

Explainable Multi-Class Osteoporosis Detection and Classification from Knee X-Ray Images Using an ROI-Guided Dual-Branch Deep Learning Framework

Chaitrashree R. S. and Dr. K. Thippeswamy

Department of Computer Science and Engineering VTU College, Mysore, Karnataka, India

Abstract—Osteoporosis is a metabolic skeletal disorder characterized by reduced bone mineral density, deterioration of bone microarchitecture, and increased fracture risk. Dual-energy X-ray absorptiometry is considered the clinical reference standard for osteoporosis diagnosis; however, its availability is limited in many rural and resource-constrained healthcare environments. Conventional radiographs are more widely available, but visual manifestations of early bone loss are subtle and may be difficult to identify consistently. This paper proposes an explainable region-of-interest-guided dual-branch deep learning framework for classifying knee X-ray images into normal, osteopenia, and osteoporosis categories. The proposed framework, named Osteoporosis ROI-guided Self-supervised Attention Network (ORSA-Net), combines bone-region segmentation, contrast-enhanced radiograph preprocessing, convolutional texture extraction, transformer-based global representation learning, and ordinal classification. A clinical-feature branch can additionally incorporate age, sex, body mass index, menopause status, fracture history, and other available risk factors. Class-balanced focal loss, ordinal consistency loss, supervised contrastive loss, and uncertainty-aware calibration are integrated to reduce class imbalance and improve the reliability of predictions. Gradient-weighted class activation mapping is used to identify image regions that influence each prediction. The proposed system is intended as an opportunistic screening and clinical decision-support tool and not as a replacement for dual-energy X-ray absorptiometry or specialist diagnosis. A patient-independent evaluation protocol with internal cross-validation and external dataset validation is described to improve reproducibility and generalizability.

Index Terms—osteoporosis, osteopenia, X-ray classification, deep learning, convolutional neural network, vision transformer, self-supervised learning, explainable artificial intelligence, bone mineral density

I. INTRODUCTION

Osteoporosis is a systemic skeletal disorder in which bone mineral density (BMD), bone mass, and bone microarchitecture deteriorate. This deterioration reduces bone strength and increases the probability of fragility fractures. The condition is frequently described as a silent disease because significant bone loss may occur before the patient experiences a fracture or other clear clinical symptom.

Dual-energy X-ray absorptiometry (DXA) is widely used to measure BMD and establish diagnostic categories through T-

scores. In general, a T-score greater than or equal to -1.0 is considered normal, a T-score between -1.0 and -2.5 indicates osteopenia, and a T-score less than or equal to -2.5 indicates osteoporosis. Although DXA is clinically important, routine screening may be affected by equipment cost, limited availability, lack of trained operators, and low patient participation. Conventional X-ray imaging is less expensive and more widely available than DXA. Knee, hand, pelvis, hip, chest, dental, and spinal radiographs are routinely acquired for several clinical indications. These images contain information related to cortical thickness, trabecular texture, bone shape, density distribution, and structural degradation. However, the radiographic signs of early osteoporosis may be subtle and can be influenced by exposure settings, scanner type, anatomical positioning, age, sex, and coexisting diseases.

Deep learning has demonstrated the ability to extract complex representations directly from radiographic images. Previous studies have used convolutional neural networks (CNNs), transfer learning, segmentation, attention mechanisms, multi-modal fusion, and self-supervised learning to predict BMD or identify osteoporosis from radiographs [1]–[8]. Despite promising results, several important challenges remain:

- Many studies use data from a single institution.
- Some studies perform only binary classification.
- Osteopenia, an important intermediate stage, is frequently excluded.
- Dataset size and class distributions are often limited.
- Patient-level separation is not always clearly described.
- Prediction confidence and uncertainty are rarely reported.
- External validation across different imaging devices is limited.
- Clinical deployment may be affected by poor interpretability.

To address these issues, this paper proposes ORSA-Net, an ROI-guided, dual-branch, explainable deep learning framework for three-class osteoporosis classification. The principal contributions are as follows:

- 1) A three-class classification framework for distinguishing normal, osteopenia, and osteoporosis cases.

- 2) Automated segmentation and ROI extraction to reduce the influence of irrelevant image regions.
- 3) A dual image branch combining local CNN texture features and global transformer representations.
- 4) Self-supervised pretraining to utilize unlabeled X-ray images and reduce dependence on large annotated datasets.
- 5) Ordinal learning that incorporates the medically meaningful progression from normal to osteopenia and osteoporosis.
- 6) Optional fusion of clinical risk factors with image features.
- 7) Class-balanced focal loss and supervised contrastive learning to address imbalance and overlapping classes.
- 8) Grad-CAM-based visual explanations and uncertainty-aware prediction calibration.
- 9) Patient-independent cross-validation and external testing to reduce data leakage and evaluate generalizability.

II. CLINICAL BACKGROUND

A. Bone Mineral Density and T-Score

BMD represents the mineral content per unit area of bone and is commonly measured in g/cm^2 . A T-score compares the patient's BMD with the mean BMD of a healthy young adult reference population:

$$T = \frac{BMD_{\text{patient}} - \mu_{\text{reference}}}{\sigma_{\text{reference}}}, \quad (1)$$

where $\mu_{\text{reference}}$ and $\sigma_{\text{reference}}$ denote the mean and standard deviation of the reference population.

The diagnostic categories used in this study are:

$$y = \{0, T > -1.0, \text{ Normal}, 1, -2.5 < T \leq -1.0, \text{ Osteopenia}, 2, T \leq -2.5, \text{ Osteoporosis}\} \quad (2)$$

The final labels should be created using verified DXA or quantitative ultrasound measurements associated with each patient rather than visual assumptions made from the X-ray image alone.

