International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 1-13
Journal homepage: http://iieta.org/journals/ts

ViT-FESEL: A Multi-Stage Vision Transformer-Based Feature Extraction and Selection Framework for Osteoporosis Classification from Knee X-Ray Images

Hasan Genç[1], Fatih Özyurt[2*]

[1] Elazig Fethi Sekin City Hospital, Elazığ 23280, Türkiye
[2] Software Engineering, Faculty of Engineering, Firat University, Elazığ 23119, Türkiye

Corresponding Author Email: fatihozyurt@firat.edu.tr

https://doi.org/10.18280/ts.430101

Received: 1 January 2026
Revised: 2 February 2026
Accepted: 24 February 2026
Available online: 28 February 2026

Keywords:
vision transformers, feature selection, ensemble, multi-stage, osteoporosis

ABSTRACT

Early and accurate detection of osteoporosis is clinically critical for reducing fracture risk and enabling appropriate treatment planning. In this study, a multi-stage, multi-branch framework, Vision Transformer–Based Feature Extraction and Selection (ViT-FESEL), is proposed that integrates Vision Transformer (ViT)-based representation learning with a systematic feature extraction and selection strategy for osteoporosis classification from knee X-ray images. The proposed method aims to refine the high-dimensional deep representations produced by ViT backbones and generate discriminative features that can be effectively utilized by both classical machine learning algorithms and ordinal learning strategies. Within the ViT-FESEL framework, Branch-A is designed for binary diagnosis, Branch-B for ordinal modeling of osteoporosis severity levels, and Branch-C for an ensemble structure that fuses multi-scale representations extracted from different ViT backbones. Experimental evaluations are conducted using a five-fold cross-validation scheme with various backbone architectures and patch sizes. The results demonstrate that the FESEL module significantly enhances inter-class separability and that the Branch-C ensemble structure achieves the highest and most stable performance with an accuracy of 89.77%. Although ordinal learning results exhibit lower absolute accuracy, they successfully capture the ordered relationships among classes in a clinically meaningful manner. Overall, this study presents a hybrid, modular, and generalizable solution for osteoporosis detection from knee X-ray images and highlights the proposed ViT-FESEL approach as a strong candidate for clinical decision-support systems.

1. INTRODUCTION

Osteoporosis is a progressive skeletal disorder characterized by a decrease in bone mineral density and deterioration of bone microarchitecture, leading to a substantially increased risk of fractures, particularly in the elderly population. Early and accurate assessment of osteoporosis severity is of critical importance for fracture prevention and the determination of appropriate treatment strategies; however, conventional diagnostic techniques such as dual-energy X-ray absorptiometry (DXA) cannot always be effectively utilized due to limitations related to accessibility, cost, and routine clinical implementation [1, 2]. In this context, plain radiographic imaging (X-ray) is among the most widely available and low-cost modalities for musculoskeletal assessment, thereby motivating the development of automated computer-aided diagnosis systems. In recent years, Vision Transformer (ViT)-based deep learning approaches have demonstrated remarkable success in medical image analysis owing to their capability to model long-range dependencies and to produce high-level visual representations [3, 4]. Nevertheless, a large proportion of existing studies rely on end-to-end learning strategies, which suffer from notable limitations, such as high annotation costs, limited interpretability, and uncertainty about the contribution of learned representations [5, 6].

Wani and Arora [7] employed transfer-learning–based CNN architectures for osteoporosis detection in knee X-ray images in their study. The dataset was labeled into three classes—normal, osteopenia, and osteoporosis—based on T-scores from a Quantitative Ultrasound (QUS) system and consisted of 381 knee X-ray images. Pretrained CNN models, including AlexNet, VGG-16, VGG-19, and ResNet-18, were comparatively evaluated, and it was demonstrated that transfer learning significantly improves performance under limited data conditions. According to the experimental results, the highest classification performance was achieved with the AlexNet architecture, achieving 91.1% accuracy, while the other architectures reported accuracies of 84.2%–86.3%. The study highlighted that although the knee region is a relatively underexplored anatomical site for osteoporosis assessment, X-ray images can be effectively used as a low-cost, accessible screening tool. Moreover, the proposed three-class disease staging differentiates this work from many existing approaches that focus solely on binary classification.

In another study, Shen [8] used deep learning techniques to diagnose osteoporosis from knee X-ray images. The author employed a dataset of 1,573 images (780 normal and 793 labeled as early-stage osteoporosis) collected from Kaggle and Mendeley, and trained a VGG19 model to classify images into three categories: healthy, osteopenia, and osteoporosis, achieving an accuracy of 89%. Shen's work occupies an important place in the literature by demonstrating the potential of the knee region for osteoporosis assessment using CNN-based approaches.

Qureshi et al. [9] proposed a framework that integrates deep learning and transfer learning techniques for osteoporosis detection from knee X-ray images. The authors performed binary classification on a dataset comprising 372 images (186 normal and 186 osteoporosis) with the objective of distinguishing between normal and osteoporotic cases. Using a ResNet-50–based model, they achieved 90% accuracy, demonstrating high performance.

Naguib et al. [10] proposed a novel deep learning model based on a superfluity mechanism for diagnosing osteoporosis and osteopenia from knee X-ray images. The study targeted a three-class categorization consisting of normal, osteopenia, and osteoporosis, and utilized two separate datasets comprising 240 images in the first dataset and 371 images in the second dataset. After applying data augmentation techniques, the authors achieved accuracies of 83.74% and 74.51% on the first and second datasets, respectively. The work of Naguib et al. presents a CNN-based architecture that aims to prevent information loss by propagating features through two parallel branches.

Hong et al. [11] conducted a comprehensive study using deep learning methods to automatically detect osteoporosis and vertebral fractures (VF) from lateral spine X-ray images. The research was conducted on a large-scale dataset comprising 26,299 lateral spine radiographs from 9,276 patients. Using the EfficientNet-B4 architecture, separate binary classification models were developed for vertebral fracture detection and osteoporosis diagnosis. On the internal test set, an accuracy of 77% (0.85 AUROC) was achieved for osteoporosis detection, while vertebral fracture detection reached an accuracy of 91% (0.93 AUROC). The study by Hong et al. is particularly notable for its large data volume and focus on spinal imaging.

Zhang et al. [2] conducted a multicenter cohort study using deep convolutional neural networks (DCNNs) to screen for osteopenia and osteoporosis in lumbar spine X-ray images. The researchers worked with a large-scale dataset comprising 2,510 radiographs (anteroposterior and lateral views) collected from 1,255 patients, and categorized subjects into three classes—normal, osteopenia, and osteoporosis—according to World Health Organization criteria. The proposed model achieved AUCs of 0.810 for osteopenia detection and 0.767 for osteoporosis detection on external test datasets, while the overall classification accuracy during training was approximately 58–60% on the validation set.

Table 1. Literature review

	Study	Dataset	Result
[7]	Transfer-learning-based CNN comparison (AlexNet, VGG16, VGG19, ResNet18) for three-class knee X-ray osteoporosis classification	381 knee X-rays; 3 classes (normal, osteopenia, osteoporosis); labels based on QUS T-scores	Best: AlexNet 91.1% Acc; Others: 84.2%–86.3% Acc
[8]	VGG19-based deep learning model for three-class knee osteoporosis classification	1,573 images (780 normal, 793 early osteoporosis); Kaggle + Mendeley; output: healthy/osteopenia/osteoporosis	89% Acc
[9]	ResNet-50 transfer learning for binary knee osteoporosis detection	372 images (186 normal, 186 osteoporosis)	90% Acc
[10]	Superfluity-mechanism dual-branch CNN for three-class knee osteoporosis/osteopenia diagnosis	Dataset-1: 240 images; Dataset-2: 371 images	83.74% Acc (D1); 74.51% Acc (D2)
[11]	EfficientNet-B4 for automatic osteoporosis and vertebral fracture detection from lateral spine X-rays	26,299 spine X-rays from 9,276 patients	Osteoporosis: 77% Acc (0.85 AUROC); VF: 91% Acc (0.93 AUROC)
[12]	Systematic review of deep learning for knee osteoarthritis grading	Review of 74 studies	KL-grade 5-class accuracy: 70.4%–82.5%
[13]	EfficientNet-based binary cervical X-ray osteoporosis detection vs. surgeons	230 patients (200 train, 30 test)	DL: 80.0% Acc; Surgeons: 60.6% Acc
ViT-FESEL	ViT-FESEL (Branch-C, XGBoost)	1,573 knee X-rays; binary (normal vs. pathological)	89.77% Acc

Yeoh et al. [12] conducted a comprehensive systematic review evaluating the effectiveness of deep learning methods for diagnosing knee osteoarthritis. A total of 74 studies focusing on the segmentation and classification of knee joint images were analyzed. The review reported that KL-grade–based five-class staging studies using knee radiographs and deep neural networks—such as VGG, ResNet, and DenseNet variants—generally achieved accuracy rates ranging from 70.4% to 82.5%.

