Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 11 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i11.56865

PAPER

Advancing Osteoporosis Diagnosis through State-of-the-Art CNNs and Vision Transformers with Ensemble Strategies

Israa S. Abed[1,2](✉), Abeer Twakol Khalil[3], Hanan M. Amer[3], Samer Mahmoud Mohamed Ali[4], Mohamed Maher Ata[5]

[1]Department of Biomedical Engineering, Al-Khwarizmi College of Engineering, University of Baghdad, Baghdad, Iraq

[2]Biomedical Engineering Program, Faculty of Engineering, Mansoura University, Mansoura, Egypt

[3]Electronics and Communications Engineering Department, Faculty of Engineering, Mansoura University, Mansoura, Egypt

[4]Orthopedic Surgery Department, Faculty of Medicine, Mansoura University, Mansoura, Egypt

[5]School of Computational Sciences and Artificial Intelligence (CSAI), Zewail City of Science and Technology, Giza, Egypt

israasafa@kecbu.uobaghdad.edu.iq

ABSTRACT

Osteoporosis is a common bone disorder marked by reduced mineral density and microarchitectural deterioration, increasing fracture risk. Early, accurate detection is vital for clinical intervention and personalized care. This study applies deep learning to binary classification of osteoporosis (normal vs. osteoporotic) using medical imaging. A curated dataset of bone-related images was used to train and evaluate advanced models and ensemble strategies. Evaluated architectures include EfficientNetB2, InceptionV3, InceptionResNetV2, ResNet50V2, Xception, Vision Transformer (ViT_B32), and Faster R-CNN. Accuracy served as the main metric. ResNet50V2 outperformed all with 97.83% accuracy, ahead of EfficientNetB2 and ViT_B32 (95.65%), InceptionV3 and Xception (95.22%), InceptionResNetV2 (93.91%), and Faster R-CNN (76.96%). Ensembles—average, weighted, and hard voting—further improved accuracy to 96.96% and 96.09%. The results validate the benefit of ensemble learning in boosting model robustness. ResNet50V2 stands out as the top single model, and ensemble techniques show strong promise for reliable, automated osteoporosis detection. These findings support deploying deep learning in clinical radiology for early diagnosis and decision support.

KEYWORDS

osteoporosis, deep learning, medical image classification, bone mineral density, ResNet50V2, ensemble learning, Vision Transformer, convolutional neural network (CNN), automated diagnosis, medical imaging

1 INTRODUCTION

Osteoporosis is a progressive bone disease characterized by reduced bone mineral density and structural deterioration, significantly increasing the risk of fractures and posing a global health burden, particularly among the elderly [1]. While DXA remains the diagnostic gold standard, its limited accessibility and high cost highlight the need for

Abed, I. S., Khalil, A. T., Amer, H. M., Mohamed Ali, S. M., Ata, M. M. (2025). Advancing Osteoporosis Diagnosis through State-of-the-Art CNNs and Vision Transformers with Ensemble Strategies. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(11), pp. 116–131. https://doi.org/10.3991/ijoe.v21i11.56865

Article submitted 2025-05-26. Revision uploaded 2025-07-10. Final acceptance 2025-07-10.

alternative approaches [2]. Advances in deep learning have enabled the development of AI-driven diagnostic tools capable of automating medical imaging tasks, including osteoporosis detection [3]. However, selecting the most effective model remains challenging due to varying trade-offs in accuracy, complexity, and generalizability. This study addresses the gap by systematically evaluating six state-of-the-art models—EfficientNetB2, InceptionV3, InceptionResNetV2, ResNet50V2, Xception, and ViT_B32 for binary classification of osteoporosis from medical images, using metrics such as accuracy, precision, recall, and F1-score. Ensemble strategies, including majority and weighted voting, were also applied to enhance predictive performance by aggregating outputs from top-performing models. The study leverages a curated dataset and assigns weights based on individual model accuracy to optimize ensemble performance. Results aim to identify the most robust and deployable model or ensemble, contributing to the development of scalable, AI-driven osteoporosis screening tools that may also generalize to broader medical imaging applications. The remainder of this paper is organized as follows: Section 2 reviews related work on deep learning in medical image analysis with a focus on osteoporosis detection; Section 3 details the methodology, including dataset, preprocessing, model architecture, and evaluation criteria; Section 4 present the experimental results, section 5 present comparison and discussion and the section 6 represent conclusion and future work.

