scientific reports

OPEN

Osteoporosis prediction from hand X-ray images using segmentation-for-classification and self-supervised learning

Ung Hwang[1], Chang-Hun Lee[2] & Kijung Yoon[1,3]✉

Osteoporosis is a prevalent metabolic bone disease that frequently remains undiagnosed due to limited access to bone mineral density (BMD) tests, such as Dual-energy X-ray absorptiometry (DXA). To address this issue, recent research explores alternative indicators from peripheral skeletal sites to enable earlier and more accessible screening. In this paper, we propose a method to predict osteoporosis using hand and wrist X-ray images, which are widely available and cost-effective, though their association with DXA-based diagnoses is not yet fully established. Our approach employs an image segmentation model utilizing a mixture of probabilistic U-Net decoders, which captures predictive uncertainty when segmenting the ulna, radius, and metacarpal bones. The segmentation task is formulated as an optimal transport (OT) problem, effectively addressing the variability inherent in medical images. Additionally, we adopt a self-supervised learning (SSL) strategy that pretrains the model on augmented, unlabeled data to learn robust, invariant feature representations. These features are subsequently fine-tuned in a supervised classification task to distinguish osteoporotic from normal cases. We evaluate our method on X-rays from 192 individuals with verified DXA diagnoses. By combining uncertainty-aware segmentation and self-supervised feature learning, our framework offers a promising vision-based strategy for early osteoporosis detection using peripheral X-ray imaging.

Osteoporosis is a common bone ailment characterized by reduced bone mineral density (BMD) or bone mass loss, leading to bones becoming fracture-prone and structurally compromised. Given its prevalence and far-reaching impact, there is a pressing need for preemptive risk assessment, early diagnosis, and effective preventive actions. While computed tomography (CT) and magnetic resonance imaging (MRI) have demonstrated potential in BMD estimation and osteoporosis screening[1–3], their clinical use is limited due to concerns about radiation exposure and associated costs. At present, the dual energy X-ray absorptiometry (DXA) is recognized as a standard and reliable instrument for osteoporosis detection and BMD analysis. Nonetheless, DXA comes with its own set of challenges: the proficiency of the operator can influence the results, the patient's posture during the test can skew accuracy, and in obese patients with significant fat mass, BMD might be overestimated[4,5].

As a solution, recent studies have highlighted the utility of X-ray images, routinely acquired in clinical settings, as an effective means for gathering comprehensive data. This approach can be particularly useful for osteoporosis detection[6–10], leveraging the widespread availability of radiographs from individuals who have not been specifically screened for the condition. By eliminating the need for expensive DXA equipment, this strategy presents a more affordable option and is preferable for patients due to its considerably lower radiation exposure compared to DXA scans. Among the various clinical approaches, the 2nd metacarpal cortical index (2MCI) in hand X-ray images has gained attention as a promising biomarker for osteoporosis screening[11,12], with the cortical thickness and cancellous bone tissue porosity of the metacarpal bone playing a crucial role in disease identification. Despite its promise, this method faces obstacles, such as the time-intensive manual measurement of metacarpal dimensions and its failure to encompass all disease characteristics. A viable alternative is automated radiographic understanding and analysis, which can offer valuable insights to orthopedists during the diagnostic process.

To address the research gap between traditional manual radiographic analysis and the recent strides in deep learning that have reached diagnostic levels comparable to those of medical experts in various fields[13], we introduce a novel automated approach for osteoporosis screening (Fig. 1). In these procedures, we innovate

[1]Department of Electronic Engineering, Hanyang University, Seoul 04763, Republic of Korea. [2]Department of Orthopedic Surgery, Hanyang University Medical Center, Seoul 04763, Republic of Korea. [3]Department of Artificial Intelligence, Hanyang University, Seoul 04763, Republic of Korea. ✉email: kiyoon@hanyang.ac.kr

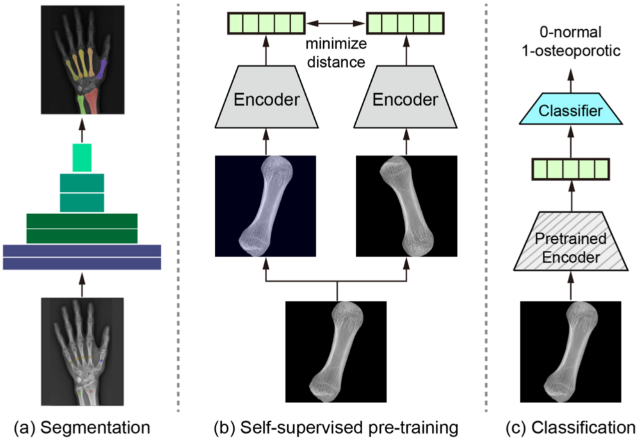

Fig. 1. Illustration of framing a hand X-ray osteoporosis screening problem through (**a**) segmentation, (**b**) self-supervised pre-training, and (**c**) downstream classification task.

by redefining the task of medical image classification as a segmentation (Fig. 1a), thereby obviating the need for manual calculations of the 2MCI. The first step in our approach involves curating the training dataset by identifying and isolating critical bones in hand and wrist radiographs, with a particular focus on the metacarpals for osteoporosis screening. This task is accomplished through a dedicated U-Net architecture[14–16] designed to effectively captures the multimodal aspects of predictive uncertainty in semantic segmentation. It is akin to the routine practices of orthopedists who depend on quantitative assessments derived from segmentation masks. Following the segmentation-based classification reformulation, our approach employs a self-supervised learning (SSL) framework (Fig. 1b) for the initial pre-training phase of our classification networks (Fig. 1c). This framework enables the models to discern nuanced, yet critical features without the dependency on explicitly labeled data, thereby improving the models' generalization capabilities, especially in the context of limited data availability, as demonstrated in our study with 192 subjects. Furthermore, we enhance the efficiency of our self-supervised learning by incorporating a customized multi-crop augmentation strategy[17], specifically adapted for hand X-ray images, to optimize the learning process.

This paper explores how easily accessible, yet unlabeled, hand radiographic data can be leveraged to detect osteoporosis on real-world medical settings. Our contributions are threefold:

- We introduce a novel automated approach for osteoporosis screening that utilizes peripheral radiographic images instead of central DXA measurements, leveraging a custom U-Net architecture to eliminate the need for traditional 2MCI calculations. This shift could significantly simplify and enhance the accessibility of osteoporosis screenings.
- We explore a SSL framework focused on segments of hand and wrist bones. This method allows our models to effectively distill critical features prior to osteoporosis prediction, demonstrating enhanced outcomes compared to conventional supervised learning approaches.
- Our research boosts model training efficiency through a specialized multi-crop augmentation strategy specifically designed for hand X-ray images, optimizing the learning process and improving model performance in scenarios with limited data availability.

Related work

Osteoporosis Prediction in Radiographic Data. Classical approaches to osteoporosis prediction have relied on manually derived radiographic measurements from either proximal or distal segments of long bones to estimate BMD and assess osteoporosis risk. Examples include indices such as cortical thickness and canal flare, as well as measures like mean cortical bone thickness and the femoral cortex index at the distal femur for prediction[18]. Other studies have reported correlations between the corticomedullary index of the distal humerus and epiphyseal BMD[19], and between mandibular cortical width and DXA-measured BMD[20]. Although these findings demonstrate the potential of plain radiographs for osteoporosis assessment, their diagnostic accuracy remains limited. Furthermore, BMD obtained from DXA provides only a partial representation of bone quality, which also depends on microstructural integrity and density; notably, certain bone remodeling biomarkers have been shown to predict fracture risk independently of BMD[21].