Tamai et al. [13] developed a deep learning–based algorithm for detecting osteopenia and osteoporosis from cervical radiography (neck X-ray) images. Using data from 230 patients (200 for training and 30 for independent testing), an EfficientNet-based model was trained in a binary classification setting defined by WHO criteria (normal vs. combined osteopenia/osteoporosis class), achieving an accuracy of 80.0%. This performance significantly exceeded the average accuracy of nine expert surgeons (60.6%) evaluated on the same dataset, highlighting the diagnostic value of deep learning approaches. The studies mentioned in the literature review are presented in Table 1.

1.1 Contribution

This study presents the following main novelties and contributions:

A novel framework (ViT-FESEL) that integrates ViT–based representations with feature extraction and selection is proposed. The approach refines high-dimensional deep representations obtained from ViT backbones through a systematic feature extraction and selection process, thereby improving classification performance and generalizability.

Osteoporosis assessment is addressed under a unified framework using binary, ordinal, and ensemble-based decision scenarios. Through the Branch-A, Branch-B, and Branch-C structures, the proposed method accommodates diverse clinical requirements—including basic diagnosis, ordinal modeling of disease severity, and multi-representation integration—within a single holistic architecture.

A CORAL-based ordinal learning strategy is employed alongside ViT representations to model osteoporosis severity levels. By explicitly considering the natural ordering of classes, the framework produces clinically more meaningful and consistent severity predictions.

An ensemble strategy that fuses multi-scale representations derived from different ViT backbones and patch sizes is developed. The proposed Branch-C structure combines complementary representation spaces, yielding more stable, higher, and more generalizable performance compared to individual models.

The impact of the FESEL module on inter-class separability and decision stability is systematically investigated through extensive experiments. Results across various backbones and learning scenarios quantitatively confirm the critical role of feature selection in ViT-based representations.

Owing to its modular design, the proposed framework provides an adaptable foundation that can be readily extended to other medical imaging problems. ViT-FESEL is designed not only for osteoporosis detection from knee X-ray images, but also as a scalable approach for other musculoskeletal disorders characterized by similar structural alterations.

2. MATERIAL AND METHOD

2.1 Dataset

In this study, the Multi-class Knee Osteoporosis X-ray Dataset [14], available as open source, was used to automatically classify knee joint osteoporosis. The dataset consists of 1,573 knee X-ray images from different individuals and is a multi-class dataset divided into three classes: clinically normal, osteopenia, and osteoporosis (Figure 1). In this study, within the scope of binary classification, the osteopenia and osteoporosis classes were combined into a single pathological class, resulting in a total of 793 images; the normal class consists of 780 images. The images do not contain any manual annotation or regional labeling; they reflect structural changes in bone mineral density and differences in trabecular patterns, incorporating the variation in resolution, contrast, and imaging conditions encountered in actual clinical practice. In this respect, the dataset provides a highly representative and realistic resource for modeling osteoporotic changes specific to the knee region. Within the scope of the study, the dataset was used in multiple ways to evaluate different clinical decision scenarios. In binary classification experiments, the osteopenia and osteoporosis classes were combined into a single pathological class, and classification was performed against the normal class. In the ordinal classification scenario, three classes were preserved to model the natural ordinal structure of disease severity. In ensemble-based experiments, the entire dataset was used to combine representations obtained from different ViT backbones. All experiments were conducted using a k-fold cross-validation scheme that preserved the class distribution; feature selection and model training were performed only on training subsets to prevent information leakage into the test data.

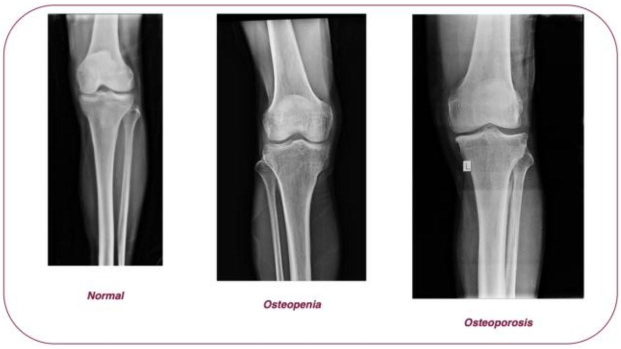

Figure 1. Examples of the dataset

The X-ray images in the dataset exhibit variability in resolution and contrast levels, reflecting the imaging conditions commonly encountered in clinical practice. The images do not contain any manual markings or regional annotations, and are organized solely by class labels. This characteristic enables the dataset to be directly applicable to both traditional machine learning and deep learning–based approaches. Furthermore, the multi-class nature of the dataset provides a suitable basis for studies aimed not only at identifying advanced-stage osteoporosis, but also at distinguishing early-stage bone loss conditions such as osteopenia.

2.2 Vision Transformer-Based Feature Extraction and Selection

In this study, a novel and comprehensive method, ViT-FESEL, is proposed for osteoporosis classification from knee X-ray images. The main motivation behind ViT-FESEL is that disease-specific findings in musculoskeletal radiographs are not limited to local tissue characteristics but are also strongly associated with the overall morphological structure of bone and global contextual relationships. In this regard, unlike convolutional neural network (CNN)–based architectures with limited receptive fields, ViT architectures leverage global self-attention, enabling more effective modeling of long-range dependencies across different image regions.

The proposed ViT-FESEL framework is a multi-branch architecture designed to address the problem from multiple clinical decision-making perspectives. This structure consists of three main components:

- Branch-A aims to perform binary classification by extracting ViT-based deep representations, applying feature selection, and employing classical machine learning classifiers.
- Branch-B focuses on ViT-based ordinal classification by explicitly accounting for the natural ordering of osteoporosis classes.
- Branch-C provides a multi-scale, more generalizable ensemble-based decision mechanism by fusing complementary representations from different ViT variants.

This multi-branch design reduces dependency on a single architecture while allowing each branch to model different levels of information. Overall, ViT-FESEL integrates transformer-based representation learning, dimensional refinement through feature selection, and task-specific classification strategies within a unified framework. By decoupling representation learning from decision-making processes and formulating the method to be adaptable to binary, ordinal, and ensemble-based scenarios, the proposed approach aims to deliver both high classification performance and clinically meaningful, stable outputs.

2.2.1 Vision Transformers–based architecture

In this study, the ViT architecture is employed during the deep representation learning stage. The primary reason for preferring ViT architectures is their ability to model images not only through local spatial windows, as in convolution-based networks, but also through global contextual relationships. In knee X-ray images, osteoporosis-related findings are not restricted to local bone tissue alterations; rather, they are closely associated with the overall bone structure and the relationships among different anatomical regions. Therefore, the self-attention mechanism in ViT offers a significant advantage for capturing such structural dependencies.

The ViT architecture divides the input image into fixed-size patches and treats these patches as a sequence [15-18]. Each patch is transformed into a fixed-dimensional embedding vector via a linear projection, which is then used as an input token to the transformer architecture [19]. To support the classification task, an additional learnable [CLS] token is appended to the sequence. This token aggregates information from all patches across the transformer layers and forms a global representation of the image [20, 21].

Since transformer architectures do not inherently encode spatial order, positional embeddings are used to encode patch locations within the image [22]. Positional embeddings are added to each patch token, enabling the model to learn the relative positions of anatomical structures in knee X-ray images [23]. Consequently, the ViT model generates representations that account for both content information and spatial arrangement. The fundamental building blocks of the ViT architecture are transformer encoder blocks, composed of Multi-Head Self-Attention and fully connected feed-forward networks [24]. The self-attention mechanism allows each patch to simultaneously attend to all other patches, enabling the model to capture correlations between distant anatomical regions and to more effectively learn global structural patterns associated with osteoporosis. Residual connections and layer normalization employed within the encoder blocks facilitate stable training of the deep architecture [25, 26].

