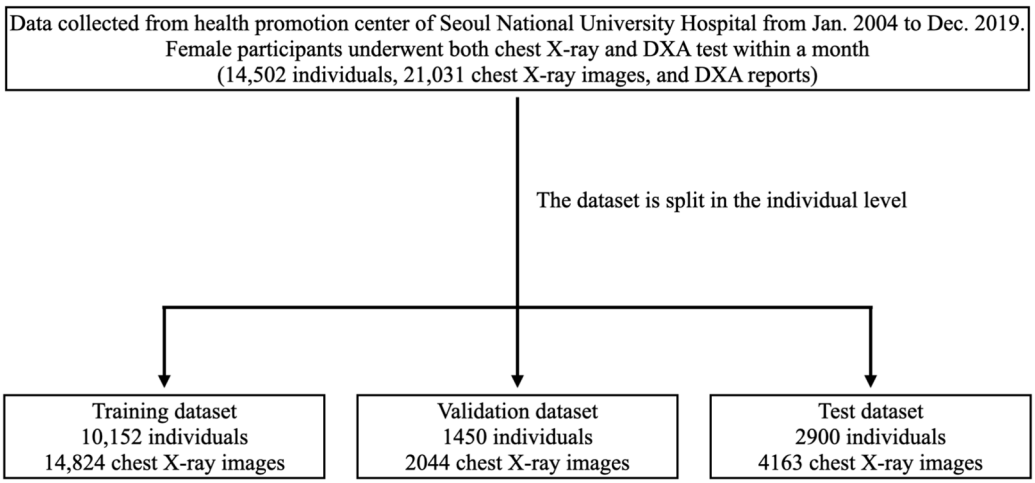

Fig. 1 Dataset configuration. DXA: dual-energy X-ray absorptiometry

models for osteoporosis screening. For linear evaluation, we kept the vision encoder frozen and fine-tuned a linear classifier standard protocol to evaluate out-of-the-box generalization capabilities of a vision encoder [25, 29, 36]. For parameter-efficient fine-tuning of the vision encoder, we explored partial fine-tuning and low-rank adaptation (LoRA) [37]. In partial fine-tuning, we trained only the last transformer block, since feature reuse is more prevalent and effective in lower layers during transfer learning to the medical domain [27]. With LoRA, decomposed low-rank metrics were injected into each transformer layer, allowing the foundation model to be fine-tuned for downstream tasks with significantly fewer parameter updates and without introducing additional inference latency [37]. Following the ablation study on different LoRA ranks (Supplemental Table 2), we selected a LoRA rank of 8 for the subsequent sections of this paper.

Quantitative evaluation of explainability

Explainability refers to the degree to which a human can understand the cause of a decision or prediction made by an AI system. In medical imaging, explainability typically involves visualizing or quantifying how model predictions relate to clinically meaningful image features, thereby allowing clinicians to assess whether the model's behavior aligns with medical rationale. Previous studies [8, 11, 17] have employed gradient-weighted class activation mapping (GradCAM) [38] to generate saliency maps that highlight class-discriminative regions. While GradCAM is widely used due to its simplicity, it remains inherently qualitative, limiting its effectiveness for comparing explainability across models. To overcome these limitations, we utilized a pretrained segmentation model [39] to extract bone regions from chest X-rays, where $bone \in \{$All bones, Spine, Sternum, Clavicles, Scapulas, Ribs$\}$. Using these extracted regions, two quantitative metrics were proposed for evaluating model explainability in osteoporosis screening.

- Occlusion analysis (Δ_{bone}): Occlusion analysis in our study evaluates how much information the model extracts from each bone region. We first masked all bones to create a baseline where only soft tissue and background are visible (i.e., No-bone AUC). Then, we selectively reintroduced a specific bone region (i.e., AUC_{bone}) and measured how much the model's prediction performance improved compared to the baseline (i.e., Δ_{bone}). A larger improvement indicates that the corresponding bone provides more useful information for the model's decision. Specifically, it is defined as follows:

$$\Delta_{bone} = \text{AUC}_{bone} - \max(\text{No-bone AUC}, 0.5) \quad (1)$$

  Thus, $\Delta_{\text{All bones}}$ equals the difference in AUC before and after masking all bones. The average Δ_{bone} across the test dataset serves as the explainability metric (see Supplemental Fig. 3 for visual examples).
- Saliency-map analysis (IoU_{bone}): Saliency-map analysis quantitatively evaluates how well the model's highlighted regions correspond to the actual bone structures relevant for osteoporosis. Using GradCAM, we generated saliency maps indicating regions most influential to the model's decision. These GradCAM outputs were then binarized using the median value and compared with the segmented bone regions to compute the intersection over union (IoU). A higher IoU indicates that the model's focus was more precisely aligned with the bone regions. Formally, IoU_{bone} was calculated as follows:

$$\text{IoU}_{bone} = \frac{\text{Intersection of } bone \text{ and GradCAM}}{\text{Union of } bone \text{ and GradCAM}} \quad (2)$$

The mean IoU_{bone} across the test dataset was used as an additional explainability metric (see Supplemental Fig. 4 for visual examples).