B. Radiographic Characteristics

Potential radiographic indicators of reduced bone density include:

- thinning of the cortical bone;
- reduced trabecular density;
- increased radiolucency;
- changes in trabecular orientation;
- widening of medullary regions;
- structural irregularity;
- loss of fine bone texture.

These changes may overlap between adjacent disease stages. Therefore, osteoporosis classification can be treated as both a multi-class and an ordinal learning problem.

III. LITERATURE REVIEW

Ho *et al.* developed DeepDXA to estimate femoral BMD from pelvis radiographs. Their framework used femur segmentation and CNN-based regression on paired pelvis X-ray and DXA examinations [1]. The study demonstrated that conventional radiographs can provide information correlated with DXA measurements. However, the model was developed using retrospective data from a single medical center.

Sato *et al.* proposed a deep learning model for predicting BMD and T-scores from chest radiographs [2]. Their research demonstrated the possibility of opportunistic screening using routinely acquired chest X-rays. Nevertheless, differences in hospital equipment, population distributions, and acquisition protocols can affect model generalization.

Golestan *et al.* estimated femoral-neck BMD using hand, knee, and pelvis radiographs from the Osteoarthritis Initiative dataset [3]. Image features were combined with age and sex. The study showed that peripheral anatomical regions may contain information useful for low-BMD screening. However, the selected cohort primarily consisted of patients with osteoarthritis, potentially limiting generalization to the broader population.

Boonrod *et al.* evaluated deep learning for osteoporosis screening from anteroposterior hip radiographs [4]. The method indicated that hip radiographs can be used to identify patients requiring further BMD testing. However, the retrospective design and dependence on a specific anatomical view can restrict wider application.

Asamoto *et al.* externally validated a chest-radiograph model for predicting femoral-neck BMD [5]. External validation is an important contribution because it evaluates performance on images acquired outside the development facility. Nevertheless, differences between scanners and patient groups continue to influence prediction accuracy.

Hwang *et al.* proposed osteoporosis prediction from hand X-rays using uncertainty-aware segmentation and self-supervised learning [6]. Their method segmented the ulna, radius, and metacarpal bones before classification. Although the segmentation-for-classification strategy improved anatomical focus, the reported dataset included only 192 individuals, limiting statistical strength.

Yoshida *et al.* used frontal and lateral chest radiographs with Inception-ResNet-V2 models to estimate BMD [7]. Bidirectional chest images improved performance compared with individual frontal or lateral views. However, paired radiographic views are not always available, and the reported specificity at high sensitivity indicated room for improvement.

Tang *et al.* proposed a multimodal osteoporosis model that fused chest X-rays and clinical features [8]. The model included wavelet features and attention-based fusion and improved the area under the receiver operating characteristic curve compared with an image-only model. The study was retrospective, lacked broad external validation, and used data from one hospital.

TABLE I
LITERATURE REVIEW OF DEEP LEARNING METHODS FOR OSTEOPOROSIS SCREENING

No.	Authors	Year	Title/Method	Main Contribution	Limitation and Research Gap
1	Ho <i>et al.</i>	2021	Application of Deep Learning Neural Network in Predicting BMD from Plain X-Ray Radiography	DeepDXA used femur segmentation and CNN regression to estimate BMD from pelvis radiographs.	Single-center retrospective data; primarily focused on BMD regression; limited cross-device and multi-population validation.
2	Sato <i>et al.</i>	2022	Deep Learning for BMD and T-Score Prediction from Chest X-Rays	Predicted BMD and T-score using routinely acquired chest radiographs from multiple hospitals.	Possible scanner and population bias; limited evidence of prospective clinical implementation and calibrated uncertainty.
3	Golestan <i>et al.</i>	2023	Approximating Femoral-Neck BMD from Hand, Knee, and Pelvis X-Rays Using Deep Learning	Combined radiographic features with age and sex to predict femoral-neck BMD from multiple anatomical sites.	OAI cohort was enriched with osteoarthritis cases; limited osteoporosis prevalence and restricted population diversity.
4	Boonrod <i>et al.</i>	2024	Deep Learning for Osteoporosis Screening Using Anteroposterior Hip Radiographs	Demonstrated screening of osteoporosis from standard hip radiographs.	Restricted to a single projection and retrospective cohort; limited model interpretability and external validation.
5	Asamoto <i>et al.</i>	2024	External Validation of a Deep Learning Model for Predicting BMD on Chest Radiographs	Evaluated a BMD prediction model at facilities outside the model-development environment.	Performance can vary among devices and populations; clinical threshold selection and uncertainty require further study.
6	Hwang <i>et al.</i>	2024	Osteoporosis Prediction from Hand X-Ray Images Using Segmentation-for-Classification and Self-Supervised Learning	Used uncertainty-aware bone segmentation and self-supervised representation learning for osteoporosis prediction.	Small cohort of 192 individuals; hand-specific framework; further multicenter evaluation is necessary.
7	Yoshida <i>et al.</i>	2025	Utility of Osteoporosis Screening Using Bidirectional Chest Radiographs with Deep Learning	Combined frontal and lateral chest images to improve BMD estimation and osteoporosis screening.	Requires paired views; specificity remained limited at high-sensitivity operating points; not designed for knee X-rays.
8	Tang <i>et al.</i>	2025	Fusion of X-Ray Images and Clinical Data for a Multimodal Deep Learning Prediction Model of Osteoporosis	Fused chest X-rays, clinical data, wavelet features, CNN representations, and attention mechanisms.	Single-center retrospective study; possible selection bias; lack of broad external validation and prospective testing.