In this study, ViT variants with different capacities and patch sizes are employed to enhance the model's multi-scale representation learning capability. Smaller patch sizes enable a more detailed capture of local tissue characteristics, whereas larger patch sizes emphasize the image's global structural organization. This design choice lays the foundation for constructing rich and complementary representation spaces that are subsequently exploited by the feature selection and ensemble strategies within the ViT-FESEL framework.

Finally, the [CLS] token representation from the transformer encoder layers is treated as a high-level summary feature vector of the image. This representation is utilized for different purposes across the branches of the ViT-FESEL method. In the binary and ensemble-based classification branches, this vector is forwarded as input to the subsequent feature selection and classical classification stages. In contrast, within the ordinal classification branch, it is directly connected to the ordinal learning head, enabling end-to-end learning.

2.2.2 ViT-based feature extraction and feature selection (FESEL)

ViT architectures can generate high-dimensional, rich representations from input images. However, not all components of these representations are equally discriminative for the classification task. In medical X-ray images, data-related noise, variations in imaging conditions, and subtle inter-class differences may lead to redundancy and low-discriminative dimensions in the learned embedding space. This issue can adversely affect the model's generalization performance, particularly in limited-data scenarios. For this reason, ViT-based feature extraction within the ViT-FESEL framework is supported by a systematic feature selection stage (FESEL).

In this study, the [CLS] token representation obtained from the ViT backbone for each knee X-ray image is regarded as a high-level summary feature vector of the image. This vector is extracted from the final encoder block of the ViT model and encodes the image's global structural information. Rather than being directly fed into the classifier, this deep representation vector is subjected to a feature selection process to improve classification performance and control model complexity.

The primary objective of the feature selection process is to identify, from the high-dimensional representation space learned by ViT, the subspace that best reflects inter-class separability and is most beneficial for generalization. Accordingly, the FESEL stage is designed with two main goals:

- preserving features with high discriminative power among classes, and
- eliminating features that contribute little to the classification process or contain noise.

This strategy reduces computational cost and mitigates the risk of overfitting.

Mathematically, the feature vector extracted from the ViT backbone can be denoted as $f \in RD$. In the FESEL stage, this vector is mapped to a lower-dimensional subspace and represented as $f* \in Rd$ with $d \ll D$. This transformation is not merely a dimensionality reduction operation, but rather a selection of features that contain discriminative information. The selected feature subset is determined solely from the training data in each cross-validation fold, thereby preventing data leakage into the test set. In the FESEL stage, a two-step dimensionality reduction is applied: in the first step, MI-based selection retains the top 512 most informative features from the high-dimensional ViT representation vector; in the second step, NCA reduces this to a final feature dimensionality of $d = 128$. These settings were kept constant across all backbone architectures and cross-validation folds.

The integration of the FESEL mechanism with ViT-based representations establishes a hybrid architecture that effectively combines the strengths of deep learning and classical machine learning approaches. While the ViT model learns complex, high-level visual patterns, the FESEL stage selects the most informative components of these patterns for classification, thereby enabling classical classifiers to learn more effectively. This decoupling enhances the modularity of the ViT-FESEL framework and allows different classification strategies to be evaluated on the same representation space.

The refined representation vectors obtained after feature selection are provided as inputs to classical machine learning–based classifiers within the Branch-A and Branch-C components of the ViT-FESEL method. By combining the strong representation-learning capabilities of deep models with the flexibility of classical algorithms for decision boundary formation, this structure aims to deliver more stable, generalizable classification performance for osteoporosis detection in knee X-ray images.

2.2.3 Branch-A: ViT-FESEL–based binary classification approach

Branch-A forms the core and reference pipeline of the proposed ViT-FESEL method and is designed for scenarios in which osteoporosis status is treated as a binary classification problem (presence/absence) based on knee X-ray images. The main objective of this branch is to demonstrate the extent to which ViT-based deep representations, systematically refined through feature selection and used in conjunction with classical machine learning classifiers, can produce effective, stable, and generalizable results. In this context, Branch-A presents both the simplest implementation of the proposed framework and provides a strong basis for comparison for more complex ordinal and ensemble-based scenarios. As the first step in this pipeline, each input X-ray image sequence is processed by the selected ViT backbone, yielding a high-dimensional representation vector that captures the image's global structural features and potentially related contextual patterns. However, such ViT-based representations, due to their high dimensionality and limited discriminative components, can lead to performance fluctuations and overfitting risks when used directly in the classification stage. Therefore, in Branch-A, the ViT-based representation learning process is supported by a feature selection (FESEL) stage.

In this pipeline, as a first step, each input X-ray image is processed by the selected ViT backbone, yielding a high-dimensional representation vector that captures the image's global structural features and potentially associated contextual patterns with osteoporosis. However, such ViT-based representations, due to their high dimensionality and limited discriminative components, can lead to performance fluctuations and overfitting risks when used directly in the classification stage. Therefore, in Branch-A, the ViT-based representation learning process is supported by a feature selection (FESEL) stage.

In the FESEL phase, deep representation vectors generated by ViT are analyzed using only training data, and the feature subset that best distinguishes between classes is determined. In this study, Mutual Information (MI) and Neighborhood Components Analysis (NCA) methods were used within the FESEL mechanism. MI-based feature selection aims to identify the most informative components for classification by measuring the statistical dependence of each feature on class labels, while the NCA method offers a feature weighting and selection strategy that maximizes the distinction between classes by considering the neighborhood relationships between examples. Thanks to these two approaches, features that are strongly related to class information and that define decision boundaries more clearly are retained, while redundant or noisy dimensions are eliminated. The feature selection process was performed independently for each cross-validation fold, thus preventing information leakage to the test data.

The more compact, discriminative feature space obtained through feature selection enables classical machine learning classifiers to learn decision boundaries more stably. This study focuses on methods that can effectively model nonlinear decision boundaries and perform well in high-dimensional feature spaces; in this context, XGBoost and RBF-kernel Support Vector Machines (SVM-RBF) were used. These classifiers can effectively distinguish between the presence and absence of osteoporosis using refined deep representations generated within the ViT-FESEL framework.

Within the scope of Branch-A, model performance was evaluated using multiple metrics that account for class imbalance. These metrics include Accuracy, macro-averaged precision (macro-Precision), macro-Recall, macro-F1 score, Balanced Accuracy, and ROC-AUC, as appropriate. This multifaceted evaluation strategy ensures a comprehensive examination of not only overall accuracy but also the model's discriminative power for each class. In conclusion, Branch-A demonstrates high discriminative power and stable performance in the binary osteoporosis classification problem when ViT-based representation learning is combined with MI and NCA-based feature selection; it also provides a robust and meaningful reference for the Branch-B and Branch-C architectures.

2.2.4 Branch-B: ViT-based ordinal classification approach

Osteoporosis assessment is not treated as a binary decision problem (“present” or “absent”) in many clinical scenarios. Instead, it is evaluated using classes that represent the disease's severity levels and have a natural ordering (e.g., normal, mild, moderate, severe). Such problems are defined as situations where there is an ordinal (sequential) relationship between classes. However, softmax-based models used in classical multi-class classification approaches ignore this sequential relationship among classes and treat all classes as equally distant from one another. This situation can lead to clinically meaningful errors not being sufficiently distinguished. To overcome this limitation, a separate pipeline, Branch-B, based on an ordinal learning approach, was designed within the ViT-FESEL framework.

The primary objective of Branch-B is to produce clinically more meaningful and consistent predictions by directly modeling the natural hierarchical structure of osteoporosis classes, enabling the learning of relative relationships between classes. In this context, an ordinal classification head based on CORAL (Consistent Rank Logits) has been integrated onto the ViT backbone [27, 28]. The CORAL approach addresses a problem with K ordered classes by converting it into $K - 1$ binary decision problems; each decision aims to predict whether the relevant example is above a certain class threshold. Thanks to this structure, the model can consistently learn not only class labels but also the ordinal relationships between classes. Thus, an error between “mild” and “moderate” osteoporosis, for example, can be less penalized than an error between “normal” and “severe” osteoporosis; this provides a more clinically meaningful assessment.