Statistical analysis

To evaluate the performance of foundation models in osteoporosis screening, the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity were calculated with 95% confidence interval (CI). To compare foundation models in terms of performance and explainability, we applied the Friedman test separately to three evaluation metrics: predictive performance (AUC), occlusion analysis ($\Delta_{\text{All bones}}$), and saliency-map analysis ($\text{IoU}_{\text{All bones}}$). The Friedman test assessed whether there were statistically significant differences among models for each metric, and the Nemenyi post hoc test was subsequently performed to determine which pairs of models differed significantly. In addition, Kendall's tau test was used to examine the association between predictive performance and explainability. Specifically, we calculated rank correlations between AUC and $\Delta_{\text{All bones}}$ and between AUC and $\text{IoU}_{\text{All bones}}$, to get further insight into the relationship between performance and explainability. We have used Scikit-Learn, Scipy, and scikit-posthocs packages in Python to conduct statistical analysis.

Implementation details

In this study, we trained foundation models for a multi-class classification task to predict normal, osteopenia, and osteoporosis cases, using cross-entropy loss with label smoothing [40]. The training dataset was used to optimize model parameters for this classification task. The validation dataset was used to evaluate predictive performance during training and to select the model with the highest AUC. The test dataset was then used for final evaluation, including both predictive performance and explainability analyses (occlusion analysis and saliency-map analysis). All results reported in this study are based on the test dataset. Similar to [41], training was conducted in two stages for efficiency: images were initially resized to 224 × 224 and trained for 80 epochs, and then fine-tuned at a higher resolution of 448 × 448 for 20 epochs. Validation and test images were consistently resized to 448 × 448 to prevent overfitting and match the final evaluation resolution. For linear evaluation, we used stochastic gradient descent (SGD) with a step learning rate decay schedule, while for partial fine-tuning and LoRA, we applied the AdamW optimizer with a cosine annealing learning rate schedule and a warm-up phase. All experiments were conducted on NVIDIA A100 GPUs with 80GB memory.

Ethics statement

This study was approved by the institutional review board at SNUH (IRB: H-2003-205-1113) and followed the Declaration of Helsinki of 1975. All data were anonymized, and the Institutional Review Board waived the requirement for written informed consent.

Results

Patients' characteristics

Our dataset consisted of 21,031 chest X-rays from 14,502 female individuals, with ages ranging from 20 to 88 years (mean age, 55.97 years). Of all chest X-rays, 12,038 showed normal results, 7936 indicated osteopenia, and 1057 indicated osteoporosis based on DXA examination within a month. Baseline characteristics of the training, validation, and test datasets are summarized in Table 1.

Performance of foundation models

The predictive performance of foundation models for osteoporosis classification (osteoporosis vs. non-osteoporosis) is summarized in Table 2. Among all models, DINOv2/LoRA achieved the highest AUC (0.93; 95% CI, 0.92–0.94), with a sensitivity of 84.36% (95% CI, 83.26–85.46) and a specificity of 87.55% (95% CI, 86.55–88.55). This was followed by DINOv2/partial fine-tuning (AUC = 0.91; 95% CI, 0.90–0.92) and OpenCLIP fine-tuned models (AUC = 0.91; 95% CI, 0.90–0.92). Notably, RAD-DINO/LoRA, despite being based on a ViT-B architecture with fewer parameters, achieved an AUC of 0.90 (95% CI, 0.89–0.91), demonstrating competitive performance relative to larger ViT-G-based models.

A Friedman test on predictive performance (AUC) revealed a statistically significant difference among the models ($\chi^2 = 472.74$, $p < 0.001$). Post hoc Nemenyi comparisons (Supplemental Fig. 5) further indicated that among the highest-performing models, DINOv2/LoRA and OpenCLIP fine-tuned models did not differ significantly ($p > 0.05$), suggesting that these architectures achieve similar predictive performance when fine-tuned effectively.

Across all architectures, LoRA and partial fine-tuning consistently outperformed linear evaluation, with AUCs ranging from 0.85 to 0.93 compared to 0.72 to 0.77 for linear models. Linear evaluation resulted in significantly lower AUCs and accuracy compared to all fine-tuned methods ($p < 0.05$), reinforcing the importance of fine-tuning strategies in optimizing model performance. Moreover, within the linear evaluation paradigm, no significant performance differences

Table 1 Demographic characteristics of the datasets

	Training set	Validation set	Test set	Overall
Participant	10,152	1450	2900	14,502
Chest radiograph images	14,824	2044	4163	21,031
Age (years), $\mu \pm \sigma$	55.97 ± 9.73	55.81 ± 9.81	56.06 ± 9.51	55.97 ± 9.69
BMD (g/cm^2), $\mu \pm \sigma$				
L_1-L_4	1.08 ± 0.17	1.08 ± 0.17	1.08 ± 0.17	1.08 ± 0.17
Femur neck	0.85 ± 0.12	0.85 ± 0.13	0.85 ± 0.13	0.85 ± 0.12
T-score, $\mu \pm \sigma$				
L_1-L_4	−0.36 ± 1.35	−0.33 ± 1.35	−0.34 ± 1.37	−0.35 ± 1.36
Femur neck	−0.43 ± 1.03	−0.45 ± 1.04	−0.43 ± 1.04	−0.43 ± 1.03
T-score categories, n (%)				
Normal	8487 (57.25%)	1140 (55.77%)	2411 (57.91%)	12,038 (57.24%)
Osteopenia	5597 (37.76%)	798 (39.04%)	1541 (37.02%)	7936 (37.73%)
Osteoporosis	740 (4.99%)	106 (5.19%)	211 (5.07%)	1057 (5.03%)