More recently, however, research in osteoporosis prediction has continuously embraced a variety of radiographic imaging techniques and computational models to enhance diagnostic accuracy. For example, Yamamoto et al.[6] and Jang et al.[9] have expanded the use of deep learning models on hip radiographs, frequently integrating clinical data to boost the models' diagnostic effectiveness. Similarly, Hsieh et al.[8] illustrated the effectiveness of traditional radiographic scans of the pelvis and lumbar spine in predicting bone mineral density (BMD) and assessing fracture risk, alongside automated tools. Ho et al.[10] also utilized plain pelvis X-rays to introduce the DeepDXA model, which uses convolution-based regression to predict BMD, highlighting the potential for broad screening applications with standard radiographic images.

Recent studies have shifted towards leveraging more readily accessible imaging types; for instance, Wang et al.[22] employed chest X-rays and developed a multi-ROI deep model with a transformer encoder to precisely estimate BMD, while Sebro et al.[23] explored the use of computed tomography in analyzing wrist and forearm CT scans for opportunistic screening, revealing positive correlations between CT attenuations of the wrist/forearm bones and DXA measurements. In a slightly different approach, Zheng et al.[24] applied texture-based descriptors to distinguish between healthy and osteoporotic subjects, supporting the role of 2D texture analysis in detecting changes in trabecular bone microarchitecture. Our work distinguishes itself by focusing on hand radiographs, leveraging these peripheral yet commonly available radiographic data. We enhance our approach by applying self-supervised learning on the hand X-rays, aiming to encourage the model to learn features within the cluster of articulating bones in each hand and wrist, which are likely directly correlated with osteoporotic labels.

Medical Image Classification as Segmentation. The recent trend in medical image analysis has seen an innovative convergence of segmentation techniques with classification tasks, considerably increasing diagnostic accuracy and model robustness. This synthesis allows for the leveraging of spatial and structural features inherent in segmentation tasks to improve the classification of various medical conditions. Early endeavors, exemplified in Wong et al.[25], introduced a curriculum learning-inspired approach where features from segmentation networks are utilized to facilitate classifying complex structures, such as brain tumors and cardiac levels, by initially training networks to understand simpler shape and structural concepts through segmentation. Further research, as discussed in Heker et al.[26] and Mojab et al.[27], continued to expand on this foundation by blending transfer and joint learning techniques to optimize feature extraction and model robustness in medical imaging tasks, such as liver lesion segmentation and classification. These studies underline the effectiveness of using segmentation as a preliminary step to classification, achieving notably better results than traditional multi-task architectures. In Gare et al.[28], authors took a different approach by employing pre-trained segmentation models for diagnostic classification, termed reverse-transfer learning, which highlighted the advantages of dense versus sparse segmentation labeling and reduced false positive rates in lung ultrasound analysis. Finally, Saab et al.[29] addressed the reliance of neural networks on spurious features by increasing spatial specificity through segmentation. By providing more precise location data of abnormalities, the models achieved greater robustness against misleading features, which was particularly important in tasks like pneumothorax and melanoma classification. Each of these studies highlights a unique dimension of how segmentation can enrich classification tasks in medical imaging, from increasing data richness to improving model reliability and reducing the need for extensive annotated datasets. Our work builds upon these advancements by specifically applying segmentation to derive features from the ulna, radius, and five metacarpals within a self-supervised learning framework for hand and wrist radiographs.

Methods

We begin with the hypothesis that directly employing raw X-ray images may not serve as the ideal input for our residual neural networks for classification due to their large dimensions, which surpasses the handling capacity of standard image classification frameworks. Moreover, merely resizing these images could pose the risk of losing essential features such as bone texture and microarchitecture[23,24,30], vital for making precise osteoporosis predictions. As an alternative, instead of using arbitrary small image patches, we propose using images based on specific bones, obtained through segmentation of the targeted areas. This approach of segment-specific analysis is backed by several pieces of research demonstrating the benefits of incorporating segmentation methods into classification tasks[25–29,31,32]. Our approach also aligns with findings from earlier research indicating a significant correlation between the 2MCI, which can be derived from segmented inputs, and BMD measurements[11,12]. Consequently, we opt to isolate individual bones in hand and wrist radiographic images to enhance the specificity of our analysis.

In this section, we outline the process for creating segmentation masks for the bones of interest, introduce multi-crop data augmentation methods, and detail our approach to applying self-supervised pre-training to segmented image inputs followed by supervised classification task.

Segmentation mask for target bones

Problem setup

Accurate segmentation of target bones in radiographic images is challenging due to annotation variability and inherent ambiguity in image interpretation. Different experts may provide slightly different but equally valid segmentations, especially near boundaries, in occluded regions, or where image contrast is low. A standard encoder–decoder architecture with a single deterministic output often collapses these possibilities into an averaged prediction, which can obscure clinically relevant details. To better capture this variability, we frame segmentation as a probabilistic prediction problem over a distribution of plausible outputs rather than a single deterministic mask. We adopt an encoder–decoder structure similar to U-Net[14], consistent with prior work[33], but extend it with a modular and uncertainty-aware design (see Fig. 2).

We begin with a labeled dataset $\mathcal{D} = (x_i, y_i^{(j)}, \beta_i^{(j)})_{i=1:D}^{j=1:M}$, where $x_i \in \mathbb{R}^{H \times W}$ represents the i-th radiographic image, and each image has M distinct segmentation annotations $y_i^{(j)} \in \{1, \ldots, C\}^{H \times W}$ with associated probabilities $\beta_i^{(j)}$ satisfying $\sum_j \beta_i^{(j)} = 1$. Here, C denotes the number of classes, and H and W indicate the image's height and width, respectively. Our goal is to learn a model that outputs a probability distribution over segmentations that matches the variability in these annotations.