2 RELATED WORKS

Osteoporosis detection using deep learning and artificial intelligence has witnessed rapid advancements in recent years, with diverse techniques enhancing diagnostic precision and clinical applicability. A.O.M and Mohan Kumar [4] proposed a hybrid ResNet50-GRU architecture for predicting osteoporosis from knee X-rays, achieving high performance with 95.65% accuracy and an F1-score of 95.49%. Similarly, Rasool et al. [5] introduced a weighted ensemble combining DenseNet121 and EfficientNetB0, reaching a validation accuracy of 96.52%, outperforming individual models through synergistic learning. Sarmadi et al. [6] demonstrated that Vision Transformers (ViTs), particularly ViT16, surpassed CNNs such as VGG16 in classifying knee images into normal, osteopenic, and osteoporotic classes, with better localization of pathological regions. Transfer learning has played a pivotal role, as shown by Sarhan et al. [7], who utilized AlexNet, VGG16, ResNet50, and XceptionNet, achieving 97.5% and 92.0% accuracy for binary and multiclass classification, respectively. Dodamani and Danti [8] used VGG16, DenseNet121, and InceptionV2, attaining 81.2% accuracy with high sensitivity but moderate specificity.

Mohammed and George [9] leveraged a 55-layer CNN to reach 98.91% accuracy on Dataset A and 96.61% on Dataset B, while Wani and Arora [1] found AlexNet most effective at 91.1%. Ensemble learning also proved beneficial, as Naguib et al. [10] developed a multi-branch CNN achieving 85.42% accuracy, and Abubakar et al. [11] showed GoogLeNet outperforming VGG-16 and ResNet50 with 90% accuracy. Hwang et al. [12] proposed MVCTNet, a multi-view CT model achieving an AUC of 0.9640, surpassing ResNet-18 and EfficientNetB0. Few-shot learning (FSL) approaches have addressed limited-data scenarios. Xie et al. [13] designed an FSL model for knee X-rays that exceeded radiologist accuracy (0.728 mean accuracy), with performance improving when combined with human interpretation. For opportunistic screening, Zhang et al. [14] introduced a broad-learning system (BLS) using lumbar spine radiomics, achieving an AUC of 0.802. Peng et al. [15] applied deep learning on CT to estimate QCT-derived BMD with high correlation ($r = 0.996$ training, 0.981 test), and Oh et al. [16] automated BMD measurement with 77.7% accuracy.

Tsai et al. [17] extended this to chest X-rays, achieving an AUC of 0.930 and identifying high-risk patients with a hazard ratio of 2.59. Feature engineering also contributed significantly. Mebarkia et al. [18] used handcrafted features such as HOG and LPQ, achieving 89.66% accuracy. Ryu et al. [19] combined semantic segmentation with multitask learning, reporting a Dice score of 0.947. Kiran and Areeckal [20] used wavelet-based texture features with k-NN, reaching 78.24% accuracy. Muzaffar et al. [21] developed OsteoNet, integrating LPQ with spatial and channel attention, achieving 74.1% accuracy on the ISBI 2014 dataset. On the population level, Je et al. [22] used machine learning (e.g., XGBoost) on a Korean female cohort, achieving 70.5% accuracy and F1-score of 0.738, outperforming conventional tools. Meta-analyses by Amani et al. [23] and Inigo et al. [24] further validated CNN-based osteoporosis detection, reporting pooled sensitivity and specificity of 86% and 89% and AUCs of 0.94 and 0.878, respectively, highlighting both the potential and the need for improved generalizability. Notable contributions in this field have been made by researchers such as those cited in [9] and [8].