BMD, bone mineral density; μ, mean; σ, standard deviation

were observed among the foundation models ($p > 0.05$), indicating that their out-of-the-box feature representations for osteoporosis screening are comparable. Detailed results for the normal and osteopenia classes are provided in Supplemental Table 3 and Supplemental Table 4.

Explainability of foundation models

Occlusion analysis

Occlusion analysis (Fig. 2A) quantifies the contribution of different bone structures to model predictions by measuring the increase in AUC when each region is revealed. The Friedman test confirmed significant differences in occlusion scores among models ($\chi^2 = 434.22$, $p < 0.001$), indicating that reliance on bone regions varies across architectures and fine-tuning strategies.

DINOv2/LoRA and DINOv2/partial fine-tuning exhibited the highest $\Delta_{\text{All bones}}$ (0.41), followed closely by OpenCLIP/LoRA (0.40), suggesting that these models strongly depend on overall bone structures for decision-making. In contrast, linear models (DINOv2/linear, OpenCLIP/linear, CheXagent/linear, and RAD-DINO/linear) demonstrated significantly lower $\Delta_{\text{All bones}}$ ($p < 0.05$), ranging from 0.12 to 0.22, indicating weaker reliance on key anatomical regions. Post hoc Nemenyi comparisons (Supplemental Fig. 6A) further confirmed that DINOv2/LoRA, DINOv2/partial fine-tuning, and OpenCLIP/LoRA were significantly different

Table 2 The osteoporosis predictive performance metrics of foundation models

Model	Method	AUC (95% CI)	Accuracy (%) (95% CI)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)
OpenCLIP	Linear	0.73 (0.72–0.74)	62.62 (61.15–64.09)	74.41 (73.08–75.73)	61.99 (60.52–63.47)
	Partial ft	0.91 (0.90–0.92)	82.30 (81.14–83.46)	83.89 (82.77–85.00)	82.21 (81.05–83.37)
	LoRA	0.91 (0.90–0.92)	79.01 (77.77–80.24)	88.63 (87.66–89.59)	78.49 (77.24–79.74)
DINOv2	Linear	0.77 (0.76–0.79)	66.66 (65.23–68.09)	76.78 (75.49–78.06)	66.12 (64.68–67.56)
	Partial ft	0.91 (0.90–0.92)	81.60 (80.42–82.78)	85.31 (84.23–86.38)	81.40 (80.22–82.58)
	LoRA	0.93 (0.92–0.94)	87.39 (86.38–88.40)	84.36 (83.26–85.46)	87.55 (86.55–88.55)
CheXagent	Linear	0.72 (0.71–0.73)	66.92 (65.49–68.35)	68.25 (66.83–69.66)	66.85 (65.42–68.28)
	Partial ft	0.90 (0.89–0.91)	82.70 (81.5683.85)	82.46 (81.31–83.62)	82.72 (81.57–83.87)
	LoRA	0.85 (0.84–0.86)	76.24 (74.95–77.54)	79.62 (78.40–80.84)	76.06 (74.77–77.36)
RAD-DINO	Linear	0.72 (0.70–0.73)	66.99 (65.57–68.42)	66.82 (65.39–68.25)	67.00 (65.58–68.43)
	Partial ft	0.89 (0.88–0.90)	85.73 (84.67–86.79)	76.30 (75.01–77.60)	86.23 (85.19–87.28)
	LoRA	0.90 (0.89–0.91)	86.69 (85.66–87.72)	76.78 (75.49–78.06)	87.22 (86.21–88.24)

AUC, area under the receiver operating characteristic curve; CI, confidence interval; ft, fine-tuning; $LoRA$, low-rank adaptation