To achieve this, we employ multiple decoder modules, each capable of representing a different mode of the segmentation distribution, and introduce controlled stochasticity through module-specific Gaussian noise in the

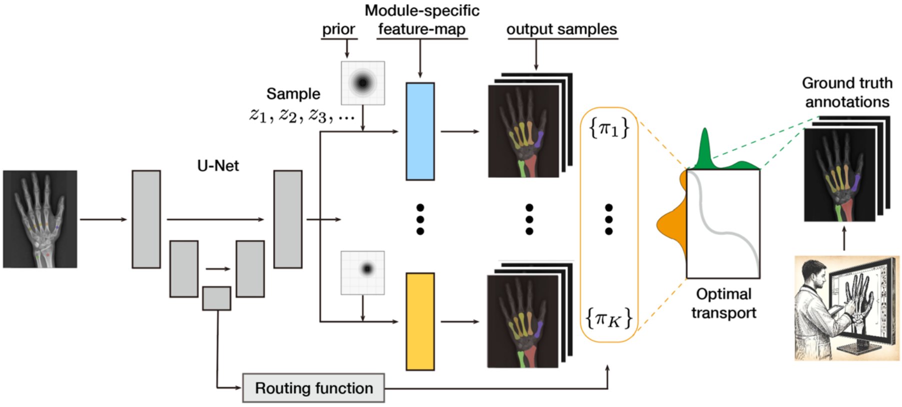

Fig. 2. Overview of the proposed uncertainty-aware segmentation framework. The model features a shared U-Net encoder and multiple decoder modules, each designed to capture a distinct mode of the segmentation distribution. For each decoder, stochasticity is introduced via a module-specific latent vector sampled from an isotropic Gaussian prior, which is transformed into feature-wise scale and bias terms before modulating the decoded feature maps. This Gaussian prior imposes no directional bias in the latent space, ensuring an unbiased representation of uncertainty. The resulting stochastic features are passed through 1×1 convolution layers to produce multiple plausible segmentation hypotheses. A routing function, operating on global encoder features, assigns input-dependent weights to each module's output. The final predictive distribution is a weighted mixture of all module outputs, trained with an optimal transport-based loss to align the predicted and annotated segmentation distributions.

feature space. This combination enables the model to produce anatomically coherent yet diverse segmentation hypotheses, which are then combined via a learned routing function into a single predictive distribution.

Uncertainty-aware segmentation with modularity

Our segmentation framework builds on a shared encoder–decoder backbone but replaces the single decoder with K parallel decoder modules. Each module is designed to capture a distinct plausible mode of the segmentation distribution, thereby modeling the multimodality induced by annotation variability and image ambiguity. The encoder $f_{\text{enc}}(x)$ maps an input x to a global feature map $u_e \in \mathbb{R}^{F_1 \times H' \times W'}$, and each decoder module $f_{\text{dec}}^{(k)}$ transforms this shared representation into a deterministic decoded feature map $u_d \in \mathbb{R}^{F_2 \times H \times W}$. To capture aleatoric uncertainty[34], we introduce a module-specific stochastic latent vector $z_k \sim \mathcal{N}(\mu_k, \sigma_k^2 I)$. The isotropic Gaussian prior ensures rotational symmetry in the latent space, avoiding bias toward any specific anatomical direction and enabling smooth, continuous variability within each module. Rather than injecting noise at the input or output pixel level, which can produce incoherent artifacts, we modulate the decoded features in a feature-wise affine manner: $\bar{u}_d^k[:, i, j] = W_s z_k \odot (u_d[:, i, j] + W_b z_k)$ where $W_s, W_b \in \mathbb{R}^{F_2 \times L}$ project the latent vector into scale and bias terms that are applied to the decoded features. This approach preserves spatial coherence while allowing structured variation within each module's predictions. The resulting stochastic feature maps are passed through three 1×1 convolution layers to produce the k-th segmentation output $s_k = f_{\text{Conv1D}}\left(\bar{u}_d^k\right)$. For each module, we draw S latent samples $\{z_k^{(j)}\}_{j=1}^{S}$ to generate S segmentation hypotheses $\{s_k^{(j)}\}_{j=1}^{S}$, which are averaged to form the module-specific predictive distribution. To combine these into the final predictive distribution, we employ a routing function $r(\cdot)$ operating on global encoder features $\bar{u}_e = \text{AvgPool}(u_e)$. This vector is processed by a multi-layer perceptron that outputs mixture weights $\pi_k(x)$, resulting in

$$q_\theta(y|x) = \sum_{k=1}^{K} \pi_k(x) \left(\frac{1}{S} \sum_{j=1}^{S} \delta(s_k^{(j)}) \right) \tag{1}$$

where θ includes all encoder, decoder, noise-projection, and routing parameters. This mixture-of-decoders architecture, combined with structured stochastic modulation, enables the model to represent rich, multimodal segmentation distributions that align with the inherent uncertainty in the data.

Optimal transport-based loss

We define the learning process of our model as the minimization of a Wasserstein loss[35] between the model prediction $q_\theta(y|x)$ and the actual annotated distribution of the image $p(y|x)$:

$$q_\theta(y|x) = \sum_{i=1}^{N} \alpha^{(i)} \delta(s^{(i)}) = \sum_{k=1}^{K} \sum_{j=1}^{S} \frac{\pi_k(x)}{S} \delta(s_k^{(j)}) \tag{2}$$

$$p(y|x) = \sum_{j=1}^{M} \beta^{(j)} \delta(y^{(j)}) \tag{3}$$

This is approached as an optimal transport (OT) problem[36,37], which determines the most cost-efficient way to reallocate the probability mass from $q_\theta(y|x)$ to the true distribution $p(y|x)$, guided by minimizing a cost matrix $\mathcal{C} \in \mathbb{R}^{N\times M}$. For the cost $\mathcal{C}_{ij}$ between pairs $s^{(i)}$ and $y^{(j)}$, we employ the generalized energy distance (GED)[15,16], which is often used to assess the alignment of the model predictions with the ground truth labels within the segmentation output space. By leveraging the optimal transport framework, the model explicitly accounts for the inherent variability in segmentations by aligning the predicted distribution with the annotated distribution. This approach ensures that the model can handle ambiguous regions and improve robustness, even when the ground truth contains inherent uncertainty or noise. In contrast to the Kullback-Leibler divergence, OT offers a well-defined distance measure that remains valid even when the compared distributions have differing supports. Based on this loss, our objective function can be written as follows:

$$\min_{\theta_s} \sum_{n=1}^{D} \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{T}^*_{ij}\, \mathcal{C}\left(s_n^{(i)}(x_n;\theta_s), y_n^{(j)}\right) \tag{4}$$

$$\text{s.t.} \quad \mathcal{T}^* = \mathop{\arg\min}_{\theta_\pi,\, \mathcal{T}\in U} \sum_{i,j} \mathcal{T}_{ij}\mathcal{C}_{ij} \tag{5}$$

$$\text{and} \quad \mathcal{U} = \left\{\mathcal{T} \in \mathbb{R}_+^{N\times M} : \mathcal{T}\mathbb{1}_M = \boldsymbol{\alpha}(x_n;\theta_\pi), \mathcal{T}^\top \mathbb{1}_N = \boldsymbol{\beta}\right\} \tag{6}$$

where $\mathcal{T}$ is a coupling matrix indicating the transportation plan, $\mathcal{U}$ is the set of all possible values of $\mathcal{T}$ which satisfy the marginal constraints, and $\mathbb{1}$ is a vector of ones. Eqs. (4–6) constitute a bilevel optimization problem, where the inner problem is responsible for identifying the optimal coupling matrix $\mathcal{T}$, while the outer problem aims to refine the model's output distribution within the framework of this coupling. In contrast to the conventional optimal transport problem, the dependency of the marginal probabilities $\alpha(x_n;\theta_\pi)$ on the routing function parameters θ_π prevents the direct use of backpropagation for optimizing the entire set of parameters $\theta \triangleq \theta_s \cup \theta_\pi$. To overcome this challenge, we use a method of imbalanced constraint relaxation[37].