In Branch-B, as in the previous branch, a high-level deep representation is obtained from the input sequence of X-ray images via the ViT backbone. The [CLS] token representation from the final encoder block of ViT serves as a summary of the image's global structural features and contextual information. However, unlike Branch-A and Branch-C, no MI or NCA-based feature selection is applied to these representations in Branch-B. This choice is deliberate and based on methodological reasoning. Since the fundamental goal in ordinal learning is to preserve the continuous and relative-intensity relationship between classes, aggressive feature selection steps aimed at increasing discriminability can break this ordered structure and lead to a loss of intensity continuity. Therefore, in Branch-B, the goal is to directly integrate ordinal information into the deep representation learning process, and ViT-based representations are connected to the end-to-end CORAL header.

When evaluating model performance under Branch-B, evaluation metrics specific to ordinal problems were used in addition to the classic accuracy metric. In particular, the Quadratic Weighted Kappa (QWK) metric was chosen to assess the extent to which model predictions align with the actual class rankings. The QWK metric weights errors by considering the ordinal distance between classes, thereby providing a clinically more meaningful performance assessment. In addition, macro-averaged precision, recall, and F1 scores were also reported to account for the effects of class imbalance. This multifaceted evaluation approach enables a comprehensive analysis of ordinal classification performance.

As a result, Branch-B, as the component of the ViT-FESEL method that addresses the scenario closest to clinical reality, provides an assessment that considers not only the presence of osteoporosis but also its severity. While Branch-A aims for high discriminative power through MI and NCA-based feature selection, Branch-B deliberately omits this step to preserve ordinal continuity. In this respect, Branch-B complements the binary classification approach and demonstrates the flexibility and comprehensive adaptability of the proposed framework to different diagnostic requirements.

2.2.5 Branch-C: Multi–vision transformer–based ensemble approach

Although ViT architectures have strong representation learning capabilities, when trained with a specific patch size or model capacity, they can become more sensitive to certain patterns in the image compared to others. Particularly in knee X-ray images, osteoporosis findings become more pronounced in some cases with local bone tissue and trabecular structure changes, and in some cases with the overall morphological structure and global arrangement of the bone. This situation can lead to models based on a single ViT variant exhibiting unstable or variable performance in certain cases. In this study, to overcome this limitation, a multi-ViT-based ensemble approach called Branch-C is proposed within the ViT-FESEL framework.

The fundamental idea behind Branch-C is based on the assumption that ViT variants with different patch sizes and model capacities can learn complementary and multi-scale representations from the image. Smaller patch sizes focus on local bone tissue and fine structural details, while larger patch sizes more effectively represent the overall morphological arrangement and global structural relationships of the image. Similarly, ViT models with different depths and parameter counts can capture osteoporosis-related patterns at different levels of abstraction. Therefore, in Branch-C, multiple ViT variants are used in parallel to create a rich, diversified, and complementary representation space.

In this context, high-dimensional deep representation vectors are extracted independently for each input sequence X-ray image from different ViT backbones. These representations are then structured to preserve the information provided by each ViT model and subsequently combined into a single unified representation vector through feature-level concatenation. This process enables the model to learn on a multi-scale and multi-perspective feature space. However, directly combining representations obtained from different ViT variants significantly increases the dimension of the feature space, which can lead to higher computational costs and overfitting risk.

Therefore, in Branch-C, a two-stage feature selection (FESEL) mechanism was applied on the combined ensemble representation vectors. In the first stage of the FESEL process, the statistical dependency of each feature with the class labels was measured using the MI method, and features that carried weak information for classification, were noisy, or redundant were eliminated. This MI-based pre-selection stage ensures the creation of a more compact and informative subspace in the high-dimensional ensemble representation space. Subsequently, NCA was applied to this feature subset selected by MI. NCA selects features that maximize class separation by considering inter-example neighborhood relationships and contribute to learning clearer decision boundaries. Thanks to this two-stage FESEL approach, complementary and discriminative features from different ViT variants are preserved, while components that make only a limited contribution to classification are systematically eliminated. For MI-based feature selection, the `mutual_info_classif` function from the scikit-learn library was used, and the continuous prediction method was preferred (discrete_features=False). This choice stems from the fact that ViT-based representation vectors contain continuous-valued features. During the MI phase, the top 512 features with the highest statistical dependence were selected. NCA, on the other hand, was implemented using scikit-learn's Neighborhood Components Analysis. NCA is a direct metric learning approach that does not involve a traditional k-neighborhood parameter; therefore, the k value is not relevant in this method. The number of NCA components was set to n_components=128, and the maximum number of iterations to max_iter=200; these values were selected through preliminary experiments.

The refined ensemble representation vectors obtained after feature selection are used in the final classification stage within the Branch-C framework. In this study, XGBoost and RBF-kernel Support Vector Machines (SVM-RBF), which can effectively model non-linear decision boundaries, were preferred. These classifiers can effectively learn complex, high-dimensional decision spaces by combining information from multiple ViT models. The primary goal of Branch-C is to provide a more stable, balanced, and generalizable performance across different data samples and classes rather than achieving the highest score with a single model.

As a result, Branch-C, as the top-level integration component of the ViT-FESEL method, provides a holistic decision mechanism that refines complementary representations from multiple ViT architectures via MI- and NCA-based feature selection, and combines them with powerful classical classifiers. This structure aims to achieve more robust, consistent, and clinically reliable osteoporosis classification results compared to approaches based on a single ViT backbone, strongly supporting the practical clinical application potential of the proposed framework.

2.2.6 Training procedure, implementation details, and overall workflow

Within the proposed ViT-FESEL framework, all experiments are structured according to principles of fair comparison and generalizability. A unified training and evaluation protocol is adopted for all branches (Branch-A, Branch-B, and Branch-C), enabling consistent comparison of different architectural configurations and classification strategies.

ViT backbones are initialized using weights pre-trained on large-scale natural image datasets. This choice facilitates faster convergence and more stable representation learning, which is particularly important given the limited size of medical imaging datasets. During training, the final layers of ViT models are updated task-specifically, while limited or progressive fine-tuning is applied to earlier layers to mitigate overfitting. All experiments are conducted using a k-fold cross-validation scheme. For each fold, training, validation, and test splits are strictly maintained, and both feature selection (FESEL) and model learning processes are performed exclusively on the corresponding training subset. This design prevents information leakage from the test data and increases the reliability of the reported performance metrics.

At the feature selection stage, high-dimensional ViT representation vectors are refined to retain only the most discriminative components for classification. The representations obtained after feature selection are used as inputs to classical machine learning classifiers in Branch A and Branch C. Hyperparameters of the classifiers employed in the classification stage are optimized using the training data and are applied consistently across all folds.

Model performance is evaluated using multiple metrics that account for class imbalance and clinical decision-making requirements. For the binary and ensemble-based branches, metrics such as accuracy, macro-averaged precision, recall, F1-score, balanced accuracy, and ROC-AUC are reported. For the ordinal classification branch, the QWK metric, which measures the degree of agreement between predicted and true ordinal labels, is additionally employed. This multi-dimensional evaluation strategy enables analysis not only of overall performance, but also of inter-class separability and clinically meaningful error patterns [29].