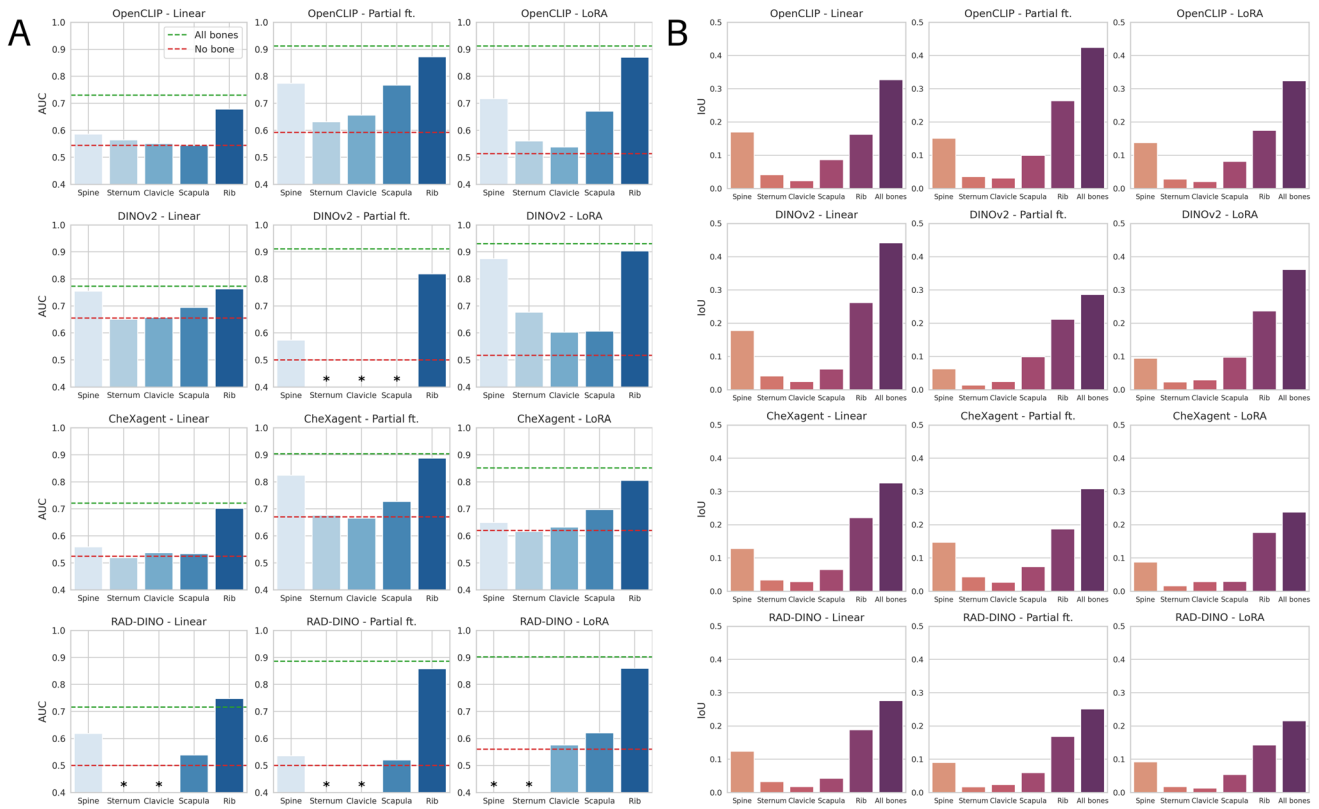

Fig. 2 Quantitative explainability evaluation of foundation models for osteoporosis screening. A Occlusion analysis (Δ_{bone}), which evaluates how much information the model extracts from bone regions by measuring the performance gain when specific bones are revealed. **B** Saliency-map analysis (IoU_{bone}), which quantifies how well the model's highlighted regions align with the actual bone structures. *AUC < 0.5 indicates performance worse than random. No-bone AUC < 0.5 is adjusted to 0.5 to align with Eq. 1. AUC, area under the receiver operating characteristic curve; IoU, intersection over union; ft., fine-tuning; LoRA, low-rank adaptation

from most other models ($p < 0.001$). Additionally, in most models, including DINOv2/LoRA, the higher scores for Δ_{Rib} and Δ_{Spine} compared to other specific bone regions suggest that the ribs and spine are the primary contributors to osteoporosis diagnosis.

Saliency-map analysis

Saliency-map analysis (Fig. 2B) evaluates the alignment between model-identified regions and actual bone structures. The Friedman test confirmed statistically significant differences in saliency-map agreement across models ($\chi^2 = 20810.46$, $p < 0.001$), indicating that explainability varies based on both architecture and fine-tuning approach.

DINOv2/linear demonstrated the highest alignment with bone structures ($\text{IoU}_{\text{All bones}} = 0.44$), followed by OpenCLIP/partial fine-tuning (0.42) and DINOv2/LoRA (0.36). Post hoc Nemenyi comparison (Supplemental Fig. 6B) further confirmed significant differences in explainability across models and fine-tuning strategies ($p < 0.001$). In contrast, RAD-DINO/LoRA exhibited the lowest overall saliency-map alignment (0.22), despite its high accuracy, suggesting that it focuses less on critical bone regions. Additionally, across most models, IoU_{Rib} and $\text{IoU}_{\text{Spine}}$ scores were consistently higher than those of other specific bone regions, aligning with occlusion analysis findings. This reinforces that the ribs and spine are the primary contributors to osteoporosis diagnosis.