Imbalanced constraint relaxation

To tackle the intractability of back-propagation for optimizing the routing parameters θ_π, we shift the constraint of the predictive marginal from Eq. 6 to the outer problem outlined in Eq. 7 as described below:

$$\min_{\theta_s,\theta_\pi} \sum_{n=1}^{D} \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{T}^*_{ij}\, \mathcal{C}\left(s_n^{(i)}(x_n;\theta_s), y_n^{(j)}\right) + \lambda D_{\mathrm{KL}}\left(\sum_{j} \mathcal{T}^*_{ij} \,\|\, \boldsymbol{\alpha}(x_n;\theta_\pi)\right) \tag{7}$$

$$\text{s.t.} \quad \mathcal{T}^* = \mathop{\arg\min}_{\mathcal{T}\in U'} \sum_{i,j} \mathcal{T}_{ij}\mathcal{C}_{ij} \tag{8}$$

$$\text{and} \quad \mathcal{U}' = \left\{\mathcal{T} \in \mathbb{R}_+^{N\times M} : \mathcal{T}\mathbb{1}_M \le \gamma \cdot \mathbb{1}_N, \mathcal{T}^\top \mathbb{1}_N = \boldsymbol{\beta}\right\} \tag{9}$$

where λ serves to balance the conventional segmentation loss (the first term in Eq. 7) and the cross-entropy loss (the second term in Eq. 7) between the predicted probabilities and a pseudo label generated from the coupling matrix. The range of γ, from $\frac{1}{N}$ to 1, is strategically chosen to avoid sub-optimal solutions during the initial phases of training. Empirically, we begin with a value $\gamma_0 < 1$ and gradually anneal it to 1 (effectively removing the constraint). The underlying idea is to use the optimally solved matrix $\mathcal{T}^*$ as supervision for the routing function in the outer problem outlined in Eq. 7.

Multi-crop data augmentation

The resultant output segmentation masks predicted in the previous section are then used to produce targeted bone image segments using a simple Hadamard multiplication. In the end, we derive 7 unique image patches from an original radiograph, corresponding to the ulna, radius, and the five metacarpals (Fig. 3). It should be noted that there were instances of deviation resulting in fewer than seven target image segments. Such deviations were typically caused by unusual positioning of the hand or the presence of accessories or medical apparatus on the patient's hand.

The bone segments created above exhibit variations in rotational angles (Fig. 3), primarily due to the inherent anatomical structure of the hand and wrist, as well as the hand position. To equip our model with the ability to reliably process these variations, we implement several data augmentation strategies. These include applying random rotations between -30 and $+30$ degrees, horizontal and vertical translations within a range of -10 to $+10\%$ of the image's width and height, flipping the image randomly in both vertical and horizontal directions, and adjusting the brightness and contrast from 50 to 150% of their original values. Furthermore, we adopt the multi-crop data augmentation method[17], which is advantageous for effective training of pretext tasks as discussed in Section . This method samples images of two distinct resolutions: randomly seletected two global

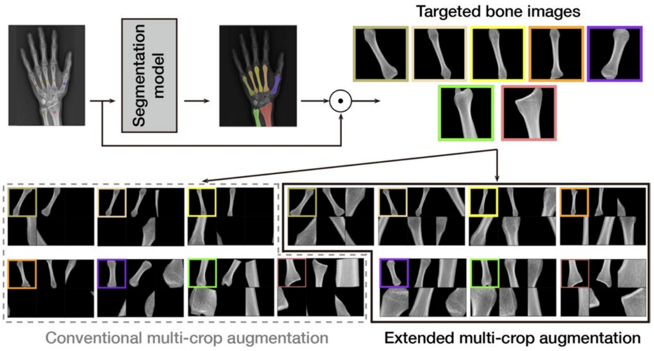
Targeted bone images
Segmentation model
Conventional multi-crop augmentation
Extended multi-crop augmentation

Fig. 3. Illustration of the process for extracting targeted bone images from segmentation outputs, and a comparison of two multi-crop augmentation methods. (Top) Seven bone image patches, each outlined by a square in the same color as the corresponding segmentation mask overlaid on the hand X-ray image. (Bottom) Each of the seven image segments undergoes two different multi-crop augmentation strategies: the conventional method (represented by a gray dashed line) and our extended method (indicated by a solid black line). A global view matching the color of the input bone patch and five local views are displayed side by side.

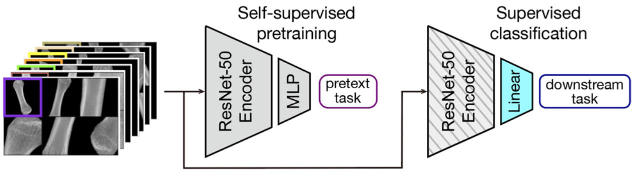

Fig. 4. Two-stage training for osteoporosis prediction. In the first stage, augmented samples are processed through an encoder for self-supervised pretraining as part of a pretext task. The backbone classification encoder network (gray trapezoid with slanting lines) is subsequently fixed and repurposed to address the downstream classification task by incorporating a trainable linear layer (cyan trapezoid).

views (224 × 224) and four local views (96 × 96). By using crops of different sizes, the model is exposed to objects at various scales, which enhances its scale invariance.

Conventional multi-crop augmentation, often applied to standard image classification benchmarks or natural images, ensures each cropped segment remains feature-rich. In contrast, our specific bone segments frequently have vast areas of zero pixels (Fig. 3), a result of the segmentation masking process. This characteristic persists even in the original radiographs before applying the segmentation method due to the sparse anatomical features in hand X-ray images. Consequently, some samples might entirely lack bone content (Fig. 3). To mitigate this, we tailor the multi-crop strategy to our needs, ensuring that the augmented samples are spatially constrained crops[38,39]. Specifically, our modified method guarantees that both global and local views contain a sufficient amount of feature data, employing iterative rejection sampling to maintain at least a 10% non-zero pixel presence in each crop. This threshold is set to preserve the integrity of the features, reducing the chance of valuable information being omitted from our samples, as visualized in Fig. 3. This adjustment secures the inclusion of meaningful content in each cropped segment, optimizing their contribution to the pre-training stage.

Contrastive self-supervised pretraining

Our next step progresses through two separate training phases, as illustrated in Fig. 4. The first phase involves a pretext task where we train the ResNet-50 encoder using the augmented views described previously. This phase operates under the premise that transformations applied to an image do not change its semantic content. Consequently, the task focuses on bringing closer the representations of different augmentations of the same image, referred to as a positive pair, in the latent space. This crucial step allows the encoder to focus on the inherent data structures and enables the model to learn robust, generalizable features from the radiographic images without relying on human-provided labels. For the optimization of this process, we employ the LARC optimizer[40], along with a cosine annealing scheduler[41] to adjust the learning rates. The entire training procedure is executed with a batch size of 128.

In this phase, we assess four well-known SSL techniques: SimCLR[42], SupCon[43], SwAV[17], and VICReg[44], all categorized under contrastive self-supervised methods. They collectively aim to maximize consistency among augmented views within a shared embedding space, ensuring that semantically similar data points are brought closer together. Specifically, SimCLR[42] concentrates on contrasting different augmented views of the same instance against other instances, while SupCon[43] extends this approach by also pulling instances of the same class closer together in the embedding space. SwAV[17] introduces a swapped prediction mechanism that bypasses the need for negative pair comparisons, making it computationally more efficient. Meanwhile, VICReg[44] stands out by imposing variance, invariance, and covariance regularization to the latent space to ensure the features

Gender	Normal	Osteoporotic	Total
Male	18	5	23
Female	115	54	169

Table 1. Number of normal and osteoporotic samples categorized by gender.