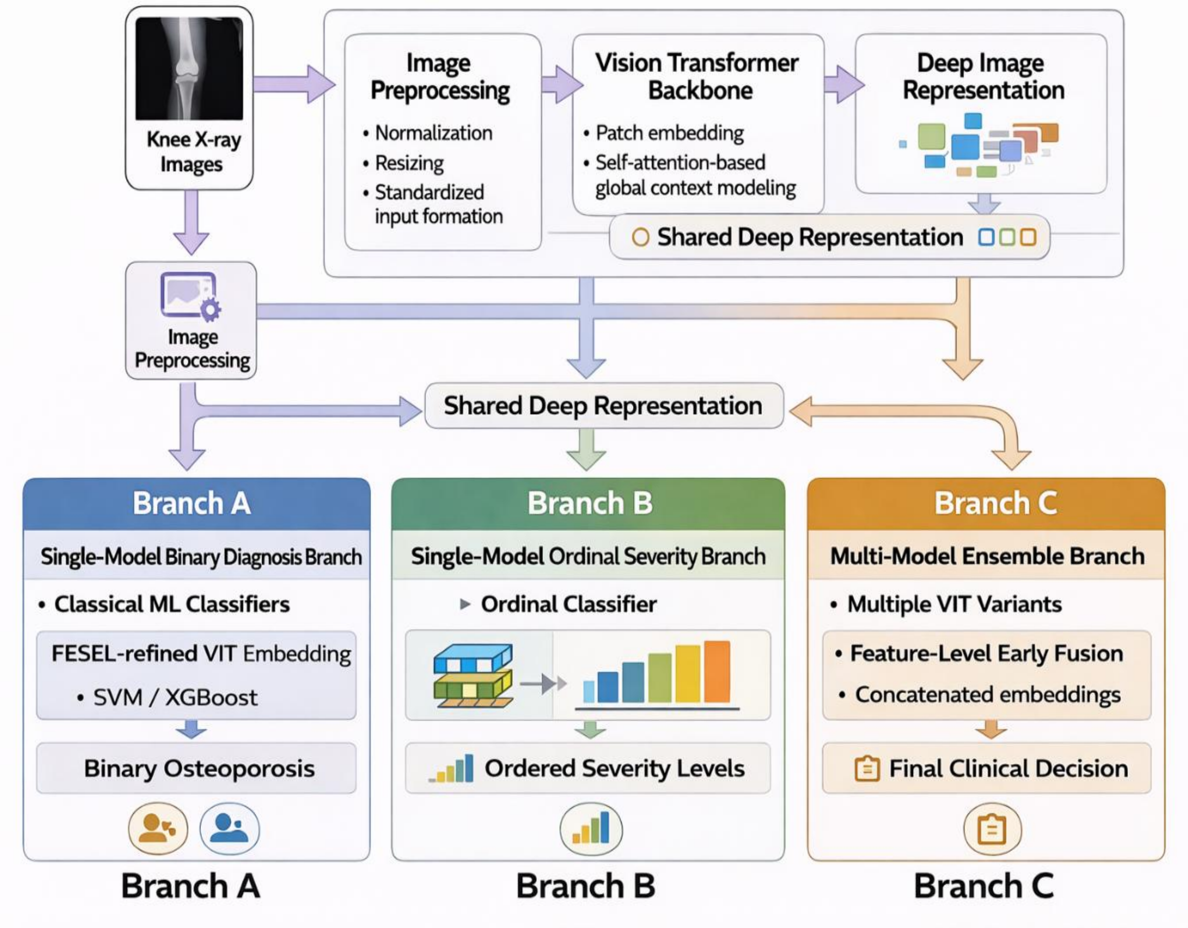
Knee X-ray Images
Image Preprocessing
• Normalization
• Resizing
• Standardized input formation
Vision Transformer Backbone
• Patch embedding
• Self-attention-based global context modeling
Deep Image Representation
Shared Deep Representation
Image Preprocessing
Shared Deep Representation
Branch A
Single-Model Binary Diagnosis Branch
• Classical ML Classifiers
FESEL-refined VIT Embedding
• SVM / XGBoost
Binary Osteoporosis
Branch A
Branch B
Single-Model Ordinal Severity Branch
▸ Ordinal Classifier
Ordered Severity Levels
Branch B
Branch C
Multi-Model Ensemble Branch
• Multiple VIT Variants
• Feature-Level Early Fusion
• Concatenated embeddings
Final Clinical Decision
Branch C

Figure 2. Graphical abstract of Vision Transformer–Based Feature Extraction and Selection (ViT-FESEL)

The overall workflow of the ViT-FESEL method begins with the preprocessing of input knee X-ray images (Figure 2). The preprocessed images are then transformed into high-level deep representations through ViT backbones. These representations are subsequently processed through different branches depending on the problem formulation. In Branch-A, ViT-based representations are passed through a feature selection stage and then classified using classical machine learning classifiers for binary classification. In Branch-B, the same representations are directly connected to an ordinal classification head that explicitly models the ordered structure of the classes. In Branch-C, representations extracted from different ViT variants are fused and refined via feature selection to form an ensemble-based decision mechanism.

Through this multi-branch design, ViT-FESEL addresses osteoporosis assessment from multiple perspectives, including binary diagnosis, severity level estimation, and multi-model integration, rather than relying on a single viewpoint. Consequently, the proposed method offers a flexible, generalizable framework that provides a holistic, reliable solution for osteoporosis classification based on knee X-ray images.

3. EXPERIMENTAL RESULTS

In this section, the experimental performance of the proposed ViT-FESEL framework for osteoporosis classification from knee X-ray images is comprehensively evaluated. The conducted experiments aim to demonstrate the effectiveness of combining ViT–based representation learning with feature extraction and selection under different decision scenarios. Accordingly, the proposed method is analyzed across three complementary configurations: binary diagnosis, ordinal severity level estimation, and ensemble-based decision-making. Model performance is reported using multiple evaluation metrics that reflect generalization capability, decision stability, and clinical relevance. In addition, the contribution of the FESEL module and the advantages provided by the multi-branch design are systematically investigated through comparative experiments.

The ViT_Base/16–Branch-A results presented in Table 2 demonstrate that ViT representations refined through the FESEL module provide high and stable performance for binary osteoporosis classification. Across five-fold cross-validation, XGBoost consistently outperforms SVM-RBF in all evaluation metrics, yielding more robust and reliable results. The accuracy values obtained with XGBoost range between 83.5% and 87.5%, with the highest performance of 87.50% achieved in the third fold. The close agreement among macro-precision, macro-recall, and balanced accuracy indicates that the model discriminates between the two classes with comparable effectiveness and is only marginally affected by class imbalance. Moreover, ROC-AUC values of 0.946–0.967 indicate strong discriminative capability when ViT-based representations are combined with FESEL to distinguish between osteoporosis-present and osteoporosis-absent cases. Overall, these findings confirm that even the final-layer representations extracted from a single ViT_Base/16 backbone can yield reliable binary diagnostic performance when integrated with appropriate feature selection and classical machine learning classifiers.

Table 2. Binary osteoporosis classification results of Branch-A based on ViT_Base/16

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	XGBoost	0.8352	0.8230	0.8294	0.8256	0.8294	0.9461
0	SVM-RBF	0.7955	0.7765	0.7964	0.7830	0.7964	0.9378
1	XGBoost	0.8494	0.8518	0.8360	0.8406	0.8360	0.9544
1	SVM-RBF	0.8693	0.8564	0.8521	0.8528	0.8521	0.9610
2	XGBoost	0.8722	0.8490	0.8564	0.8515	0.8564	0.9670
2	SVM-RBF	0.8722	0.8470	0.8590	0.8503	0.8590	0.9620
3	XGBoost	0.8750	0.8605	0.8784	0.8677	0.8784	0.9621
3	SVM-RBF	0.8665	0.8457	0.8711	0.8539	0.8711	0.9581
4	XGBoost	0.8636	0.8482	0.8387	0.8430	0.8387	0.9566
4	SVM-RBF	0.8466	0.8235	0.8356	0.8286	0.8356	0.9516

Table 3. Binary osteoporosis classification results of Branch-A based on ViT_Base/32

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	XGBoost	0.8125	0.7945	0.8078	0.7994	0.8078	0.9482
0	SVM-RBF	0.8182	0.7950	0.8174	0.8017	0.8174	0.9510
1	XGBoost	0.8580	0.8518	0.8375	0.8418	0.8375	0.9526
1	SVM-RBF	0.8636	0.8462	0.8445	0.8449	0.8445	0.9597
2	XGBoost	0.8693	0.8538	0.8515	0.8521	0.8515	0.9629
2	SVM-RBF	0.8580	0.8296	0.8391	0.8327	0.8391	0.9546
3	XGBoost	0.9034	0.8947	0.9021	0.8972	0.9021	0.9638
3	SVM-RBF	0.8778	0.8591	0.8807	0.8668	0.8807	0.9629
4	XGBoost	0.8750	0.8686	0.8591	0.8634	0.8591	0.9545
4	SVM-RBF	0.8636	0.8419	0.8524	0.8464	0.8524	0.9471

Table 3 presents the Branch-A binary classification results obtained using features extracted from the ViT_Base/32 backbone and refined through the FESEL module. Across five-fold cross-validation, both the XGBoost and SVM-RBF models exhibit high and consistent performance, with XGBoost achieving the highest accuracy of 90.34% in the third fold. Overall, XGBoost demonstrates superior performance, particularly in terms of accuracy, macro-F1 score, and balanced accuracy, while SVM-RBF yields competitive results in some folds. The close alignment between macro-averaged precision and recall indicates that both classes are learned in a balanced manner and that the impact of class imbalance remains limited. Furthermore, ROC-AUC values exceeding 0.94 across all folds confirm that ViT_Base/32-based representations, when combined with FESEL, provide strong discriminative capability for distinguishing between osteoporosis-present and osteoporosis-absent cases. These findings verify that the ViT_Base/32 backbone, despite using a larger patch size, can deliver reliable, generalizable binary diagnostic performance when integrated with appropriate feature selection and classical machine learning classifiers.