Qualitative analysis

Figure 3 presents the averaged GradCAMs for each foundation model, providing a qualitative visualization of the regions that contribute most to the model's predictions. Among the models with the highest alignment to bone structures, DINOv2/linear primarily focuses on the spine and ribs (Fig. 3D), OpenCLIP/partial fine-tuning highlights the lateral ribs, scapula, and heart (Fig. 3B), and DINOv2/LoRA directs attention to the scapula, clavicle, and heart (Fig. 3F). In contrast, RAD-DINO/LoRA, the model with the lowest

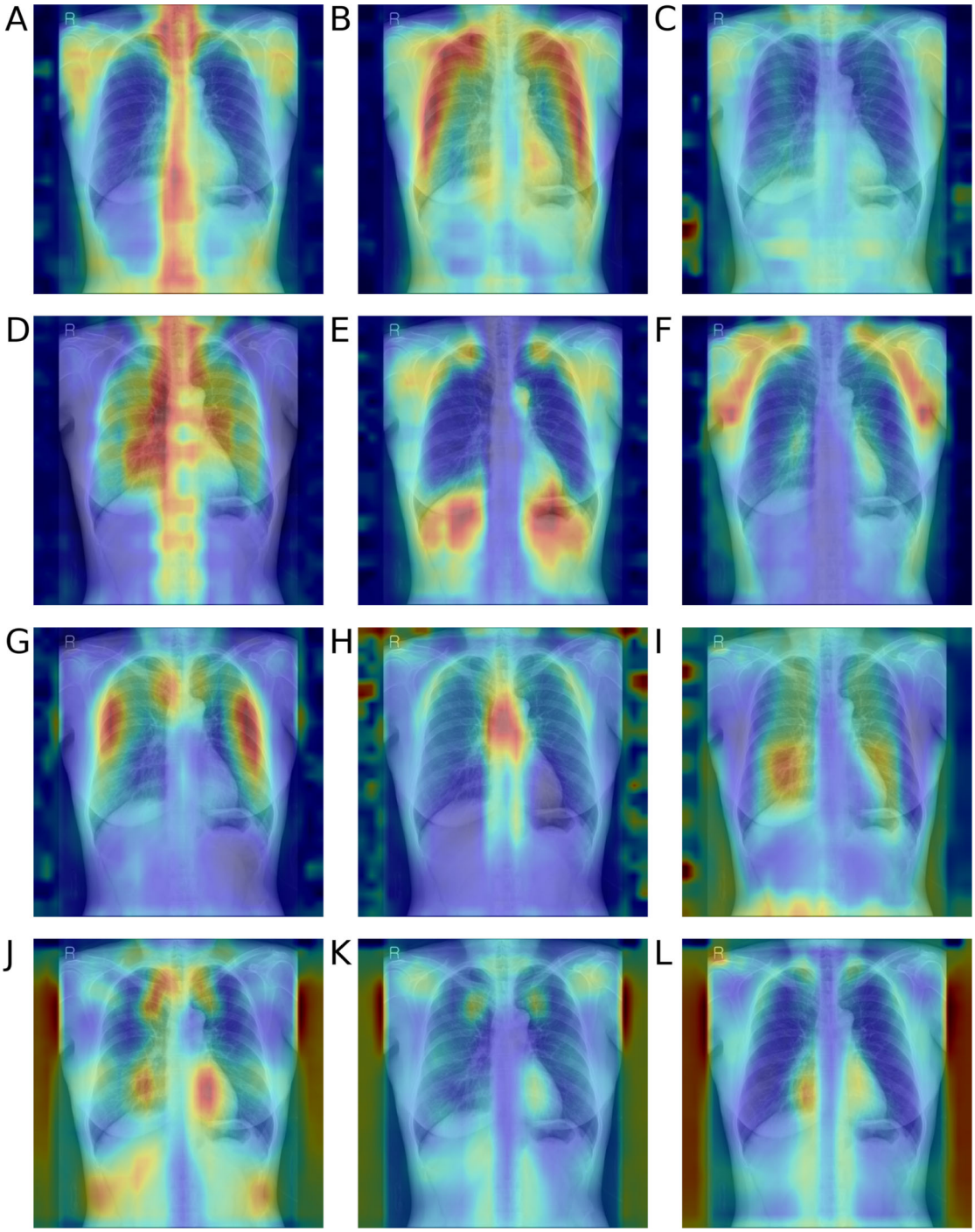

Fig. 3 Averaged GradCAMs for each ground truth label in the test dataset, overlaid with a chest X-ray. GradCAM highlights image regions most influential for the model's prediction. **A** OpenCLIP/Linear. **B** OpenCLIP/Partial ft. **C** OpenCLIP/LoRA. **D** DINOv2/Linear. **E** DINOv2/Partial ft. **F** DINOv2/LoRA. **G** CheXagent/Linear. **H** CheXagent/Partial ft. **I** CheXagent/LoRA. **J** RAD-DINO/Linear. **K** RAD-DINO/Partial ft. **L** RAD-DINO/LoRA. The color map follows the jet scale, where red indicates high importance and blue represents low importance. Figure best viewed in color. GradCAM, gradient-weighted class activation mapping; ft., fine-tuning; LoRA, low-rank adaptation

saliency-map alignment, predominantly focuses on soft tissues and regions outside the anatomical area of interest (Fig. 3L), indicating reduced reliance on relevant skeletal structures. The result suggests that saliency-map analysis ($IoU_{All\ bones}$) serves as an effective surrogate quantification method for conventional saliency-map interpretation.