IV. RESEARCH GAP

The literature review identifies the following unresolved problems:

- 1) **Binary prediction:** Several studies differentiate only normal and low-BMD cases, while osteopenia is either merged with osteoporosis or ignored.
- 2) **Limited anatomical focus:** Whole-image classification can allow the model to learn labels, positioning markers, implants, image borders, and scanner-specific patterns instead of bone characteristics.
- 3) **Small labeled datasets:** Medical image annotation requires DXA or clinically verified labels, limiting the number of usable samples.
- 4) **Class imbalance:** Osteopenia, normal, and osteoporosis images may not be equally represented.
- 5) **Single-feature representation:** CNN-only systems may capture local texture effectively but can be less effective at modeling long-range relationships.
- 6) **Weak ordinal modeling:** Normal, osteopenia, and osteoporosis have a natural severity order that is not incorporated by ordinary categorical cross-entropy.
- 7) **Insufficient uncertainty reporting:** A high-confidence

but incorrect prediction can negatively affect clinical trust.

- 8) **Limited explainability:** Many models do not clearly demonstrate whether predictions are based on anatomically meaningful bone regions.
- 9) **Data leakage risk:** Splitting individual images rather than patients can place radiographs belonging to the same patient in training and testing sets.
- 10) **Limited external validation:** Models are frequently evaluated only on internal test data.

V. PROPOSED ORSA-NET FRAMEWORK

A. System Overview

The proposed system accepts a knee X-ray image and optional clinical information. Its output consists of:

- predicted class: normal, osteopenia, or osteoporosis;
- confidence score for each class;
- uncertainty estimate;
- Grad-CAM explanation;
- referral suggestion for uncertain or high-risk cases.

The complete workflow contains the following stages:

- 1) image quality checking;

- 2) patient-independent dataset splitting;
- 3) intensity normalization and contrast enhancement;
- 4) bone-region segmentation;
- 5) ROI extraction;
- 6) self-supervised encoder pretraining;
- 7) local CNN feature extraction;
- 8) global transformer feature extraction;
- 9) optional clinical-feature encoding;
- 10) attention-guided feature fusion;
- 11) ordinal three-class prediction;
- 12) uncertainty calibration and explainability.

B. Image Preprocessing

Each input image is converted into a standard grayscale representation. Patient identifiers, textual annotations, and empty borders are removed wherever possible. The image is resized while preserving the anatomical aspect ratio.

Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to improve the visibility of cortical and trabecular structures without excessively amplifying noise. Let I represent the original image. The preprocessed image is:

$$I_p = N(C(I)), \quad (3)$$

where C denotes CLAHE and N denotes intensity normalization.

Pixel values are normalized using:

$$I_n = \frac{I_p - \mu_I}{\sigma_I + \epsilon}, \quad (4)$$

where μ_I and σ_I are the image mean and standard deviation, and ϵ prevents division by zero.

Recommended augmentation operations include:

- rotation within $\pm 10^\circ$;
- horizontal flipping when anatomically valid;
- translation and mild scaling;
- brightness and contrast modification;
- Gaussian noise;
- random erasing outside important bone regions;
- MixUp and CutMix with controlled probabilities.

Vertical flipping should generally be avoided because it can produce anatomically unrealistic knee images.

C. Bone ROI Segmentation

A lightweight U-Net or Attention U-Net is trained to segment the knee bone region. Let the segmentation network be S_{θ_s} :

$$M = S_{\theta_s}(I_n), \quad (5)$$

where M is the predicted binary bone mask.

The segmentation objective combines binary cross-entropy and Dice loss:

$$L_{seg} = \lambda_{bce} L_{BCE} + \lambda_{dice} L_{Dice}. \quad (6)$$

Dice loss is calculated as:

$$L_{Dice} = 1 - \frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon}, \quad (7)$$

where p_i and g_i denote predicted and ground-truth mask values.

The ROI image is obtained using:

$$I_{roi} = I_n \odot M, \quad (8)$$

where $\odot$ denotes element-wise multiplication. A small margin is retained around the segmented bone to preserve contextual information.

D. Self-Supervised Pretraining

A large number of X-ray images may be available without osteoporosis labels. Therefore, self-supervised contrastive learning is used to initialize the image encoder.

For two augmented views v_i and v_j of the same X-ray, the normalized temperature-scaled cross-entropy loss is:

$$L_{SSL} = -\log \sum_{k=i} \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\exp(\text{sim}(z_i, z_k)/\tau)}, \quad (9)$$

where z_i and z_j are projected representations, sim denotes cosine similarity, and τ is the temperature parameter.

Self-supervised pretraining is expected to improve feature quality when the labeled dataset is small.

E. Local CNN Branch

The local feature branch uses EfficientNet-B0, ConvNeXt-Tiny, or ResNet50 as its backbone. The CNN branch focuses on local characteristics such as cortical thickness, edges, texture, and trabecular patterns.

$$F_{local} = E_{cnn}(I_{roi}), \quad (10)$$

where E_{cnn} is the CNN encoder and F_{local} is the local feature vector.

ConvNeXt-Tiny is recommended because it provides strong convolutional representation while retaining computational efficiency. EfficientNet-B0 can be used when the deployment environment has limited memory.

F. Global Transformer Branch

The global branch uses a Swin Transformer or compact Vision Transformer. The ROI is divided into non-overlapping patches. For each patch x_p , a patch embedding is generated:

$$Z_0 = [x_{class}; x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos}, \quad (11)$$

where E is the patch projection matrix and E_{pos} is positional information.

The transformer branch models long-distance interactions between different portions of the knee:

$$F_{global} = E_{vit}(I_{roi}). \quad (12)$$

The CNN and transformer branches are complementary: the CNN captures local texture while the transformer captures global structural relationships.