The ViT_Large/16–based Branch-A results presented in Table 4 indicate that a higher-capacity ViT backbone, when combined with the FESEL module, delivers strong and stable performance for binary osteoporosis classification. Across five-fold cross-validation, XGBoost and SVM-RBF yield similarly competitive results, with XGBoost achieving the highest accuracy of 88.92% in the third fold. The close correspondence among macro-averaged precision, recall, and F1-score suggests that both classes are learned in a balanced manner and that the effect of class imbalance on performance remains limited. In particular, ROC-AUC values ranging from 0.94 to 0.96 demonstrate that ViT_Large/16-based representations provide high discriminative power for distinguishing between osteoporosis-present and osteoporosis-absent cases. Overall, these findings confirm that the deeper, higher-parameter ViT_Large/16 backbone can achieve reliable, generalizable binary diagnostic performance when integrated with appropriate feature selection and classical machine learning classifiers.

The ViT_Large/32–based Branch-A results presented in Table 5 demonstrate that a ViT backbone with a larger patch size and high model capacity, when combined with FESEL, achieves high and stable performance for binary osteoporosis classification. Across five-fold cross-validation, both XGBoost and SVM-RBF achieve competitive results, with XGBoost achieving the highest accuracy of 89.77% in the third fold. The close alignment of macro-averaged precision, recall, and F1-score across all folds indicates balanced learning of both classes and a limited influence of class imbalance. In particular, ROC-AUC values ranging between 0.94 and 0.96 confirm the strong discriminative capability of ViT_Large/32-based representations for distinguishing between osteoporosis-present and osteoporosis-absent cases. Overall, these findings verify that the ViT_Large/32 backbone, despite its larger patch size, can achieve reliable, generalizable binary diagnostic performance when integrated with appropriate feature selection and classical machine learning classifiers.

Table 4. Binary osteoporosis classification results of Branch-A based on ViT-Large/16

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	XGBoost	0.8182	0.7965	0.8121	0.8026	0.8121	0.9447
0	SVM-RBF	0.8295	0.8055	0.8240	0.8121	0.8240	0.9416
1	XGBoost	0.8693	0.8609	0.8548	0.8570	0.8548	0.9556
1	SVM-RBF	0.8636	0.8507	0.8473	0.8480	0.8473	0.9574
2	XGBoost	0.8551	0.8427	0.8427	0.8425	0.8427	0.9633
2	SVM-RBF	0.8750	0.8508	0.8640	0.8543	0.8640	0.9635
3	XGBoost	0.8892	0.8777	0.8847	0.8805	0.8847	0.9608
3	SVM-RBF	0.8750	0.8538	0.8808	0.8617	0.8808	0.9575
4	XGBoost	0.8494	0.8380	0.8354	0.8366	0.8354	0.9548
4	SVM-RBF	0.8438	0.8202	0.8252	0.8224	0.8252	0.9411

Table 5. Binary osteoporosis classification results of Branch-A based on ViT-Large/32

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	XGBoost	0.8068	0.7849	0.7948	0.7888	0.7948	0.9499
0	SVM-RBF	0.8239	0.7983	0.8166	0.8045	0.8166	0.9445
1	XGBoost	0.8693	0.8625	0.8493	0.8546	0.8493	0.9495
1	SVM-RBF	0.8750	0.8637	0.8486	0.8543	0.8486	0.9611
2	XGBoost	0.8636	0.8451	0.8469	0.8457	0.8469	0.9643
2	SVM-RBF	0.8665	0.8418	0.8571	0.8468	0.8571	0.9608
3	XGBoost	0.8977	0.8834	0.8945	0.8880	0.8945	0.9616
3	SVM-RBF	0.8778	0.8562	0.8725	0.8624	0.8725	0.9645
4	XGBoost	0.8608	0.8525	0.8283	0.8380	0.8283	0.9536
4	SVM-RBF	0.8750	0.8579	0.8565	0.8569	0.8565	0.9522

Table 6 reports the results of the CORAL-based ordinal classification approach integrated on top of the ViT_Base/16 backbone within Branch-B. Across five-fold cross-validation, accuracy values range from 62.22% to 68.75%, exhibiting relatively consistent performance among folds. The close correspondence among macro-averaged precision, recall, and F1-score indicates that the model achieves balanced learning across ordinal classes. Notably, higher accuracy and macro-F1 values observed in the second and fourth folds suggest that the CORAL approach is able to capture, to a certain extent, the ordered relationships among classes. Compared to binary classification, the lower absolute performance values reflect the inherently more challenging nature of ordinal classification, where decision boundaries are constrained by ordered relationships and error costs are distributed by class proximity. Overall, these findings confirm that the ViT_Base/16-based CORAL-ordinal structure represents a clinically more meaningful yet more demanding scenario for modeling osteoporosis severity levels, and that Branch-B serves as a complementary component to Branch-A.

Table 7 presents the Branch-B results obtained using the ViT_Base/32 backbone with the CORAL-based ordinal learning approach. Across five-fold cross-validation, accuracy values range between 65.34% and 70.17%, indicating generally consistent performance among folds. The close agreement among macro-averaged precision, recall, and F1-score suggests that the model achieves balanced learning across different osteoporosis severity levels. Notably, QWK values in the range of 0.59–0.69 demonstrate that the model not only predicts the correct class labels but also shows substantial agreement with the ordinal relationships among classes. The highest QWK value of 0.6903 is obtained in the second fold, further confirming that the CORAL approach is capable of modeling severity levels in a manner consistent with their ordered structure. Overall, these findings verify that the ViT_Base/32-based Branch-B configuration can generate clinically more meaningful and consistent predictions for ordinal osteoporosis assessment, which represents a more challenging problem compared to binary classification.

Table 8 presents the Branch-B results obtained using the ViT_Large/16 backbone with the CORAL-based ordinal learning approach. Across five-fold cross-validation, accuracy ranges from 61.93% to 68.75%, with generally consistent performance across folds. The close correspondence among macro-averaged precision, recall, and F1-score indicates that the model achieves balanced learning across different osteoporosis severity levels. QWK values range from 0.539 to 0.641, indicating that the model captures the ordinal relationships among classes to a meaningful degree. The highest QWK value of 0.6414 is obtained in the fourth fold, suggesting that the CORAL approach can partially model the ordered structure of osteoporosis severity using ViT_Large/16 representations. Overall, these findings confirm that the ViT_Large/16-based Branch-B configuration produces clinically meaningful ordinal predictions, albeit within a more challenging decision space compared to binary classification.