Age group	Normal	Osteoporotic	Total
10–19	0	2	2
20–29	8	0	8
30–39	9	2	11
40–49	24	4	28
50–59	51	12	63
60–69	23	21	44
70–79	13	13	26
80–89	4	5	9
90–99	1	0	1

Table 2. Number of normal and osteoporotic samples categorized by age.

captured are diverse and not merely reflective of trivial patterns. Exploring these effective SSL methods in peripheral radiographic images still remains nascent, making our experimental results potentially useful for clinical applications in other medical domains.

Supervised fine-tuning

Following the initial phase, the pre-trained encoder is fine-tuned on a dataset that is explicitly labeled for our classification task. In this downstream phase, the encoder serves as an efficient feature extractor, with only its final linear layer being trained to precisely predict osteoporosis. Importantly, the inputs to this pre-trained encoder are not raw hand X-ray or cropped images, but rather bone segments isolated by the segmentation model. We initially aim to perform subject-wise osteoporosis screening; however, the current training model is not suited for such subject-wise predictions as it outputs multiple sub-binary decisions for different bone segment inputs from each individual. To address this, we maintain the bone segment-wise input approach but revise the evaluation protocol by aggregating multiple sub-decisions to form a final subject-wise decision. Specifically, each sub-decision from the bone segments yields a probability score for each class. The final decision is derived from the average probability across the seven image segments of the ulna, radius, and five metacarpals:

$$\hat{y} = \underset{y_j}{\arg\max}\ \langle p(y_j|x_i) \rangle_{i=1:7} \tag{10}$$

where $p(y_j|x_i)$ represents the probability predicted by the i-th bone image segment x_i for class y_j, and the final class prediction $\hat{y}$ is determined by the class with the highest average probability. Through this approach, the classifier is trained to uniformly identify osteoporosis across all bone segments during the fine-tuning phase. At the time of evaluation, the assessment criteria are adjusted so that the model's prediction is considered accurate even if only a few bone segments are diagnosed as positive. This adjustment ensures that the model's ability to detect osteoporosis is evaluated based on its overall performance across various segments, rather than requiring a unanimous positive diagnosis from all segments.

Datasets & metrics

Hand X-ray datasets

We leverage two primary datasets of hand X-ray images. The first dataset consists of hand radiographic images from 192 individual subjects, featuring a resolution of 2515 × 3588. For precise classification, we correlate our data with DXA scans. DXA measures BMD by assessing the variation in X-ray attenuation across different bone areas, quantifying the concentration of calcium and other minerals in these segments. The derived BMD values are used to compute the T-score, which benchmarks a patient's BMD against the average BMD of a healthy 30-year-old adult. A negative T-score indicates a BMD lower than average, while a positive T-score denotes a higher than average BMD. In our study, this T-score is employed to annotate radiographic images, categorizing an X-ray as indicative of osteoporosis if the T-score falls below −2.5, and as normal otherwise. Table 1 shows that out of 192 total samples, 23 belong to males (18 normal and 5 osteoporotic), while 169 belong to females (115 normal and 54 osteoporotic), indicating a higher prevalence of osteoporosis among females. Table 2 further breaks down the samples by age groups, highlighting that the majority of osteoporotic cases are found in the 50-59, 60-69, and 70-79 age groups, while younger age groups (e.g., 10-19 and 20-29) show significantly fewer osteoporotic cases. All experiments were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects and/or their legal guardian(s) prior to participation. This study was approved by the Institutional Review Board of Hanyang University Hospital (IRB No.: 2023-10-011).

Additionally, we use a publicly available dataset[45] containing 1154 hand X-ray images, each measuring 1400 × 900. Originally compiled for an object detection benchmark and lacking osteoporosis labels, this dataset is exclusively utilized for training our bone segmentation model and pretraining the backbone classification encoder network. We apply a subject-wise random split of 80% for training, 10% for validation, and 10% for testing, maintaining this ratio consistently across all experiments and different seeds.

To generate the ground truth labels required for training our segmentation model, we leverage the Segment Anything Model (SAM)[46], a foundational model for image segmentation. Our initial experiences indicate that SAM necessitates manual interventions, such as placing positive and negative point prompts inside and outside the target bone regions respectively, to produce segmentation masks that match the expert-level precision of radiologists and orthopedists. Our ultimate goal is to develop a fully automated osteoporosis screening model that integrates classification through segmentation without relying on manual segmentation steps. Therefore, we employ SAM as an auxiliary tool for creating ground truth bone segments, guided by expert feedback.

Metrics

All test results are obtained from the optimal model identified during training, which is determined by achieving the highest validation macro-F1 score. This measure provides a reliable assessment of model performance, especially under conditions of imbalanced labels[47]. For evaluation purposes, we employ conventional classification performance metrics, such as precision, recall, F1 score, AUC, and accuracy rate. We report them specifically for the positive label. Particular emphasis is placed on the AUC because of the class imbalance in our dataset. A higher AUC value indicates that the classifier is more effective at distinguishing between positive and negative cases of osteoporosis screening across various classification thresholds.

Results

Evaluation on hand X-ray images

The results presented in Table 3 compare the performance of four different self-supervised learning models—SimCLR, SupCon, SwAV, and VICReg—on the task of osteoporosis prediction using five different metrics previously described. Each model was evaluated across three trials with distinct random seeds, providing an average for each metric. Among these models, SimCLR achieves the highest F1 score of 0.68 ± 0.03 and an AUC of 0.85 ± 0.01, indicating strong predictive performance and consistency in ranking the positive class. In contrast, SwAV appears to struggle relative to the other models, with both its F1 and AUC scores the lowest at 0.54 ± 0.13 and 0.68 ± 0.04 respectively, suggesting potential challenges in handling our specific dataset or task. VICReg and SupCon, showing a balanced performance across precision, recall, and F1, exhibits a reasonably good AUC, demonstrating effective but not optimal performance.

Effect of individual bones on osteoporosis detection

While the previous experiment demonstrated the model's accuracy in diagnosing osteoporosis among individuals, we aimed to further investigate the model's decision-making process by visualizing the predicted probabilities of osteoporosis for each bone involved in the final prediction. We gathered probability values from individual bone images of true positive cases, specifically from our best-performing SimCLR-based classification model. Our analysis indicated that, on average, the 2nd metacarpal, radius, and ulna are the primary contributors to detecting osteoporosis (Fig. 5). This outcome is in line with studies that recognize the 2MCI as a reliable biomarker, and it corroborates clinical findings that the radius and ulna—integral parts of the forearm—are critical in osteoporosis diagnosis.