3 MATERIALS AND METHODS

The block diagram in Figure 1 outlines the deep learning workflow for osteoporosis classification, starting with data preprocessing (including histogram equalization, edge enhancement, resizing, and normalization) on four datasets. Data is split into training, validation, and testing sets (70-15-15). Models including EfficientNetB2, InceptionV3, InceptionResNetV2, ResNet50V2, Xception, ViT_B32 and Faster RCNN are evaluated using metrics such as accuracy, F1-score, ROC AUC, specificity, and Jaccard index. Ensemble methods (average, hard, and weighted voting) are applied to improve performance. Finally, model results are compared and analyzed.

3.1 Dataset collection

The dataset used for osteoporosis classification is composed of knee X-ray images aggregated from four different public sources on Kaggle: the Osteoporosis Knee X-ray Dataset [25], the Osteoporosis Knee dataset (Preprocessed 128×256) [26], the Osteoporosis dataset [27], and [28] from Mendeley database. These sources collectively provided a total of 1567 images, which were labeled as either osteoporotic or normal (healthy). The images from all sources were combined and then split into training, validation, and testing sets. The splitting strategy allocated 1101 images for training (70%), 236 for validation (15%), and 130 for testing (15%), ensuring that the training set had the most data for model learning while keeping sufficient data for validation and unbiased testing. Table 1 summarizes the dataset distribution across the splits and classes.

Table 1. Distribution of images in the dataset by split and class

Dataset Split	Ratio	#Images
Training	70%	1101
Validation	15%	236
Testing	15%	230
Overall	100%	1567

Due to the absence of patient identifiers in the public datasets, the split was performed at the image level, and patient-level overlap could not be explicitly controlled.

3.2 Data preprocessing

Proper preprocessing was crucial before feeding the X-ray images into the deep learning models. The following steps were carried out in sequence on the dataset.