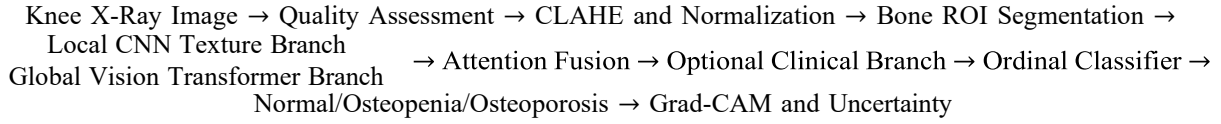


Fig. 1. High-level architecture of the proposed ORSA-Net framework. Replace this conceptual block with a professionally drawn architecture figure in the final paper.

G. Clinical Feature Branch

When clinical variables are available, the following factors can be included:

- age;
- sex;
- body mass index;
- menopause status and menopause age;
- fracture history;
- smoking and alcohol history;
- physical activity;
- calcium and vitamin D information;
- diabetes;
- thyroid disease;
- family history;
- medication history.

Continuous variables are standardized, and categorical variables are one-hot encoded. Missing values can be processed using median imputation and an additional missingness indicator.

The clinical representation is:

$$F_{clinical} = E_{mlp}(C), \quad (13)$$

where C denotes the clinical feature vector.

The proposed model must also support an image-only mode so that it can operate when clinical features are unavailable.

H. Attention-Based Feature Fusion

The image features are concatenated:

$$F_i = [F_{local}; F_{global}]. \quad (14)$$

A gating function determines the contribution of each branch:

$$G = \sigma(W_g F_i + b_g), \quad (15)$$

$$F_{image} = G \odot F_{local} + (1 - G) \odot F_{global}. \quad (16)$$

When clinical information is available:

$$F_{fusion} = \text{LayerNorm}([F_{image}; F_{clinical}]). \quad (17)$$

When clinical data are unavailable:

$$F_{fusion} = F_{image}. \quad (18)$$

I. Ordinal Multi-Class Classification

Standard three-class classification treats all class errors equally. However, predicting osteoporosis as osteopenia is less severe than predicting osteoporosis as normal. Therefore, an ordinal classification head is used.

For three ordered classes, two threshold probabilities are estimated:

$$q_1 = P(y > 0 | F_{fusion}), \quad (19)$$

$$q_2 = P(y > 1 | F_{fusion}). \quad (20)$$

The class probabilities can be derived as:

$$P(y = 0) = 1 - q_1, \quad (21)$$

$$P(y = 1) = q_1 - q_2, \quad (22)$$

$$P(y = 2) = q_2. \quad (23)$$

A monotonicity constraint is applied so that:

$$q_1 \geq q_2. \quad (24)$$

This design incorporates disease progression directly into the learning process.

J. Class-Balanced Focal Loss

For class c , the effective-number class weight is:

$$w_c = \frac{1 - \beta}{1 - \beta^{n_c}}, \quad (25)$$

where n_c is the number of training samples in class c , and β is generally close to one.

The focal loss is:

$$L_{focal} = - \sum_{c=1}^C w_c (1 - p_c)^{\gamma} y_c \log(p_c), \quad (26)$$

where γ controls the focus on difficult examples.

K. Supervised Contrastive Loss

Supervised contrastive learning pulls representations from the same disease class closer and pushes representations of different classes apart:

$$L_{supcon} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \sum_{d=i} \frac{\exp(z_i \cdot z_p / \tau)}{\exp(z \cdot z_q / \tau)}. \quad (27)$$

This loss is useful because normal, osteopenia, and osteoporosis images may have overlapping visual characteristics.

L. Combined Objective Function

The total loss is:

$$L_{total} = \lambda_1 L_{focal} + \lambda_2 L_{ordinal} + \lambda_3 L_{supcon} + \lambda_4 L_{seg} + \lambda_5 L_{consistency}. \quad (28)$$

Here:

- L_{focal} handles class imbalance;
- $L_{ordinal}$ preserves disease order;
- L_{supcon} improves feature separation;
- L_{seg} supervises bone segmentation;
- $L_{consistency}$ encourages consistent predictions under valid augmentations.

Suggested initial weights are:

$$\lambda_1 = 1.0, \quad \lambda_2 = 0.5, \quad \lambda_3 = 0.2, \quad \lambda_4 = 0.3, \quad \lambda_5 = 0.1. \quad (29)$$

These values must be tuned using the validation set rather than the final test set.

M. Uncertainty Estimation

Monte Carlo dropout or a deep ensemble can be used to estimate predictive uncertainty. For T stochastic forward passes:

$$\bar{p} = \frac{1}{T} \sum_{t=1}^T p_t. \quad (30)$$

Predictive entropy is:

$$H(\bar{p}) = - \sum_{c=1}^C \bar{p}_c \log(\bar{p}_c). \quad (31)$$

If entropy exceeds a selected threshold or maximum class confidence is below a safe threshold, the system should output an uncertain result and recommend DXA assessment rather than making a definite classification.

N. Explainability

Grad-CAM is applied to the final convolutional feature maps. The importance weight for feature map A^k is:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad (32)$$

and the Grad-CAM heatmap is:

$$L_{Grad-CAM}^c = \text{ReLU} \sum_k \alpha_k^c A^k. \quad (33)$$

A valid model should primarily focus on anatomically meaningful cortical and trabecular areas rather than labels, borders, or image artifacts.

VI. PROPOSED ALGORITHM

VII. DATASET DESCRIPTION

A. Primary Knee X-Ray Osteoporosis Database

The Knee X-Ray Osteoporosis Database published through Mendeley Data can be used as a primary source. It contains knee X-ray images, clinical risk factors, and T-score-related information. The clinical information includes variables such as age, sex, menopause information, diabetes, thyroid disease, fracture history, and lifestyle factors.