Ablation studies

To gain more insights into the features that drive the model's predictions, we performed an ablation on our proposed model framework. This process involves selectively removing one of the following steps in a series: i) segmenting raw radiological images, ii) applying enhanced multi-crop augmentation to the masked images, iii) pre-training encoders through self-supervised learning, and iv) fine-tuning classifiers in a supervised manner. Initially, to determine whether bone segmentation is essential for osteoporosis prediction, we redesigned an experiment in which the segmentation step is omitted. In this setup, raw X-ray images are directly used as inputs for both the pre-training and fine-tuning stages. Consequently, the final class prediction relies on a single model prediction probability rather than an average of multiple predictions. Despite these changes, pre-training proceeds with the same data augmentation and SSL methods as in previous experiments. To ensure a fair comparison with the original experiment, we maintained the same number of augmented image samples

	Precision	Recall	F1	AUC	Accuracy
SimCLR	**0.71 ± 0.04**	**0.66 ± 0.06**	**0.68 ± 0.03**	**0.85 ± 0.01**	**0.81 ± 0.03**
SupCon	0.61 ± 0.01	0.60 ± 0.11	0.60 ± 0.06	0.79 ± 0.03	0.75 ± 0.04
SwAV	0.54 ± 0.12	0.54 ± 0.16	0.54 ± 0.13	0.68 ± 0.04	0.72 ± 0.05
VICReg	0.64 ± 0.07	0.64 ± 0.07	0.64 ± 0.07	0.80 ± 0.03	0.77 ± 0.03

Table 3. Performance comparison of osteoporosis prediction using the proposed model framework on held-out radiographic test images. The table reports the precision, recall, F1-score, AUC, and accuracy for four different methods (SimCLR, SupCon, SwAV, and VICReg), with each metric averaged over three trials initiated with distinct random seeds. The highest performance for each metric is highlighted in bold.

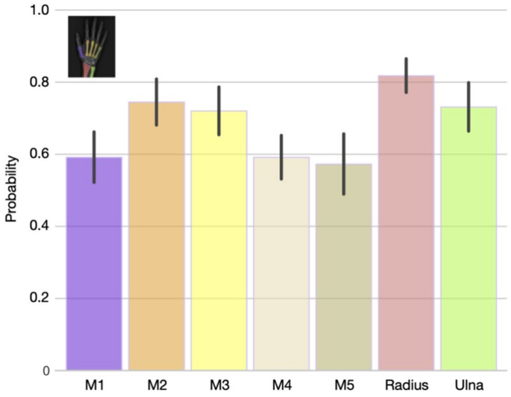

Fig. 5. Individual contributions of bone segments to osteoporosis detection. Predicted probabilities for positive labels, illustrating the role of each bone segment, including metacarpals (M1–M5), radius, and ulna, in detecting osteoporosis. Error bars indicate the uncertainty of each segment's prediction. The bar colors correspond to the segmentation mask colors shown in the inset.

	F1	Δ	AUC	Δ	Accuracy	Δ
SimCLR	**0.54 ± 0.09**	−0.12	**0.78 ± 0.04**	−0.07	**0.74 ± 0.01**	−0.07
SupCon	0.51 ± 0.06	−0.09	0.73 ± 0.04	−0.06	0.70 ± 0.06	−0.05
SwAV	0.52 ± 0.07	−0.02	0.65 ± 0.06	−0.03	0.71 ± 0.05	−0.01
VICReg	**0.54 ± 0.07**	−0.10	0.74 ± 0.03	−0.06	0.70 ± 0.08	−0.07

Table 4. Results of osteoporosis prediction after omitting the segmentation step, detailed across three key performance metrics: F1, AUC, and accuracy. The table illustrates the change in model performance (denoted by Δ) when segmentation is excluded, based on averages from three trials with different random seeds. The most significant performance drops are underlined for emphasis, while the highest performance for each metric is highlighted in bold.

	F1	Δ	AUC	Δ	Accuracy	Δ
SimCLR	**0.62 ± 0.05**	−0.06	**0.80 ± 0.02**	−0.05	**0.77 ± 0.04**	−0.04
SupCon	0.56 ± 0.07	−0.04	0.75 ± 0.04	−0.04	0.72 ± 0.05	−0.03
SwAV	0.51 ± 0.10	−0.03	0.65 ± 0.05	−0.03	0.70 ± 0.05	−0.02
VICReg	0.60 ± 0.06	−0.04	0.76 ± 0.04	−0.04	0.74 ± 0.04	−0.03

Table 5. Osteoporosis prediction performance of the conventional single-decoder U-Net and the corresponding change in model performance (Δ) compared to the proposed multi-decoder segmentation. Negative Δ values indicate performance drops relative to our proposed framework. Averages are computed over three trials with different random seeds. The most significant performance drops are underlined for emphasis, while the highest performance for each metric is highlighted in bold.

provided to the encoder during pre-training by randomly sampling seven image patches from each hand X-ray image. Table 4 displays the prediction results, focusing on three key metrics: F1 score, AUC, and accuracy. Across all models, the omission of bone segmentation results in a noticeable decline in model performance. Specifically, SimCLR experienced the largest decrease in F1 score by 0.12, AUC by 0.07, and accuracy by 0.07, followed closely by VICReg, which showed similar declines in both metrics. These findings underscore the importance of the segmentation step in improving model accuracy and reliability, demonstrating that precise localization of relevant bone features significantly contributes to the performance of osteoporosis screening. Conversely, the results suggest that task-irrelevant features, such as soft tissues including muscles, fat, skin, and nerves, which absorb X-rays less than bones, may introduce interference in distinguishing osteoporosis cases.

To further examine the performance gains achieved by our proposed uncertainty-aware, multi-decoder segmentation framework, we conducted a comparative experiment using a conventional U-Net architecture with a single decoder head (Table 5). This baseline was trained under identical conditions and dataset splits as our proposed model, with all other components of the pipeline, including multi-crop augmentation, SSL pretraining, and supervised fine-tuning, remained unchanged. Across the four SSL backbones, the single-decoder U-Net consistently underperformed relative to our multi-decoder approach, with SimCLR showing the most pronounced differences: F1 score decreased from 0.68 to 0.62, AUC dropped from 0.85 to 0.80, and accuracy

	F1	Δ	AUC	Δ	Accuracy	Δ
Supervised	0.53 ± 0.04	−0.15	0.67 ± 0.03	−0.18	0.72 ± 0.01	−0.09

Table 6. Impact of removing data augmentation and self-supervised pre-training on osteoporosis prediction. The table presents the F1-score, AUC, and accuracy of the supervised model. The change (Δ) in performance, compared to the baseline model, is also provided for each metric, based on the average results across multiple trials.

	F1	Δ	AUC	Δ	Accuracy	Δ
SimCLR	**0.65 ± 0.03**	−0.03	**0.82 ± 0.02**	−0.03	**0.79 ± 0.05**	−0.02
SupCon	0.56 ± 0.04	−0.04	0.74 ± 0.04	−0.05	0.72 ± 0.08	−0.03
SwAV	0.51 ± 0.09	−0.03	0.66 ± 0.05	−0.02	0.70 ± 0.03	−0.02
VICReg	0.59 ± 0.02	−0.05	0.78 ± 0.03	−0.02	0.74 ± 0.05	−0.03

Table 7. Counter-effect of conventional multi-crop augmentation on osteoporosis screening. The highest performance for each metric is highlighted in bold.