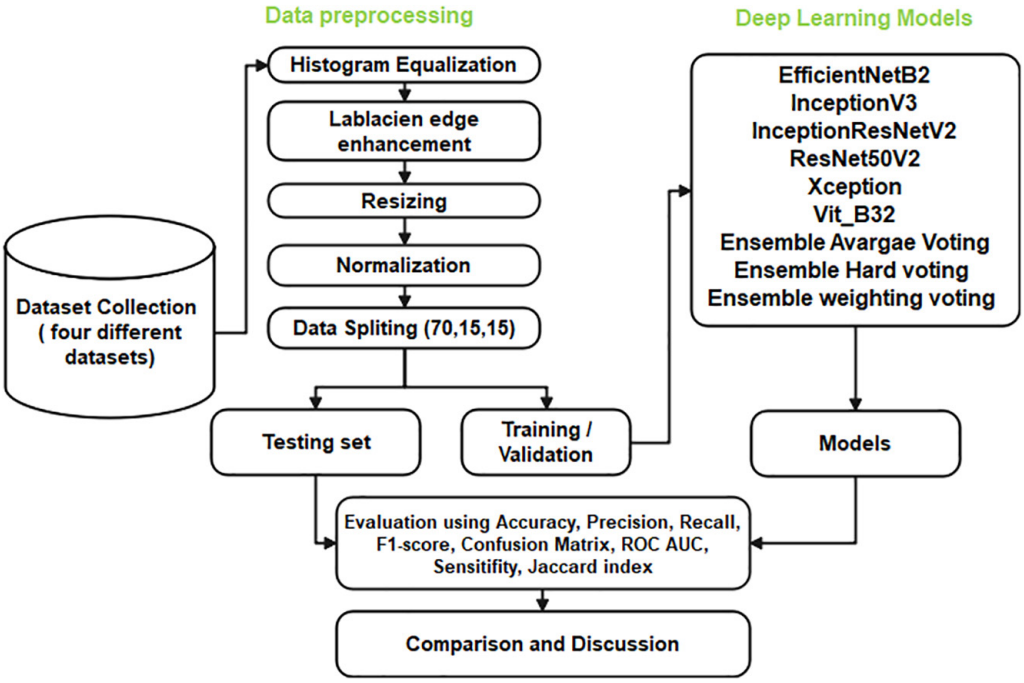

Fig. 1. Block diagram for the osteoporosis classification

Fig. 2. Image before and after histogram equalization

Contrast enhancement. A histogram equalization technique was applied to each image to improve contrast. Histogram equalization redistributes the pixel intensity values to span the full range, making bone structures more distinguishable. This step is particularly helpful for X-ray images, as it enhances the visibility of edges and texture in bone regions. In fact, histogram equalization is widely used in the medical imaging field—it is "typically applied to X-ray scans and CT scans to improve the radiograph's contrast," aiding doctors in better interpreting the images [29]. By enhancing contrast, the models can more easily learn discriminative features between healthy and osteoporotic bone tissue. Figure 2 represents an image before and after histogram equalization.

The process involves dividing the input image into a grid of tiles, computing a histogram for each tile, and applying a clip limit to prevent noise amplification. Excess pixels are redistributed to other bins, ensuring balanced contrast enhancement across the image. Let $H_i(k)$ be the histogram of the i-th tile, where k is the intensity level (from 0 to $L-1$ for an L-level image). Let *ClipLimit* be the predefined threshold. For each tile, the number of pixels exceeding the *ClipLimit* is in Eq (1):

$$Excess_i = \sum_{k=0}^{L-1} max(0, H_i(k) - ClipLimit) \tag{1}$$

The $Excess_i$ pixels are then redistributed among all the bins of the histogram for the i-th tile. A common approach is to distribute them uniformly, but other methods can be used. The number of pixels to be added to each bin can be calculated as in Eq (2):

$$Add_i = \frac{Excess_i}{N_{bins}} \tag{2}$$

where N_{bins} is the number of bins in the histogram (equal to L for a full histogram). This redistribution effectively flattens the histogram and thus reduces the local contrast.

After clipping and redistribution, the cumulative distribution function (CDF) for each tile's histogram is calculated. The CDF, $CDF_i(k)$, represents the cumulative probability of pixel values up to level k in the i-th tile in Eq (3):

$$CDF_i(k) = \sum_{j=0}^{k} P_i(j) = \sum_{j=0}^{k} \frac{H_i'(j)}{N_{pixels_i}} \tag{3}$$

where $H_i'(j)$ is the adjusted histogram of the i-th tile after clipping and redistribution, and N_{pixels_i} is the total number of pixels in the i-th tile. CDF is then used as a transformation function to map the original pixel values in each tile to new intensity values, thus performing local histogram equalization with contrast limiting as in Eq (4):

$$Pixel_{out_i} = (L-1).CDF_i(Pixel_{in_i}(x, y)) \tag{4}$$

Laplacian edge enhancement. Once the Laplacian of an image is computed as in Eq (5), it can be used to enhance edges. The Laplacian of the image at a pixel location (x, y) is equal to the sum of the intensity values of all its eight neighboring pixels (right, left, top, bottom, and the four diagonals), minus eight times the intensity value of the pixel itself.

$$\nabla^2 I(x, y) = I(x+1, y) + I(x-1, y) + I(x, y+1) + I(x, y-1) - 4I(x, y) \tag{5}$$

A simple edge enhancement technique involves adding a scaled version of the Laplacian to the original image as in Eq (6):

$$I_{enhanced}(x, y) = I(x, y) - c \cdot \nabla^2 I(x, y) \tag{6}$$

Where $I_{enhanced}(x, y)$ is the intensity of the enhanced image at pixel location (x, y), $I(x, y)$ is the intensity of the original image at pixel location (x, y), $\nabla^2 I(x, y)$ is the Laplacian of the original image at pixel location (x, y) and c is a scaling factor that controls the strength of the edge enhancement as in Figure 3.