Dataset link:

<https://data.mendeley.com/datasets/fxjm8fb6mw/2>

The dataset is distributed under a CC BY 4.0 license. Researchers must inspect the files, documentation, labels, duplicate images, and participant identifiers before using the data.

B. Multi-Class Knee Osteoporosis X-Ray Dataset

A prepared three-class dataset containing normal, osteopenia, and osteoporosis knee radiographs is available through Zenodo. The published description reports 1,947 knee X-ray images, including 780 normal, 374 osteopenia, and 793 osteoporosis images.

Dataset link:

<https://zenodo.org/records/20053147>

The dataset documentation states that augmentation and balancing were used in preparing the released data. Therefore, researchers should identify original and augmented files carefully. Augmented versions derived from the same source image must not be distributed across training, validation, and test partitions.

C. Kaggle Dataset

A multi-class knee osteoporosis dataset is also listed on Kaggle:

<https://www.kaggle.com/datasets/mohamedgobara/multi-class-knee-osteoporosis-x-ray-dataset>

Kaggle data should be used only after checking:

Algorithm 1 Training of the Proposed ORSA-Net Model

Require: Patient-level X-ray dataset D , optional clinical data C , DXA/QUS-derived labels Y

Ensure: Trained segmentation model S_{θ_s} and osteoporosis classifier F_{θ}

```

1: Remove corrupted, duplicate, and non-diagnostic images
2: De-identify images and remove textual annotations where possible
3: Group all images by unique patient identifier
4: Split patients into training, validation, and testing groups
5: Compute class weights using only the training group
6: for each training X-ray  $I$  do
7:   Convert  $I$  to standardized grayscale format
8:   Crop empty borders and resize with aspect ratio preserved
9:   Apply CLAHE and intensity normalization
10:  Apply medically valid augmentation
11: end for
12: Train Attention U-Net using manually annotated bone masks
13: Generate bone mask  $M = S_{\theta_s}(I)$ 
14: Obtain ROI image  $I_{roi} = I \odot M$ 
15: Pretrain image encoder using self-supervised contrastive learning
16: for epoch = 1 to  $E$  do
17:   for each mini-batch  $(I_{roi}, C, Y)$  do
18:      $F_{local} \leftarrow E_{cnn}(I_{roi})$ 
19:      $F_{global} \leftarrow E_{vit}(I_{roi})$ 
20:     if clinical data are available then
21:        $F_{clinical} \leftarrow E_{mlp}(C)$ 
22:        $F_{fusion} \leftarrow \text{AttentionFusion}(F_{local}, F_{global}, F_{clinical})$ 
23:     else
24:        $F_{fusion} \leftarrow \text{AttentionFusion}(F_{local}, F_{global})$ 
25:     end if
26:      $\hat{Y} \leftarrow \text{OrdinalClassifier}(F_{fusion})$ 
27:     Calculate class-balanced focal loss
28:     Calculate ordinal consistency loss
29:     Calculate supervised contrastive loss
30:     Calculate augmentation consistency loss
31:     Combine losses into  $L_{total}$ 
32:     Update model parameters using AdamW
33:   end for
34:   Evaluate macro F1-score on validation patients
35:   Save model when validation macro F1-score improves
36:   Apply early stopping when performance does not improve
37: end for
38: Calibrate saved model using temperature scaling
39: Select referral threshold using validation data
40: Evaluate once on untouched internal and external test sets
41: Generate Grad-CAM maps and uncertainty scores

```

Algorithm 2 Inference and Clinical Referral Decision

Require: New knee X-ray I and optional clinical variables C

Ensure: Predicted class, probabilities, uncertainty, and explanation

```

1: Check image resolution, orientation, and exposure quality
2: if image quality is unacceptable then
3:   return "Image quality insufficient; reacquire X-ray"
4: end if
5: Preprocess and normalize image
6: Generate bone mask and extract ROI
7: Run  $T$  stochastic forward passes
8: Compute mean probabilities  $\bar{p}$ 
9: Compute predictive entropy  $H(\bar{p})$ 
10: Generate Grad-CAM heatmap
11:  $k \leftarrow \arg \max_c \bar{p}_c$ 
12: if  $H(\bar{p}) > \delta_H$  then
13:   return "Uncertain result; recommend expert review/DXA"
14: else if  $k = \text{Osteoporosis}$  then
15:   return "High-risk screening result; recommend DXA"
16: else if  $k = \text{Osteopenia}$  then
17:   return "Intermediate risk; consider clinical risk assessment"
18: else
19:   return "Normal screening result; interpret with clinical factors"
20: end if

```

- original source and data license;
- presence of duplicate images;
- whether images are original or augmented;
- availability of patient identifiers;
- label-generation methodology;
- possible overlap with Mendeley or Zenodo data.

D. Osteoarthritis Initiative Dataset

The Osteoarthritis Initiative (OAI) is a large longitudinal collection containing knee, hand, pelvis, clinical, and DXA-related data. Access may require registration and acceptance of data-use conditions.

Dataset information:

<https://nda.nih.gov/oai/>

The OAI dataset can support external validation or BMD regression. However, it was designed mainly for osteoarthritis research. This population characteristic must be discussed when interpreting osteoporosis results.

E. MURA Dataset for Self-Supervised Pretraining

MURA contains 40,561 musculoskeletal radiographs from 14,863 studies and 12,173 patients. It includes elbow, finger, forearm, hand, humerus, shoulder, and wrist images [9]. MURA does not provide osteoporosis labels and must not be used directly as the labelled osteoporosis test set. It may

be used for self-supervised musculoskeletal representation learning.