Model	ROI extraction method	Architecture	F1	AUC	Accuracy
Jang et al.[9]	Manual cropping	VGG16	0.43 ± 0.05	0.63 ± 0.03	0.62 ± 0.04
Hsieh et al.[8]	Landmark-based cropping	VGG16	0.48 ± 0.04	0.65 ± 0.04	0.65 ± 0.03
Wang et al.[22]	Landmark-based cropping	VGG16 + Transformer	0.46 ± 0.06	0.64 ± 0.05	0.64 ± 0.03
Ho et al. (DeepDXA)[10]	Segmentation-based	ResNet18	**0.51 ± 0.03**	**0.66 ± 0.02**	**0.67 ± 0.02**
Ours (full framework)	Segmentation-based + SSL + enhanced aug	ResNet50	**0.68 ± 0.03**	**0.85 ± 0.01**	**0.82 ± 0.02**

Table 8. Comparison of our proposed model against existing supervised baselines. All baselines are trained end-to-end without SSL or enhanced augmentation. The first and second highest performance for each metric is highlighted in bold.

fell from 0.81 to 0.77. Similar trends were observed for SupCon, SwAV, and VICReg, with average performance declines of approximately 0.04 in F1 score and 0.03–0.05 in AUC. These results indicate that explicitly modeling segmentation uncertainty through multiple decoders not only improves localization of bone structures but also enhances downstream classification performance, by providing more robust and feature-consistent inputs for SSL-based representation learning.

In the second ablation study shown in Table 6, we examined the implications of excluding multi-crop augmentation and SSL pre-training phases while retaining bone segmentation within the model framework. When compared to the optimal results achieved with the SimCLR-based model in Table 3, this current setup demonstrates a considerable drop in performance: the F1 score decreases by 0.15 to 0.53 ± 0.04, AUC by 0.18 to 0.67 ± 0.03, and accuracy by 0.09 to 0.72 ± 0.01. A critical observation is the supervised model's relatively low accuracy of 0.72, which nearly aligns with the proportion of non-osteoporotic patients, suggesting that the model might predominantly predict one class regardless of variations in the input radiographic images. This result emphasizes the limitations of a purely supervised approach in learning meaningful representations for effectively distinguishing osteoporosis from normal cases, while reinforcing the importance of SSL in learning robust, generalizable representations from unlabeled data to counteract the constraints of a limited labeled dataset.

Table 7 presents the outcomes of the last ablation study within the original model framework, with a key modification: the implementation of conventional multi-crop augmentation rather than an extended version. The results show a general decline in performance metrics across all models when compared to previous experiments that allowed more bone content in the augmented images. Specifically, SimCLR, which previously showed robust performance, now records a minor dip in every metric. A similar downward trend is observed across all SSL methods. These quantitative reduction suggest that our customized multi-crop augmentation, tailored for hand radiographic images, produces more favorable results, highlighting the significance of pinpointing the sufficient and pertinent bone regions.

Comparative analysis with prior supervised approaches

To contextualize the effectiveness of our proposed framework, we compare it against existing supervised learning baselines that have addressed osteoporosis or bone mineral density prediction tasks using radiographic images. These baselines differ in terms of backbone architecture, input preparation strategies, and augmentation pipelines, but all rely solely on supervised training without SSL or our tailored augmentation scheme.

Table 8 summarizes the results. Jang et al.[9] trained a VGG16 model on manually cropped hip radiographs, where the cropping process localized the region of interest but was subject to human variability and potential omission of subtle features. Hsieh et al.[8] and Wang et al.[22] employed landmark-based cropping, in which anatomical keypoints detected by an automated model defined the ROI boundaries for hips or lumbar vertebrae. This method ensures reproducible cropping but still depends on accurate landmark detection and may exclude peripheral cues outside the predefined boxes. Ho et al.[10] adopted a segment-based approach, isolating bone structures before classification, which more closely resembles our anatomical preprocessing but without our enhanced augmentation or SSL pre-training stages.

Among these baselines, the ResNet18-based DeepDXA model of Ho et al. achieves the strongest supervised performance. Beyond being deeper than VGG16, ResNet18 incorporates a residual learning structure with identity shortcut connections, enabling more efficient gradient flow, mitigating vanishing gradients, and producing more stable optimization. Interestingly, the VGG16+Transformer design of Wang et al., despite leveraging modeling inter-ROI relationships and landmark-based ROIs, does not surpass simpler CNN-based approaches, likely due to overfitting in a small-data regime. Building on this, our proposed framework outperforms all baselines by a substantial margin, improving F1 score by 0.17, AUC by 0.19, and accuracy by 0.15 over the best-performing supervised method (DeepDXA). The comparative analysis suggests that these gains can be attributed to the combined benefits of augmentation strategies tailored for sparse bone regions, robust representation learning through SSL pre-training, and effective supervised fine-tuning.

Discussion

The prevalence of osteoporosis, a disease that weakens bones and increases the risk of fractures, is rising alongside an aging global population. Diagnosing this condition typically requires analysis of BMD, most commonly through DXA tests, which can be costly and not widely available. Radiographic images, commonly generated in clinical settings, offer a rich, underutilized dataset for detecting osteoporotic changes in bone structure using less invasive methods.

In this study, we explored the potential of using hand X-ray images to determine osteoporosis status by introducing the robust segmentation-for-classification framework and leveraging self-supervised pre-training. Our work has a broader impact because this system could greatly reduce the clinician's workload by swiftly pinpointing likely positive cases. Moreover, it could improve patient care by detecting osteoporotic cases that might initially go unnoticed, thus prompting clinicians to reconsider initial assessments and address potential diagnostic biases. This approach could also be integrated into routine workflows without significant infrastructure changes, making it adaptable across diverse healthcare environments and resource-constrained regions. Furthermore, the method's scalability allows for large-scale population screening, which is essential for early detection and prevention strategies.

Although the current study was trained on a relatively small dataset of a few hundred samples, our goal was to demonstrate the feasibility of using peripheral hand X-rays for osteoporosis detection, particularly in settings with limited resources. The limited dataset size poses challenges in fully assessing the model's generalizability to other populations or imaging modalities, and addressing this constraint is a priority for future work. We plan to expand the dataset to include tens of thousands of unlabeled radiographs, which could enhance performance, strengthen predictive capabilities, and improve clinical applicability. Despite these current limitations, we remain optimistic about the potential of this approach to enable more precise and user-friendly osteoporosis screening methods, which we aim to refine and develop further.

Data availability

The clinical hand X-ray dataset with corresponding DXA-based T-score annotations, used and/or analyzed during the current study, is available from the corresponding author upon reasonable request. The publicly available hand X-ray dataset used for segmentation and pretraining, originally compiled for object detection tasks, is accessible via the Roboflow repository: https://universe.roboflow.com/hf-w1rwi/hand-xray.