Relationship between performance and explainability

To assess the relationship between model performance and explainability, Kendall's tau correlation was computed between AUC and both occlusion analysis ($\Delta_{\text{All bones}}$) and saliency-map analysis ($\text{IoU}_{\text{All bones}}$). A moderate positive correlation was observed between performance and occlusion analysis ($\tau = 0.58$, $p = 0.01$), indicating that models with higher performance tend to rely more strongly on clinically relevant bone structures during the decision-making process. In contrast, no significant correlation was found between performance and saliency-map analysis ($\tau = 0.18$, $p = 0.46$), suggesting that higher performance does not necessarily translate to better visual alignment with bone structures in saliency maps.

Discussion

Using 14,502 chest X-ray images paired with DXA measurements from HPC-SNUH, this study is the first, to our knowledge, to apply foundation models to osteoporosis diagnosis and compare them in terms of both performance and explainability. In medical AI applications, high predictive accuracy is the primary criterion for clinical deployment, especially when the model's outputs are validated across populations. Explainability, while not always essential for clinical decision-making, remains valuable as a complementary tool to enhance clinical trust, facilitate model auditing, and generate new scientific hypotheses. Among the evaluated models, DINOv2/LoRA achieved the highest predictive performance (AUC = 0.93), with OpenCLIP/partial fine-tuning and OpenCLIP/LoRA exhibiting similar performance. In occlusion analysis, DINOv2/LoRA demonstrated the highest $\Delta_{\text{All bones}}$ score and significantly outperformed OpenCLIP/partial fine-tuning. In saliency-map analysis, while ranking among the top three in $\text{IoU}_{\text{All bones}}$, DINOv2/LoRA showed a statistically significant advantage over OpenCLIP/LoRA. These results suggest that DINOv2/LoRA could be identified as the optimal model that balances between performance and explainability for osteoporosis screening.

The variability in model performance and explainability can be examined from two perspectives: pre-training data domain and fine-tuning methods. From the pre-training data domain perspective, it is often assumed that models trained on natural image datasets struggle with the complexities of medical data, which has motivated the development of medical foundation models [42]. However, our findings challenge this assumption in the context of opportunistic osteoporosis screening. Despite medical foundation models such as CheXagent and RAD-DINO having shown promise in diagnosing cardiopulmonary diseases [32, 33], no significant differences were observed among foundation models under linear evaluation. This indicates that out-of-the-box representations from natural-domain foundation models are well-suited for osteoporosis screening, eliminating the necessity of medical foundation models in this specific task. From the fine-tuning perspective, LoRA and partial fine-tuning consistently outperformed linear evaluation across all foundation models, underscoring the insufficiency of out-of-the-box representations for osteoporosis screening and the necessity of further fine-tuning. With appropriate fine-tuning strategies, medical foundation models did not exhibit advantages in either performance or explainability. DINOv2 and OpenCLIP not only achieved strong predictive performance but also provided more clinically relevant insights, as evidenced by better GradCAM alignment with bone regions and stronger reliance on bone structures for decision-making. These findings suggest that fine-tuning natural-domain foundation models may be sufficient for certain medical tasks, reducing the need for costly data collection and pre-training of specialized medical foundation models.

An important insight from this study is that strong predictive performance does not necessarily imply high explainability. For instance, mid-tier models such as CheXagent/partial fine-tuning and RAD-DINO/LoRA achieved competitive predictive performance (both AUC = 0.90, ranking fifth and sixth, respectively), yet their explainability metrics were weaker (ranking seventh and fifth in occlusion analysis, and seventh and twelfth in saliency-map analysis, respectively). Conversely, strong explainability does not always translate to high predictive accuracy. DINOv2/linear exhibited the highest $\text{IoU}_{\text{All bones}}$ score and primarily focused on the spine and ribs, yet its predictive performance was suboptimal (AUC = 0.77). Statistical analysis further supports this finding. While AUC and $\Delta_{\text{All bones}}$ were positively correlated, no significant correlation was observed between AUC and $\text{IoU}_{\text{All bones}}$. These results emphasize the need for a comprehensive evaluation of both predictive performance and explainability when developing AI models for medical applications, as optimizing one does not guarantee improvements in the other.

Quantitative evaluation of explainability identified the ribs and spine as the most influential bone regions for osteoporosis classification (Fig. 2). These findings are clinically relevant, as fractures in the spine and ribs are strongly associated with osteoporosis [43]. Evaluating the decision-making process of AI models is essential for auditing their alignment with clinical reasoning and ethical standards [44]. For example, some models may inadvertently focus on the waist region, as studies have reported correlations between waist

circumference and osteoporosis [45, 46]. Such unintended biases highlight the importance of explainability assessment to ensure the clinical validity of AI models before deployment in real-world settings.