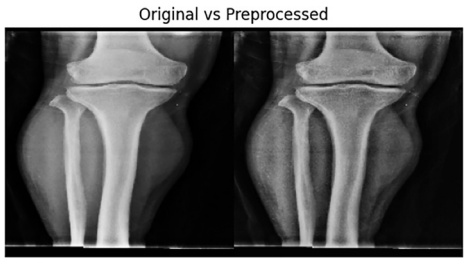

Fig. 3. Image before and after Laplacian edge enhancement

Resizing. All images were resized to a uniform dimension of 196×196 pixels with 3 color channels. This size was chosen to reduce computational load and memory usage while retaining essential features. The 3-channel format was used to match the input requirements of pre-trained CNN models (which expect RGB inputs). For grayscale X-ray images, this meant replicating the single channel into an RGB image or reading the image in RGB mode so that the input shape becomes 196×196×3.

Normalization. The normalized intensity (Inormalized) of a pixel in Eq (7) is obtained by dividing the original intensity (Ioriginal) of that pixel by the value two hundred and fifty-five. This equation scales the pixel intensity values from the range [0, 255] to the range [0, 1].

$$I_{normalized} = \frac{I_{original}}{255} \tag{7}$$

This type of normalization is often used in machine learning and image processing as it can help to ensures that input data has a consistent range, enhancing algorithm performance and preventing features with larger values from dominating the outcome.

3.3 Deep learning models

To classify osteoporosis from the knee X-rays, we leveraged several deep learning models via transfer learning. Six different state-of-the-art convolutional neural network (CNN) architectures (and one Transformer-based model) pre-trained on ImageNet were employed as feature extractors: EfficientNetB2 [30], InceptionV3 [31], InceptionResNetV2 [32], ResNet50V2 [33], Xception [33], and Vision Transformer (ViT-B32) [34]. These diverse architectures were chosen to capture a wide range of feature representations from the images. All models were used with an input shape of (196, 196, 3), but VIT uses (224, 224, 3), include_top = False (meaning the final classification layer of the original pre-trained model is removed), and weights = 'imagenet' to initialize the convolutional layers with learned features from the large ImageNet dataset. There are three other ensemble models, Ensemble Deep learning (average using the five models), hard voting, and weighted voting. Using ImageNet-pretrained weights provides a strong starting point for feature extraction, as these models have already learned to detect general shapes, edges, and textures that can be relevant to medical imaging as well. Each model was customized by adding a new classification head atop the pre-trained base: a GlobalAveragePooling2D layer to convert feature maps into a 1D vector, reducing parameters and overfitting—followed by a Dense layer with 2 softmax-activated neurons for binary classification (osteoporotic vs. normal). The resulting architecture was: Pre-trained CNN (no top) → GlobalAveragePooling2D → Dense (2, softmax). All six models were trained individually, and three ensemble strategies—Average Voting, Hard Voting, and Weighted Voting were applied to

enhance classification performance. The following algorithm illustrates the ensemble average voting implementation. To improve classification robustness, we use ensemble learning with hard voting, combining predictions from EfficientNet, Xception, ResNet, InceptionV3, and InceptionResNet, where the final output is based on majority vote. To further boost performance, weighted voting is applied, assigning more influence to models with higher validation accuracy. Additionally, we implement the Vision Transformer (ViT-B32), which uses self-attention to capture global image features. It is integrated into a custom architecture with normalization and dense layers for binary classification and trained using standard optimization techniques.