Table 6. CORAL-based Branch-B ordinal osteoporosis classification results using ViT_Base/16

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	CORAL-Ordinal	0.6222	0.4767	0.5164	0.4821	0.5164	0.5219
1	CORAL-Ordinal	0.6761	0.5955	0.5857	0.5726	0.5857	0.6203
2	CORAL-Ordinal	0.6875	0.6103	0.6106	0.6082	0.6106	0.6840
3	CORAL-Ordinal	0.6676	0.5611	0.5699	0.5539	0.5699	0.6298
4	CORAL-Ordinal	0.6875	0.6213	0.5860	0.5682	0.5860	0.6199

Table 7. CORAL-based Branch-B ordinal osteoporosis classification results using ViT_Base/32

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	CORAL-Ordinal	0.6534	0.5958	0.5891	0.5864	0.5891	0.6183
1	CORAL-Ordinal	0.7017	0.6721	0.6449	0.6454	0.6449	0.6371
2	CORAL-Ordinal	0.6903	0.6292	0.6268	0.6259	0.6268	0.6903
3	CORAL-Ordinal	0.6619	0.6131	0.6089	0.6082	0.6089	0.6506
4	CORAL-Ordinal	0.6534	0.5631	0.5633	0.5543	0.5633	0.5985

Table 8. CORAL-based Branch-B ordinal osteoporosis classification results using ViT_Large/16

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	CORAL-Ordinal	0.6193	0.4900	0.5178	0.4888	0.5178	0.5394
1	CORAL-Ordinal	0.6875	0.6347	0.6167	0.6149	0.6167	0.6282
2	CORAL-Ordinal	0.6761	0.6028	0.6013	0.5988	0.6013	0.6340
3	CORAL-Ordinal	0.6705	0.5707	0.5697	0.5492	0.5697	0.6120
4	CORAL-Ordinal	0.6761	0.5822	0.5848	0.5749	0.5848	0.6414

Table 9 presents the Branch-B results obtained using the ViT_Large/32 backbone with the CORAL-based ordinal learning approach. Across five-fold cross-validation, accuracy values range between 65.62% and 72.16%, with a noticeable performance improvement particularly in the second and third folds. The close correspondence among macro-averaged precision, recall, and F1-score indicates that the model achieves balanced learning across different osteoporosis severity levels. QWK values in the range of 0.59–0.71 demonstrate substantial agreement between model predictions and the ordinal relationships among classes, with the highest QWK value of 0.7089 obtained in the second fold. These results confirm that the ViT_Large/32 backbone, characterized by a larger patch size and higher model capacity, is able to capture the ordered structure of osteoporosis severity more effectively and to provide clinically more consistent severity predictions within the Branch-B configuration.

Overall, the results obtained within Branch-B clearly demonstrate that the ordinal nature of osteoporosis classification creates a more challenging decision space than the binary diagnosis scenario. Although the CORAL-based ordinal learning approach yields lower absolute performance metrics than the high accuracy and ROC-AUC values observed in Branch-A, the consistently moderate-to-high QWK scores indicate that the model captures the ordered relationships among classes to a meaningful extent. This finding emphasizes that the primary objective of Branch-B is not merely correct class assignment, but rather the clinically consistent modeling of relative disease severity.

While Branch-A provides higher discriminative power and clearer decision boundaries for binary diagnosis, Branch-B offers a more cautious yet clinically more realistic assessment, highlighting the complementary nature of these two approaches. These observations suggest that relying on a single decision strategy may be limiting and that integrating diverse backbones and learning paradigms can yield more stable and generalizable outcomes. Motivated by this insight, the next section presents the experimental results of the Branch-C ensemble approach, which combines multiple models and decision strategies within a unified framework.

Table 10 presents the performance results of the Branch-C ensemble structure, constructed by fusing representations extracted from four different ViT backbones, using XGBoost and SVM-RBF classifiers. Across five-fold cross-validation, both classifiers exhibit high and stable performance, with XGBoost achieving higher accuracy, macro-F1 score, and balanced accuracy. The highest accuracy of 89.77% is achieved with XGBoost in the third fold, where the macro-F1 score also reaches 0.8875. ROC-AUC values exceeding 0.94 in all folds, and approaching 0.96 in some folds, indicate that ensemble representations substantially enhance inter-class discriminability. The SVM-RBF model yields results close to XGBoost in several folds, further demonstrating the generalizability of the ensemble representations across different classifiers.

The Branch-C results show that the limitations observed in Branch-A and Branch-B are largely mitigated. Compared to the single-backbone Branch-A configuration, the complementary and multi-scale representations derived from multiple ViT backbones significantly improve decision stability and generalization capability. Likewise, while Branch-B, which models the ordinal structure, provides a more cautious but relatively lower absolute performance, Branch-C achieves both high accuracy and balanced class performance, representing the strongest and most stable decision scenario.

Overall, the experimental findings clearly indicate that the proposed ViT-FESEL framework offers complementary advantages under different decision scenarios. The results obtained in Branch-A demonstrate that representations extracted from a single ViT backbone and refined through

FESEL provide high discriminative power for binary osteoporosis diagnosis. The Branch-B results reveal that the ordinal learning strategy yields a more conservative yet clinically more realistic evaluation. In contrast, the Branch-C ensemble structure integrates the high discriminability of Branch-A with the representational richness of Branch-B, achieving the most stable and highest performance across all metrics. The fusion of multi-scale and complementary representations obtained from different ViT backbones and patch sizes alleviates the limitations of individual models, while the FESEL module enables more consistent learning of decision boundaries. These findings confirm that multi-representation and ensemble-based decision strategies produce more reliable and generalizable outcomes for medical imaging problems involving both local and global structural changes, such as osteoporosis, and strongly support the practical clinical applicability of the proposed ViT-FESEL approach.

Table 9. CORAL-based Branch-B ordinal osteoporosis classification results using ViT_Large/32

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	CORAL-Ordinal	0.6562	0.5917	0.5831	0.5801	0.5831	0.6291
1	CORAL-Ordinal	0.6676	0.6226	0.5925	0.5882	0.5925	0.5944
2	CORAL-Ordinal	0.7159	0.6464	0.6399	0.6368	0.6399	0.7089
3	CORAL-Ordinal	0.7216	0.6738	0.6388	0.6384	0.6388	0.6840
4	CORAL-Ordinal	0.6818	0.6228	0.6086	0.6089	0.6086	0.6290

Table 10. Binary classification results of the Branch-C Multi–ViT-based ensemble approach (XGBoost and SVM)

Fold	Model	acc	macroPre	macroRec	macroF1	balAcc	roc_aoc_ovr
0	ENSMB-XGB	0.8324	0.8114	0.8265	0.8175	0.8265	0.9490
1	ENSMB-XGB	0.8665	0.8597	0.8497	0.8538	0.8497	0.9481
2	ENSMB-XGB	0.8636	0.8439	0.8467	0.8443	0.8467	0.9658
3	ENSMB-XGB	0.8977	0.8806	0.8969	0.8875	0.8969	0.9661
4	ENSMB-XGB	0.8636	0.8546	0.8442	0.8489	0.8442	0.9512
0	ENSMB-SVM	0.8295	0.8058	0.8241	0.8123	0.8241	0.9447
1	ENSMB-SVM	0.8636	0.8468	0.8447	0.8444	0.8447	0.9575
2	ENSMB-SVM	0.8665	0.8418	0.8516	0.8448	0.8516	0.9626
3	ENSMB-SVM	0.8722	0.8495	0.8757	0.8567	0.8757	0.9613
4	ENSMB-SVM	0.8551	0.8311	0.8427	0.8356	0.8427	0.9447

4. DISCUSSION AND CONCLUSION

This study proposes a multi-stage, multi-branch framework, termed ViT-FESEL, that integrates ViT–based representation learning with a systematic feature extraction and selection strategy for osteoporosis detection in knee X-ray images. The dataset's class labels are based on publicly available sources and have not been validated by independent clinicians or T-score-based labeling; this constitutes a significant limitation of the study. The experimental results clearly demonstrate that deep learning–based representations can serve not only for end-to-end classification but also as powerful, discriminative features that can be effectively exploited by classical machine learning algorithms. In particular, incorporating the FESEL module reduces redundancy in high-dimensional ViT representations and enhances inter-class separability, thereby improving stability and generalizability across binary, ordinal, and ensemble-based decision scenarios. From this perspective, the study presents a hybrid and modular solution that unifies deep representation learning with traditional learning paradigms in the medical imaging domain.

A joint analysis of the results from Branch-A, Branch-B, and Branch-C indicates that relying on a single decision paradigm may be insufficient for modeling complex, multi-level diseases such as osteoporosis. Branch-A provides high discriminative power and clear decision boundaries, serving as a strong baseline for fundamental diagnostic scenarios. Branch-B, through ordinal learning, addresses the ordered nature of disease severity and enables clinically more meaningful modeling of osteoporosis progression. In contrast, the Branch-C ensemble structure aggregates complementary representations from multiple ViT backbones, achieving the most balanced outcomes in both accuracy and decision stability. These findings highlight that, given the nature of osteoporosis, which involves both local tissue alterations and global structural patterns, multi-scale and multi-representation approaches are critical for reliable clinical decision-making.