Received: 23 June 2025; Accepted: 19 August 2025
Published online: 26 September 2025

References

1. Chou, S. H. & LeBoff, M. S. Vertebral imaging in the diagnosis of osteoporosis: a clinician's perspective. *Current Osteoporosis Reports* **15**, 509–520 (2017).
2. Pickhardt, P. J. et al. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Ann. Iinter. Medi.* **158**, 588–595 (2013).
3. Gausden, E. B., Nwachukwu, B. U., Schreiber, J. J., Lorich, D. G. & Lane, J. M. Opportunistic use of ct imaging for osteoporosis screening and bone density assessment: a qualitative systematic review. *JBJS* **99**, 1580–1590 (2017).
4. Watts, N. B. Fundamentals and pitfalls of bone densitometry using dual-energy x-ray absorptiometry (dxa). *Osteoporosis Int.* **15**, 847–854 (2004).
5. Messina, C. et al. Prevalence and type of errors in dual-energy x-ray absorptiometry. *Eur. Radiol.* **25**, 1504–1511 (2015).
6. Yamamoto, N. et al. Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates. *Biomolecules* **10**, 1534 (2020).
7. Yasaka, K., Akai, H., Kunimatsu, A., Kiryu, S. & Abe, O. Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. *Eur. Radiol.* **30**, 3549–3557 (2020).
8. Hsieh, C.-I. et al. Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nat. Commun.* **12**, 5472 (2021).
9. Jang, R. et al. Prediction of osteoporosis from simple hip radiography using deep learning algorithm. *Sci. Reports* **11**, 19997 (2021).
10. Ho, C.-S. et al. Application of deep learning neural network in predicting bone mineral density from plain x-ray radiography. *Arch. Osteoporosis* **16**, 153 (2021).
11. Schreiber, J. J., Kamal, R. N. & Yao, J. Simple assessment of global bone density and osteoporosis screening using standard radiographs of the hand. *J. Hand Surg.* **42**, 244–249 (2017).

12. Patel, B., Aqil, A., Riaz, O., Jeffers, R. & Dickson, D. The 2nd metacarpal cortical index as a simple screening tool for osteopenia. *J. Bone Metabolism* **27**, 261 (2020).
13. Huang, S.-C. et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit. Med.* **6**, 74 (2023).
14. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).
15. Kohl, S. *et al.* A probabilistic u-net for segmentation of ambiguous images. *Adv. Neural Inform. Processi. Syst.* **31** (2018).
16. Baumgartner, C.F. *et al.* Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 119–127 (Springer, 2019).
17. Caron, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inform. Process. Syst.* **33**, 9912–9924 (2020).
18. Nguyen, B. N., Hoshino, H., Togawa, D. & Matsuyama, Y. Cortical thickness index of the proximal femur: a radiographic parameter for preliminary assessment of bone mineral density and osteoporosis status in the age 50 years and over population. *Clin. Orthopedic Surg.* **10**, 149 (2018).
19. Clavert, P. et al. How to determine the bone mineral density of the distal humerus with radiographic tools?. *Surg. Radiol. Anatomy* **38**, 389–393 (2016).
20. Samelson, E. J. et al. Cortical and trabecular bone microarchitecture as an independent predictor of incident fracture risk in older women and men in the bone microarchitecture international consortium (bomic): a prospective study. *The Lancet Diabetes & Endocrinol.* **7**, 34–43 (2019).
21. Vasikaran, S. et al. Markers of bone turnover for the prediction of fracture risk and monitoring of osteoporosis treatment: a need for international reference standards. *Osteoporosis Int.* **22**, 391–420 (2011).
22. Wang, F. et al. Lumbar bone mineral density estimation from chest x-ray images: anatomy-aware attentive multi-roi modeling. *IEEE Trans. Med. Imagi.* **42**, 257–267 (2022).
23. Sebro, R. & De la Garza-Ramos, C. Machine learning for opportunistic screening for osteoporosis from ct scans of the wrist and forearm. *Diagnostics* **12**, 691 (2022).
24. Zheng, K. & Makrogiannis, S. Bone texture characterization for osteoporosis diagnosis using digital radiography. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2016).
25. Wong, K. C., Syeda-Mahmood, T. & Moradi, M. Building medical image classifiers with very limited data using segmentation networks. *Med. Image anal.* **49**, 105–116 (2018).
26. Heker, M. & Greenspan, H. Joint liver lesion segmentation and classification via transfer learning. arXiv preprint arXiv:2004.12352 (2020).
27. Mojab, N., Philip, S.Y., Hallak, J.A. & Cvs, D.Y. Classification via segmentation for small datasets. arXiv preprint arXiv:2111.00042 (2021).
28. Gare, G.R. *et al.* Dense pixel-labeling for reverse-transfer and diagnostic learning on lung ultrasound for covid-19 and pneumonia detection. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1406–1410 (IEEE, 2021).
29. Saab, K. *et al.* Reducing reliance on spurious features in medical image classification with spatial specificity. In *Machine Learn. Healthcare Confer.*, 760–784 (PMLR, 2022).
30. Lespessailles, E. et al. Clinical interest of bone texture analysis in osteoporosis: a case control multicenter study. *Osteoporosis Int.* **19**, 1019–1028 (2008).
31. Gare, G. R. et al. W-net: Dense and diagnostic semantic segmentation of subcutaneous and breast tissue in ultrasound images by incorporating ultrasound rf waveform data. *Med. Image Anal.* **76**, 102326 (2022).
32. Hooper, S. *et al.* A case for reframing automated medical image classification as segmentation. *Adv. Neural Inform. Process. Syst.* **36** (2024).
33. Hu, S. *et al.* Supervised uncertainty quantification for segmentation with multiple annotations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 137–145 (Springer, 2019).
34. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learn.* **110**, 457–506 (2021).
35. Rüschendorf, L. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields* **70**, 117–129 (1985).
36. Villani, C. *et al. Optimal transport: old and new*, vol. 338 (Springer, 2009).
37. Peyré, G. et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning* **11**, 355–607 (2019).
38. Van Gansbeke, W., Vandenhende, S., Georgoulis, S. & Gool, L. V. Revisiting contrastive methods for unsupervised learning of visual representations. *Adv. Neural Inform. Process. Syst.* **34**, 16238–16250 (2021).
39. Peng, X., Wang, K., Zhu, Z., Wang, M. & You, Y. Crafting better contrastive views for siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16031–16040 (2022).
40. You, Y., Gitman, I. & Ginsburg, B. Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 (2017).
41. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016).
42. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).
43. Khosla, P. et al. Supervised contrastive learning. *Adv. Neural Inform. Process. Syst.* **33**, 18661–18673 (2020).
44. Bardes, A., Ponce, J. & LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021).
45. HF. Hand x-ray dataset. https://universe.roboflow.com/hf-w1rwi/hand-xray (2022).
46. Kirillov, A. *et al.* Segment anything. *arXiv:2304.02643* (2023).
47. Fiorillo, L., Favaro, P. & Faraci, F. D. Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE Trans. Neural Syst. Rehabilitat. Eng.* **29**, 2076–2085 (2021).

Author contributions

U.H. performed data curation, methodology, and software implementation. C.L. contributed to conceptualization and data curation. K.Y. contributed to conceptualization, supervision, investigation, and writing—original draft. All authors reviewed the manuscript.

Funding

This work was supported in part by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00337092), the Institute of Information & communications Technology Planning & Evaluation (IITP) grants (RS-2020-II201373, Artificial Intelligence Graduate School Program; RS-2023-00253914, Artificial Intelligence

Semiconductor (Hanyang University)) funded by the Korean government (MSIT), and in part by Samsung Electronics Co., Ltd.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.Y.

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.