Dataset link:

<https://stanfordmlgroup.github.io/competitions/mura/>

VIII. DATASET PREPARATION PROTOCOL

A. Data Cleaning

The following cleaning steps should be completed before training:

- 1) remove unreadable and corrupted images;
- 2) verify image-label correspondence;
- 3) remove exact duplicates using file hashes;
- 4) detect near duplicates using perceptual hashing;
- 5) identify augmented images generated from the same original image;
- 6) remove images with severe cropping or missing anatomy;
- 7) review implants, fractures, and extreme deformities;
- 8) verify class labels using T-score information;
- 9) maintain a data-cleaning audit report.

B. Patient-Level Data Splitting

All radiographs belonging to one patient must remain in the same subset. A recommended split is:

- training: 70%;
- validation: 15%;
- internal testing: 15%.

Alternatively, five-fold stratified group cross-validation can be used. The patient identifier must be supplied as the grouping variable.

Let P_{train} , P_{val} , and P_{test} represent the patient sets. The following condition must hold:

$$P_{train} \cap P_{val} = P_{train} \cap P_{test} = P_{val} \cap P_{test} = \emptyset \quad (34)$$

C. External Testing

The external test dataset must not be used for:

- model training;
- hyperparameter selection;
- early stopping;
- threshold selection;
- calibration.

It should be evaluated only after the model and operating threshold are finalized.

IX. EXPERIMENTAL SETUP

A. Hardware and Software

A possible implementation environment is:

- Python 3.10 or later;
- PyTorch 2.x;
- MONAI or TorchVision;
- OpenCV;
- Albumentations;

- scikit-learn;
- CUDA-compatible GPU with at least 8 GB memory;
- MLflow or Weights & Biases for experiment tracking.

B. Suggested Hyperparameters

Hyperparameters should be selected using validation data or nested cross-validation. The final test set must not be repeatedly inspected during model development.

C. Baseline Models

The proposed method should be compared with:

- 1) custom CNN;
- 2) ResNet50;
- 3) DenseNet121;
- 4) EfficientNet-B0;
- 5) ConvNeXt-Tiny;
- 6) Swin Transformer-Tiny;
- 7) CNN with clinical-feature concatenation;
- 8) proposed dual-branch ORSA-Net;
- 9) proposed ORSA-Net without segmentation;
- 10) proposed ORSA-Net without ordinal loss;
- 11) proposed ORSA-Net without self-supervised pretraining.

X. EVALUATION METRICS

Accuracy alone is not sufficient for an imbalanced medical dataset. The following metrics should be reported.

A. Precision, Sensitivity, and Specificity

$$Precision = \frac{TP}{TP + FP} \quad (35)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (36)$$

$$Specificity = \frac{TN}{TN + FP} \quad (37)$$

B. F1-Score

$$F1 = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \quad (38)$$

Macro F1-score should be the primary classification metric because it gives equal importance to all three classes.

C. Balanced Accuracy

$$BalancedAccuracy = \frac{1}{K} \sum_{c=1}^K Recall_c \quad (39)$$

D. Area Under the ROC Curve

One-vs-rest AUC should be calculated for each class, along with macro-averaged AUC.

TABLE II
CANDIDATE DATASETS FOR THE PROPOSED STUDY

Dataset	Approximate Size	Data Type	Suggested Use	Important Consideration
Knee X-Ray Osteoporosis Database	350 images reported in the derived dataset documentation	Knee X-rays, clinical variables, T-score information	Primary supervised training or multimodal experiments	Small and imbalanced; inspect patient IDs, labels, and duplicate images.
Multi-Class Knee Osteoporosis X-Ray Dataset	1,947 images	Normal, osteopenia, and osteoporosis knee X-rays	Three-class image classification	Contains prepared or augmented samples; source-level splitting is necessary.
Osteoarthritis Initiative	Large longitudinal cohort	Knee, hand, pelvis, DXA, and clinical data	External validation or BMD regression	Population is enriched for osteoarthritis risk and may not represent general screening populations.
MURA	40,561 images	Musculoskeletal radiographs	Self-supervised pretraining	Contains abnormal/normal labels, not osteoporosis-specific labels.

TABLE III
SUGGESTED INITIAL HYPERPARAMETERS

Parameter	Initial Value
Input size	384×384
Batch size	8–32
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Epochs	80–150
Scheduler	Cosine annealing
Warm-up epochs	5
Focal-loss γ	2.0
Dropout	0.2–0.4
Early-stopping patience	15 epochs
Mixed precision	Enabled
MC-dropout passes	20–30

E. Quadratic Weighted Kappa

Because the classes are ordered, quadratic weighted kappa is useful for penalizing severe errors more heavily:

$$\kappa_w = 1 - \frac{\sum_{ij} w_{ij} O_{ij}}{\sum_{ij} w_{ij} E_{ij}}. \quad (40)$$

F. Calibration Metrics

The Brier score is:

$$BS = \frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C (p_{ic} - y_{ic})^2. \quad (41)$$

Expected calibration error should also be reported. Reliability diagrams can be used to compare confidence and observed correctness.

G. Confidence Intervals

Performance estimates should include 95% confidence intervals obtained through patient-level bootstrapping. Statistical comparisons between models can be conducted using:

- DeLong test for AUC;
- McNemar test for paired classification outcomes;

- bootstrap confidence intervals for macro F1-score;
- permutation testing for ablation studies.

XI. POSSIBLE AND EXPECTED OUTCOMES

The following values are proposed research targets and must not be reported as achieved results before conducting experiments:

- internal-test macro F1-score greater than 0.88;
- balanced accuracy between 0.87 and 0.93;
- macro AUC greater than 0.93;
- osteoporosis sensitivity greater than 0.90;
- osteopenia sensitivity greater than 0.82;
- quadratic weighted kappa greater than 0.80;
- expected calibration error below 0.05 after calibration;
- external-test performance reduction below 8 percentage points;
- inference time below 500 ms per image on a GPU;
- clinically relevant Grad-CAM localization in the majority of correctly classified cases.