Future work may extend the proposed ViT-FESEL framework in several directions. First, its generalizability can be further validated using larger and multi-center datasets. Second, more advanced ordinal loss functions or uncertainty-aware learning strategies may be incorporated into Branch-B to further improve ordinal classification performance. For Branch-C, dynamic weighting schemes or learnable ensemble mechanisms could be explored to increase decision fusion flexibility. Moreover, integrating explainable artificial intelligence (XAI) techniques into the ViT-FESEL framework would enable clinicians to interpret model predictions more transparently and with greater trust. In this regard, the proposed approach is considered not only applicable to osteoporosis detection but also as a scalable foundation for other musculoskeletal disorders characterized by similar structural changes.

DECLARATION ON THE USE OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

The authors acknowledge that generative artificial intelligence tools were used to assist with language editing and translation of this manuscript.

REFERENCES

[1] Hong, N., Cho, S.W., Shin, S., Lee, S., Jang, S.A., Roh, S., Kim, K.M. (2020). Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography. Journal of Bone and Mineral Research, 38(6): 887-895. https://doi.org/10.1002/jbmr.4814

[2] Zhang, B., Yu, K., Ning, Z., Wang, K., et al. (2020). Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. Bone, 140: 115561. https://doi.org/10.1016/j.bone.2020.115561

[3] Xue, L., Qin, G., Chang, S., Luo, C., et al. (2023). Osteoporosis prediction in lumbar spine X-ray images using the multi-scale weighted fusion contextual transformer network. Artificial Intelligence in Medicine, 143: 102639. https://doi.org/10.1016/j.artmed.2023.102639

[4] Sarmadi, A., Razavi, Z.S., van Wijnen, A.J., Soltani, M. (2024). Comparative analysis of vision transformers and convolutional neural networks in osteoporosis detection from X-ray images. Scientific Reports, 14(1): 18007. https://doi.org/10.1038/s41598-024-69119-7

[5] Liu, J., Wang, H., Shan, X., Zhang, L., et al. (2024). Hybrid transformer convolutional neural network-based radiomics models for osteoporosis screening in routine CT. BMC Medical Imaging, 24(1): 62. https://doi.org/10.1186/s12880-024-01240-5

[6] Huang, J., Gao, J., Li, J., Gao, S., Cheng, F., Wang, Y. (2025). Transformer-based deep learning model for predicting osteoporosis in patients with cervical cancer undergoing external-beam radiotherapy. Expert Systems with Applications, 273: 126716. https://doi.org/10.1016/j.eswa.2025.126716

[7] Wani, I.M., Arora, S. (2023). Osteoporosis diagnosis in knee X-rays by transfer learning based on convolution neural network. Multimedia Tools and Applications, 82(9): 14193-14217. https://doi.org/10.1007/s11042-022-13911-y

[8] Shen, M. (2024). Utilizing deep learning for osteoporosis diagnosis through knee X-ray analysis. In 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024), pp. 553-560. https://doi.org/10.2991/978-94-6463-512-6_58

[9] Qureshi, M.B., Sani, M., Raza, A., Qureshi, M.S., et al. (2025). Deep-learning based osteoporosis classification in knee X-rays using transfer-learning approach. Scientific Reports, 15(1): 38448. https://doi.org/10.1038/s41598-025-24338-4.

[10] Naguib, S.M., Saleh, M.K., Hamza, H.M., Hosny, K.M., Kassem, M.A. (2024). A new superfluity deep learning model for detecting knee osteoporosis and osteopenia in X-ray images. Scientific Reports, 14(1): 25434. https://doi.org/10.1038/s41598-024-75549-0

[11] Hong, N., Cho, S.W., Lee, Y.H., Kim, C.O., Kim, H.C., Rhee, Y., Kim, K.M. (2025). Deep learning-based identification of vertebral fracture and osteoporosis in lateral spine radiographs and DXA vertebral fracture assessment to predict incident fracture. Journal of Bone and Mineral Research, 40(5): 628-638. https://doi.org/10.1093/jbmr/zjaf050

[12] Yeoh, P.S.Q., Lai, K.W., Goh, S.L., Hasikin, K., Hum, Y.C., Tee, Y.K., Dhanalakshmi, S. (2021). Emergence of deep learning in knee osteoarthritis diagnosis. Computational Intelligence and Neuroscience, 2021(1): 4931437. https://doi.org/10.1155/2021/4931437

[13] Tamai, K., Imanishi, K., Terakawa, M., Uematsu, M., et al. (2025). Deep learning algorithm for identifying osteopenia/osteoporosis using cervical radiography. Scientific Reports, 15(1): 25274. https://doi.org/10.1038/s41598-025-11285-3

[14] Multi-Class Knee Osteoporosis X-Ray Dataset. https://www.kaggle.com/datasets/mohamedgobara/multi-class-knee-osteoporosis-x-ray-dataset, accessed on Jan. 26, 2026.

[15] Genç, H., Koç, C., Yüzgeç Özdemir, E., Özyurt, F. (2025). An innovative approach to classify meniscus tears by reducing vision transformers features with elasticnet approach. The Journal of Supercomputing, 81(4): 602. https://doi.org/10.1007/s11227-025-07103-2

[16] Wang, F., Ren, S., Zhang, T., Neskovic, P., Bhattad, A., Xie, C., Yuille, A. (2026). ViT-5: Vision Transformers for The Mid-2020s. arXiv preprint arXiv:2602.08071. https://doi.org/10.48550/arXiv.2602.08071

[17] Balezo, G., Trullo, R., Planas, A.P., Decencière, E., Walter, T. (2026). MIPHEI-ViT: Multiplex immunofluorescence prediction from H&E images using ViT foundation models. Computers in Biology and Medicine, 206: 111564. https://doi.org/10.1016/j.compbiomed.2026.111564

[18] Özyurt, F., Koç, C., Özdemir, E.Y. (2025). Segment-aware contrastive representation learning with vision transformers: TransCon-Skin. IEEE Access, 14: 763-777. https://doi.org/10.1109/ACCESS.2025.3645694

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint:arXiv:2010.11929.

[20] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pp. 10347-10357. https://doi.org/10.48550/arXiv.2012.12877

[21] Li, Y., Huang, Y., He, N., Ma, K., Zheng, Y. (2024). Improving vision transformer for medical image classification via token-wise perturbation. Journal of Visual Communication and Image Representation, 98: 104022. https://doi.org/10.1016/j.jvcir.2023.104022

[22] Das, B.K., Zhao, G., Islam, S., Re, T.J., Comaniciu, D., Gibson, E., Maier, A. (2024). Co-ordinate-based positional embedding that captures resolution to enhance transformer's performance in medical image analysis. Scientific Reports, 14(1): 9380. https://doi.org/10.1038/s41598-024-59813-x

[23] Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 12901: 36-46. https://doi.org/10.1007/978-3-030-87193-2_4

[24] Gao, Y., Zhou, M., Metaxas, D.N. (2021). UTNet: A hybrid transformer architecture for medical image segmentation. International conference on medical image computing and computer-assisted intervention, 12903: 61-71. https://doi.org/10.1007/978-3-030-87199-4_6

[25] Lai, T. (2024). Interpretable medical imagery diagnosis with self-attentive transformers: A review of explainable AI for health care. BioMedInformatics, 4(1): 113-126. https://doi.org/10.3390/biomedinformatics4010008

[26] Hu, Y., Mu, N., Liu, L., Zhang, L., Jiang, J., Li, X. (2024). Slimmable transformer with hybrid axial-attention for medical image segmentation. Computers in biology and medicine, 173: 108370. https://doi.org/10.1016/j.compbiomed.2024.108370.

[27] Tang, W., Yang, Z., Song, Y. (2023). Disease-grading networks with ordinal regularization for medical imaging. Neurocomputing, 545: 126245. https://doi.org/10.1016/j.neucom.2023.126245

[28] Gao, Z., Zhao, H., Wu, Z., Wang, Y., et al. (2024). Coral-CVDs: A consistent ordinal regression model for cardiovascular diseases grading. International Workshop on Ophthalmic Medical Image Analysis, 15188: 73-82. https://doi.org/10.1007/978-3-031-73119-8_8

[29] Swiecicki, A., Li, N., O'Donnell, J., Said, N., et al. (2021). Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. Computers in biology and medicine, 133: 104334. https://doi.org/10.1016/j.compbiomed.2021.104334