This study has several limitations. First, validation was conducted using data from a single institution, which may introduce population bias. Additionally, restricting the dataset to female patients minimized gender-related biases but resulted in an age distribution skewed toward older individuals, as younger women are less likely to undergo DXA scans due to cost barriers. Consequently, the generalization of these findings to broader populations remains uncertain. Future research should examine the impact of explainability on model generalization across external datasets. Furthermore, the WHO classification of osteoporosis, osteopenia, and normal is increasingly viewed as insufficient for guiding treatment decisions, and recent guidelines emphasize fracture risk prediction using tools such as FRAX [47–49]. However, BMD remains an essential component of fracture risk assessment and can serve as a useful surrogate in opportunistic screening settings where outcome data or comprehensive clinical information is lacking. Our approach aims to bridge this gap by offering a scalable, low-cost method to detect BMD-defined osteoporosis from widely available chest X-rays. Future work could incorporate clinical variables and longitudinal fracture outcomes to build more comprehensive risk prediction models. In such settings, explainability can clarify whether predictions are based on known risk factors or unexpected signals. Our current explainability assessment, focused on bone structures, serves as a surrogate and may need to be expanded to better capture the model's decision logic in future multimodal applications. Finally, while the quantitative explainability metrics used in this study effectively enable relative comparisons, their interpretation as absolute values warrants further investigation, potentially through clinician surveys.

Conclusion

This study utilized pre-trained bone segmentation models to quantitatively assess explainability through occlusion and saliency-map analysis. By comparing natural and medical foundation models across various fine-tuning methods, our results identify DINOv2/LoRA as the optimal model, demonstrating strong predictive performance while maintaining a clear focus on overall bone structures for osteoporosis prediction. Moreover, our framework, particularly the approach for quantifying explainability, is adaptable to medical imaging tasks with well-defined anatomical structures, enabling broader applications of deep learning in clinical practice. As current AI advancements often prioritize performance over explainability, our findings emphasize the importance of integrating both aspects to enhance the transparency and reliability of AI-driven medical diagnostics.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00198-025-07727-3.

Funding Open Access funding enabled and organized by Seoul National University. This research was supported by Seoul National University Hospital, Seoul, Republic of Korea (grant no. 04-2024-0790)

Data availability Data that support the findings of this study are available on request from the corresponding author.

Declarations

Conflicts of interest None.

References

1. Consensus N (2000) Osteoporosis prevention, diagnosis, and therapy. NIH Consens Statement 17:1–45
2. Gregson CL, Armstrong DJ, Bowden J, Cooper C, Edwards J, Gittoes NJL, et al (2022) UK clinical guideline for the prevention and treatment of osteoporosis. Arch Osteopor. 17(1)
3. Bolland MJ, Grey AB, Gamble GD, Reid IR (2010) Effect of osteoporosis treatment on mortality: a meta-analysis. J Clinic Endocrinol Metabol 95(3):1174–1181
4. Force* UPST (2011) Screening for osteoporosis: US preventive services task force recommendation statement. Annal Internal Med 154(5):356–364
5. Nicholson WK, Silverstein M, Wong JB, Chelmow D, Coker TR, Davis EM et al (2025) Screening for osteoporosis to prevent fractures: US preventive services task force recommendation statement. JAMA 333(6):498–508
6. Onizuka N, Onizuka T (2024) Disparities in osteoporosis prevention and care: understanding gender, racial, and ethnic dynamics. Curr Rev Musculoskelet Med 17(9):365–372
7. Choksi P, Gay BL, Haymart MR, Papaleontiou M (2023) Physician-reported barriers to osteoporosis screening: a nationwide survey. Endocr Pract 29(8):606–611
8. Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N (2020) Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset. J Bone Miner Res 37(2):369–377
9. Sato Y, Yamamoto N, Inagaki N, Iesaki Y, Asamoto T, Suzuki T et al (2022) Deep learning for bone mineral density and t-score