A plausible experimental results table is provided as a template. All entries must be replaced with actual measurements.

The osteopenia category is expected to be the most difficult class because it lies between normal bone density and osteoporosis. Misclassification analysis should therefore focus on normal-to-osteopenia and osteopenia-to-osteoporosis transitions.

XII. ABLATION STUDY

The contribution of each proposed component should be evaluated independently.

The following hypotheses can be tested:

- ROI segmentation improves robustness by removing irrelevant background.
- Self-supervised pretraining improves performance when labeled data are limited.
- Transformer fusion improves global structural modeling.
- Ordinal loss reduces severe normal-to-osteoporosis errors.
- Clinical fusion improves classification of borderline osteopenia cases.

TABLE IV
TEMPLATE FOR EXPERIMENTAL RESULTS. REPLACE DASHES WITH ACTUAL VALUES

Model	Accuracy	Balanced Acc.	Macro Precision	Macro Recall	Macro F1	Macro AUC
ResNet50	–	–	–	–	–	–
DenseNet121	–	–	–	–	–	–
EfficientNet-B0	–	–	–	–	–	–
ConvNeXt-Tiny	–	–	–	–	–	–
Swin Transformer-Tiny	–	–	–	–	–	–
ORSA-Net without clinical data	–	–	–	–	–	–
ORSA-Net with clinical data	–	–	–	–	–	–

TABLE V
TEMPLATE FOR CLASS-WISE PERFORMANCE

Class	Precision	Recall	F1	AUC
Normal	–	–	–	–
Osteopenia	–	–	–	–
Osteoporosis	–	–	–	–

- Temperature scaling improves probability calibration without changing accuracy.

XIII. NOVELTY OF THE PROPOSED WORK

The novelty of ORSA-Net is based on the integration of several complementary strategies specifically designed for three-stage osteoporosis screening:

- 1) **ROI-guided hybrid representation:** A segmentation model restricts feature extraction to relevant bone anatomy before processing the image through CNN and transformer branches.
- 2) **Local-global feature combination:** The CNN branch models local cortical and trabecular texture, while the transformer branch models global structural relationships.
- 3) **Three-class ordinal learning:** Unlike binary models, the proposed system distinguishes normal, osteopenia, and osteoporosis while preserving their natural severity order.
- 4) **Self-supervised musculoskeletal pretraining:** Unlabeled X-rays can be used to learn domain-specific features before supervised fine-tuning.
- 5) **Optional multimodal operation:** The system can operate using only X-rays or combine image representations with clinical risk factors.
- 6) **Uncertainty-aware referral:** Uncertain cases are flagged for DXA or specialist review rather than being assigned an overconfident prediction.
- 7) **Explainability validation:** Grad-CAM localization is evaluated against the segmented bone region to verify that the model uses anatomically meaningful evidence.
- 8) **Patient-independent and external evaluation:** The protocol explicitly prevents patient overlap and includes testing on a dataset not used during model development.

The novelty claim should be refined after conducting a final systematic literature search. It should not claim that no

previous study has ever combined any two of these components. Instead, the contribution is the unified, experimentally evaluated framework for explainable three-class knee X-ray osteoporosis screening.

XIV. EXPECTED IMPROVEMENT OVER EXISTING LITERATURE

XV. ETHICAL, PRIVACY, AND CLINICAL CONSIDERATIONS

The system should be developed using de-identified data. Institutional ethics approval and informed-consent requirements depend on the source of the data and local regulations. Public dataset availability does not remove the obligation to comply with the corresponding license and data-use agreement.

The following safeguards are recommended:

- remove personal identifiers from DICOM metadata;
- store images using encrypted systems;
- restrict access using role-based authorization;
- maintain audit logs;
- evaluate performance across age and sex subgroups;
- assess scanner and institution bias;
- avoid autonomous diagnosis;
- display a clear decision-support disclaimer;
- refer uncertain and high-risk patients for DXA;
- obtain radiologist review before clinical deployment.

Because osteoporosis labels are related to DXA or other reference measurements, the proposed model should be described as a screening system. A normal model prediction must not be used to rule out osteoporosis when the patient has significant clinical risk factors.

XVI. LIMITATIONS OF THE PROPOSED STUDY

The proposed method has several anticipated limitations:

- 1) Public datasets may be small, imbalanced, or derived from a limited geographic population.
- 2) Some released datasets may contain augmented rather than original images.
- 3) Knee radiographs are affected by osteoarthritis, implants, fractures, positioning, and exposure.
- 4) BMD measured at the hip or lumbar spine may not perfectly correspond to the visible knee region.
- 5) Self-supervised learning increases training complexity.
- 6) Segmentation errors may propagate to the classification stage.

TABLE VI
TEMPLATE FOR ABLATION STUDY

Configuration	ROI Seg.	SSL	Transformer	Ordinal Loss	Macro F1
CNN baseline	No	No	No	No	–
CNN + ROI	Yes	No	No	No	–
CNN + ROI + SSL	Yes	Yes	No	No	–
Dual branch + ROI	Yes	No	Yes	No	–
Dual branch + ROI + SSL	Yes	Yes	Yes	No	–
Complete ORSA-Net	Yes	Yes	Yes	Yes	–