To integrate object detection with classification, we combine Faster R-CNN for region detection with a custom ResNet50V2-based CNN for binary classification, leveraging both localization and classification strengths. All models, including CNNs and ensembles, were trained under consistent conditions: batch size of 16, 100 epochs, 3-channel inputs (RGB or grayscale), categorical cross-entropy loss, and Adam optimizer with a learning rate of 1e-3. Pre-trained base models were fully fine-tuned to adapt ImageNet weights to osteoporosis-specific features, with only the classification head modified. Several callback mechanisms were employed to optimize training. ModelCheckpoint saved the model whenever the validation loss reached a new minimum, ensuring the best-performing model was retained. Reduce LROnPlateau adjusted the learning rate when the validation loss plateaued for multiple epochs, enabling fine-tuning at lower rates. Early stopping halted training if validation loss did not improve for 25 epochs, preventing excessive overfitting and unnecessary computation. The final model used for evaluation was the one saved by ModelCheckpoint, ensuring optimal generalization to unseen data. The selected models were chosen for their strong performance in medical image analysis. CNNs such as ResNet50V2, EfficientNetB2, InceptionV3, InceptionResNetV2, and Xception offer effective hierarchical feature extraction, suitable for detecting bone degradation. EfficientNetB2 was favored for its balance of accuracy and efficiency, while ViT-B32 was included to assess the benefits of attention-based mechanisms. Faster R-CNN was tested despite being detection-oriented, to examine its potential in classification. Ensemble methods were used to enhance prediction robustness through model combination.

With this training strategy, each model was trained to minimize validation loss. Fine-tuning all layers meant each model could learn osteoporosis-specific features, while the callbacks ensured a well-tuned training process. All training was performed on the training set, with model selection guided by validation set performance. The independent test set was kept completely unseen during training for the final evaluation of each model and the ensembles.

3.4 Evaluation metrics

To assess the performance of the classification models, we used a comprehensive set of evaluation metrics. These include basic metrics derived from the confusion matrix (accuracy, precision, recall, and F1-score) as well as more diagnostic tools (confusion matrix itself and the ROC AUC) [35]. Below we define each metric, provide its mathematical formula, and explain how it reflects the model's performance: A confusion matrix is a 2×2 table that summarizes the prediction outcomes of a binary classifier against the true labels. It has four entries: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

In our context (osteoporosis detection), we can define "positive" as the osteoporotic class. In a confusion matrix table, the rows often represent the actual class and the columns represent the predicted class (or vice-versa) as shown in Table 2.

Table 2. Confusion matrix explanation

Actual/Predicted	Osteoporotic (Predicted Positive)	Normal (Predicted Negative)
Osteoporotic (Actual Positive)	TP	FP
Normal (Actual Negative)	FN	TN

From these four basic outcomes, various metrics are calculated to evaluate the classifier:

Accuracy, is the proportion of all predictions that the model got correct. It is a global measure of how often the classifier is right, considering both classes. Mathematically, accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Precision, also known as Positive Predictive Value, measures the accuracy of positive predictions–when the model predicts an image as osteoporotic, how often is it correct? It is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall, also called Sensitivity or True Positive Rate, measures the model's ability to identify all positive instances—i.e., how many of the truly osteoporotic cases did the model catch? Recall is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

F1-score, is the harmonic mean of precision and recall. It provides a single metric that balances both false positives and false negatives. The F1-score is given by:

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \tag{11}$$

The F1-score ranges from 0 to 1, with 1 indicating perfect precision and recall. It balances precision and recall, helping assess classification performance. A higher F1 suggests the model effectively identifies osteoporosis cases while minimizing false alarms. Beyond numeric metrics, the confusion matrix provides insight into raw counts of *TP*, *TN*, *FP*, and *FN*, helping interpret errors. More false negatives indicate low recall, while excess false positives suggest precision issues.