prediction from chest X-rays: a multicenter study. Biomedicines 10(9):2323
10. Wang F, Zheng K, Lu L, Xiao J, Wu M, Kuo CF et al (2022) Lumbar bone mineral density estimation from chest X-ray images: anatomy-aware attentive multi-ROI modeling. IEEE Trans Med Imaging 42(1):257–267
11. Asamoto T, Takegami Y, Sato Y, Takahara S, Yamamoto N, Inagaki N et al (2024) External validation of a deep learning model for predicting bone mineral density on chest radiographs. Arch Osteoporos 19(1):1–10
12. Tsai DJ, Lin C, Lin CS, Lee CC, Wang CH, Fang WH (2024) Artificial intelligence-enabled chest X-ray classifies osteoporosis and identifies mortality risk. J Med Syst 48(1):12
13. Yamamoto N, Shiroshita A, Kimura R, Kamo T, Ogihara H, Tsuge T (2024) Diagnostic accuracy of chest X-ray and CT using artificial intelligence for osteoporosis: systematic review and meta-analysis. J Bone Mineral Metabol 42(5):483–491
14. Bilbily A, Syme CA, Adachi JD, Berger C, Morin SN, Goltzman D, et al (2024) Opportunistic screening of low bone mineral density from standard X-rays. J Am College Rad 21(4):633–639
15. Tsai DJ, Lin C, Lin CS, Lee CC, Wang CH, Fang WH (2024) Artificial intelligence-enabled chest X-ray classifies osteoporosis and identifies mortality risk. J Med Syst 48(1)
16. Tseng SC, Lien CE, Lee CH, Tu KC, Lin CH, Hsiao AY, et al (2024) Clinical validation of a deep learning-based software for lumbar bone mineral density and T-score prediction from chest X-ray images. Diagnostics 14(12)
17. Lin C, Tsai DJ, Wang CC, Chao YP, Huang JW, Lin CS, et al (2024) Osteoporotic precise screening using chest radiography and artificial neural network: the OPSCAN randomized controlled trial. Rad 311(3)
18. Liebsch C, Hübner S, Palanca M, Cristofolini L, Wilke HJ (2021) Experimental study exploring the factors that promote rib fragility in the elderly. Sci Rep 11(1):9307
19. Gulam M, Thornton MM, Hodsman AB, Holdsworth DW (2000) Bone mineral measurement of phalanges: comparison of radiographic absorptiometry and area dual X-ray absorptiometry. Radiology 216(2):586–591
20. Cummings SR, Melton LJ (2002) Epidemiology and outcomes of osteoporotic fractures. The Lancet 359(9319):1761–1767
21. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
22. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al (2023) Llama: open and efficient foundation language models. arXiv:2302.13971
23. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al (2023) Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; p 4015–4026
24. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR p 8748–8763
25. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al (2023) DINOv2: learning robust visual features without supervision. arXiv:2304.07193
26. Matsoukas C, Haslum JF, Söderberg M, Smith K (2023) Pretrained ViTs yield versatile representations for medical images. arXiv:2303.07034
27. Matsoukas C, Haslum JF, Sorkhei M, Söderberg M, Smith K (2022) What makes transfer learning work for medical images: feature reuse & other factors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; p 9225–9234
28. Deng R, Cui C, Liu Q, Yao T, Remedios LW, Bao S, et al (2023) Segment anything model (SAM) for digital pathology: assess zero-shot segmentation on whole slide imaging. arXiv:2304.04155
29. Huix JP, Ganeshan AR, Haslum JF, Söderberg M, Matsoukas C, Smith K (2024) Are natural domain foundation models useful for medical image classification? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; p 7634–7643
30. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR et al (2023) A foundation model for generalizable disease detection from retinal images. Nature 622(7981):156–163
31. Wang Z, Liu C, Zhang S, Dou Q (2023) Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; p 101–111
32. Chen Z, Varma M, Delbrouck JB, Paschali M, Blankemeier L, Van Veen D, et al (2024) Chexagent: towards a foundation model for chest x-ray interpretation. arXiv:2401.12208
33. Pérez-García F, Sharma H, Bond-Taylor S, Bouzid K, Salvatelli V, Ilse M, et al (2024) RAD-DINO: exploring scalable medical image encoders beyond text supervision. arXiv:2401.10815
34. Dimai HP (2017) Use of dual-energy X-ray absorptiometry (DXA) for diagnosis and fracture risk assessment; who-criteria, T-and Z-score, and reference databases. Bone 104:39–43
35. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al (2021) On the opportunities and risks of foundation models. arXiv:2108.07258
36. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; p 9729–9738
37. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al (2021) Lora: Low-rank adaptation of large language models. arXiv:2106.09685
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; p 618–626
39. Seibold C, Jaus A, Fink MA, Kim M, Reiß S, Herrmann K, et al (2023) Accurate Fine-Grained Segmentation of Human Anatomy in Radiographs via Volumetric Pseudo-Labeling. Available from: arXiv:2306.03934
40. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p 2818–2826
41. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. PMLR. p 10347–10357
42. Khan W, Leem S, See KB, Wong JK, Zhang S, Fang R (2024) A Comprehensive Survey of Foundation Models in Medicine. arXiv:2406.10729
43. Warriner AH, Patkar NM, Curtis JR, Delzell E, Gary L, Kilgore M et al (2011) Which fractures are most attributable to osteoporosis? J Clin Epidemiol 64(1):46–53
44. Mökander J (2023) Auditing of ai: legal, ethical and technical approaches. Digital Society 2(3):49
45. Pan R, Wang R, Zhang Y, Ji H, Liang X, Zhao Y (2024) The association of waist circumference with bone mineral density and risk of osteoporosis in US adult: National health and nutrition examination survey. BONE. 185
46. Cui LH, Shin MH, Kweon SS, Choi JS, Rhee JA, Lee YH et al (2014) Sex-related differences in the association between waist circumference and bone mineral density in a korean population. BMC Musculoskelet Disord 15:1–8
47. Kanis JA, Johnell O, Odén A, Johansson H, McCloskey E (2008) Frax™ and the assessment of fracture probability in men and women from the uk. Osteoporos Int 19(4):385–397

48. Kanis JA, Oden A, Johansson H, Borgström F, Ström O, McCloskey E (2009) Frax® and its applications to clinical practice. Bone 44(5):734–743
49. Mitchell P (2011) Fracture liaison services: the uk experience. Osteoporos Int 22(Suppl 3):487