TABLE VII
COMPARISON OF LITERATURE LIMITATIONS AND PROPOSED IMPROVEMENTS

Literature Limitation	Effect on Existing Systems	Proposed Improvement
Binary classification	Osteopenia may be merged with another class, reducing early-stage clinical usefulness.	Three-class ordinal classification for normal, osteopenia, and osteoporosis.
Whole-image learning	Model may learn text labels, borders, scanner artifacts, or unrelated anatomy.	Bone segmentation and ROI-guided classification.
CNN-only feature extraction	Local texture is captured, but long-distance structural relationships may be underrepresented.	Dual CNN-transformer feature extraction.
Small labelled datasets	Increases overfitting and reduces generalization.	Self-supervised pretraining using unlabeled musculoskeletal radiographs.
Class imbalance	Majority classes dominate the optimization process.	Class-balanced focal loss and supervised contrastive learning.
No severity-aware loss	A severe error may receive the same penalty as an adjacent-class error.	Ordinal threshold learning and severity-aware loss.
Limited interpretability	Clinicians cannot verify the anatomical basis of predictions.	Grad-CAM, ROI-overlap analysis, and explanation reporting.
Uncalibrated probabilities	Prediction confidence may not represent actual correctness.	Temperature scaling, Brier score, calibration curves, and uncertainty estimation.
Internal testing only	Model may fail on another scanner, institution, or population.	External dataset testing and scanner-wise subgroup analysis.
Image-only prediction	Useful clinical risk factors are ignored.	Optional fusion of demographic and clinical variables.

- 7) Clinical variables may be missing or inconsistently recorded.
- 8) External validation cannot fully replace prospective clinical evaluation.
- 9) Grad-CAM provides supportive visualization but does not prove causal reasoning.

These limitations should be discussed transparently, and model performance should not be generalized beyond the evaluated populations.

XVII. FUTURE WORK

Future improvements can include:

- prospective multicenter clinical evaluation;
- validation using Indian and geographically diverse populations;
- direct BMD and T-score regression;
- joint osteoporosis and fracture-risk prediction;
- federated learning across hospitals;
- domain adaptation for different X-ray devices;
- vision-language reporting;
- model quantization for mobile or edge deployment;
- integration with hospital PACS systems;
- comparison with radiologists and orthopedic specialists;

- longitudinal prediction of osteoporosis progression;
- calibration monitoring after deployment.

XVIII. CONCLUSION

This paper presented ORSA-Net, an explainable ROI-guided dual-branch deep learning framework for detecting and classifying osteoporosis from knee X-ray images. The framework classifies radiographs into normal, osteopenia, and osteoporosis categories and combines bone segmentation, self-supervised learning, CNN-based local feature extraction, transformer-based global representation learning, ordinal classification, optional clinical-feature fusion, and uncertainty estimation.

The proposed method addresses several limitations identified in previous studies, including binary-only prediction, limited anatomical localization, class imbalance, absence of severity-aware learning, weak confidence calibration, and insufficient external validation. Grad-CAM explanations and uncertainty-based referral logic are included to improve transparency and clinical safety.

The system is intended for low-cost opportunistic screening and clinical decision support, particularly in settings where DXA equipment is not readily available. However, it should

not replace DXA measurements, radiologist interpretation, or specialist assessment. Actual clinical usefulness must be established through patient-independent testing, external validation, prospective studies, and ethical approval.

REFERENCES

- [1] C. S. Ho *et al.*, "Application of deep learning neural network in predicting bone mineral density from plain X-ray radiography," *Archives of Osteoporosis*, vol. 16, 2021, doi: 10.1007/s11657-021-00985-8.
- [2] Y. Sato *et al.*, "Deep learning for bone mineral density and T-score prediction from chest X-rays: A multicenter study," 2022. Available through PubMed, PMID: 36140424.
- [3] K. Golestan, C. A. Syme, A. Bilbily, S. Zuberi, M. Volkovs, T. Poutanen, and M. D. Cicero, "Approximating femoral neck bone mineral density from hand, knee, and pelvis X-rays using deep learning," *Journal of Medical Artificial Intelligence*, vol. 6, 2023, doi: 10.21037/jmai-23-10.
- [4] A. Boonrod *et al.*, "Deep learning for osteoporosis screening using anteroposterior hip radiographs," 2024. Available through PubMed, PMID: 38896146.
- [5] T. Asamoto *et al.*, "External validation of a deep learning model for predicting bone mineral density on chest radiographs," 2024. Available through PubMed, PMID: 38472499.
- [6] U. Hwang, C.-H. Lee, and K. Yoon, "Osteoporosis prediction from hand X-ray images using segmentation-for-classification and self-supervised learning," *arXiv preprint arXiv:2412.05345*, 2024.
- [7] Y. Yoshida, Y. Sato, Y. Kai, Y. Hirono, T. Sato, and T. Kasai, "Utility of osteoporosis screening based on estimation of bone mineral density using bidirectional chest radiographs with deep learning models," 2025. Available through PubMed, PMID: 40206487.
- [8] J. Tang, X. Yin, J. Lai, K. Luo, and D. Wu, "Fusion of X-ray images and clinical data for a multimodal deep learning prediction model of osteoporosis: Algorithm development and validation study," *JMIR Medical Informatics*, 2025, doi: 10.2196/70738.
- [9] P. Rajpurkar *et al.*, "MURA: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.
- [10] I. M. Wani and S. Arora, "Knee X-ray osteoporosis database," Mendeley Data, Version 2, 2021. [Online]. Available: <https://data.mendeley.com/datasets/fxjm8fb6mw/2>.
- [11] "Multi-class knee osteoporosis X-ray dataset," Zenodo, 2026. [Online]. Available: <https://zenodo.org/records/20053147>.
- [12] International Osteoporosis Foundation, "Osteoporosis information and resources." [Online]. Available: <https://www.osteoporosis.foundation/>.
- [13] National Institute of Arthritis and Musculoskeletal and Skin Diseases, "Osteoporosis: Causes, risk factors, and symptoms." [Online]. Available: <https://www.niams.nih.gov/health-topics/osteoporosis>.