Also, we used the Jaccard Index, often referred to as intersection over union (IoU) in the context of tasks such as image segmentation and object detection, provides a measure of the overlap between the model's predictions and the actual ground truth. It quantifies how well the predicted region aligns with the true region of interest.

$$IoU = \frac{TP}{TP + FP + FN} \tag{12}$$

Additionally, we evaluate the ROC curve and AUC. The ROC curve illustrates the trade-off between true positive rate (Recall) and false positive rate across classification thresholds, with a better classifier pushing the curve toward the upper-left corner [36].

4 EXPERIMENTAL RESULTS

The performance of various deep learning models for osteoporosis classification was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, Area Under the Curve (AUC), specificity, and the Jaccard index. Table 3 summarizes the results for each model. As depicted in Table 3, the performance evaluation of various deep learning models for osteoporosis classification reveals significant insights into their effectiveness across multiple metrics. Among the individual models, ResNet50V2 emerged as the top performer, achieving an outstanding accuracy of 97.83%. This model also demonstrated superior precision, recall, and F1-score, each at 97.83% or higher, indicating a well-balanced ability to correctly identify both positive and negative cases. Its area under the curve (AUC) reached 99.60%, reflecting excellent discriminative capability, while its specificity of 99.10% suggests a very low false positive rate. The Jaccard Index, which measures the similarity between predicted and actual labels, was also the highest among all models at 95.83%, further confirming the model's robustness in classification tasks. Close behind ResNet50V2 were EfficientNetB2 and ViTB32, both achieving an accuracy of 95.65%. These models also maintained high precision, recall, and F1-scores, all hovering around 95.6%, with AUC values of 99.68% and 99.50%, respectively. EfficientNetB2 slightly outperformed ViTB32 in terms of the Jaccard Index (92.25% vs. 92.06%), suggesting a marginally better overlap between predicted and actual positive cases. These results highlight the effectiveness of both compound scaling in EfficientNet and attention mechanisms in ViTs for medical image classification tasks. Models such as InceptionV3, InceptionResNetV2, and Xception performed well, with accuracies between 93.91% and 95.22% and AUCs above 98%, though they showed slightly lower specificity and Jaccard Index values, indicating more false positives.

Table 3. Performance comparison of deep learning models

Models	Accuracy	Precision	Recall	F1-Score	AUC	Specificity	Jaccard Index
EfficientNetB2	0.9565	0.9599	0.9565	0.9564	0.9968	0.9099	0.9225
InceptionV3	0.9522	0.9524	0.9522	0.9521	0.9887	0.94	0.9127
InceptionResNetV2	0.9391	0.9396	0.9391	0.9391	0.9944	0.92	0.8906
ResNet50V2	0.9783	0.9786	0.9783	0.9783	0.9960	0.9910	0.9583
Xception	0.9522	0.9530	0.9522	0.9521	0.9944	0.93	0.9134
Vit_B32	0.9565	0.9570	0.9565	0.9565	0.9950	0.94	0.9206
Faster RCNN	76.96	0.8024	0.7696	0.7595	0.7634	0.5856	0.6788
Ensemble (average)	0.9696	0.9704	0.9696	0.9695	0.9688	0.95	0.9440
Ensemble (hard voting)	0.9609	0.9609	0.9609	0.9609	0.9607	0.95	0.9274
Ensemble (weighted voting)	0.9696	0.9704	0.9696	0.9695	0.9688	0.95	0.9440

In contrast, Faster R-CNN performed poorly across all metrics, with just 76.96% accuracy, 75.95% F1-score, and 58.56% specificity, highlighting its limitations in fine-grained classification tasks such as osteoporosis detection. To boost performance, three ensemble strategies average, hard, and weighted voting were used. Both average and weighted voting delivered the best results, each achieving 96.96% across accuracy, precision, recall, F1-score, and an AUC of 96.88%, highlighting the strength of ensemble learning. Hard voting followed closely with 96.09% accuracy. Figure 4 represents the ROC AUC score and curves for all proposed models.

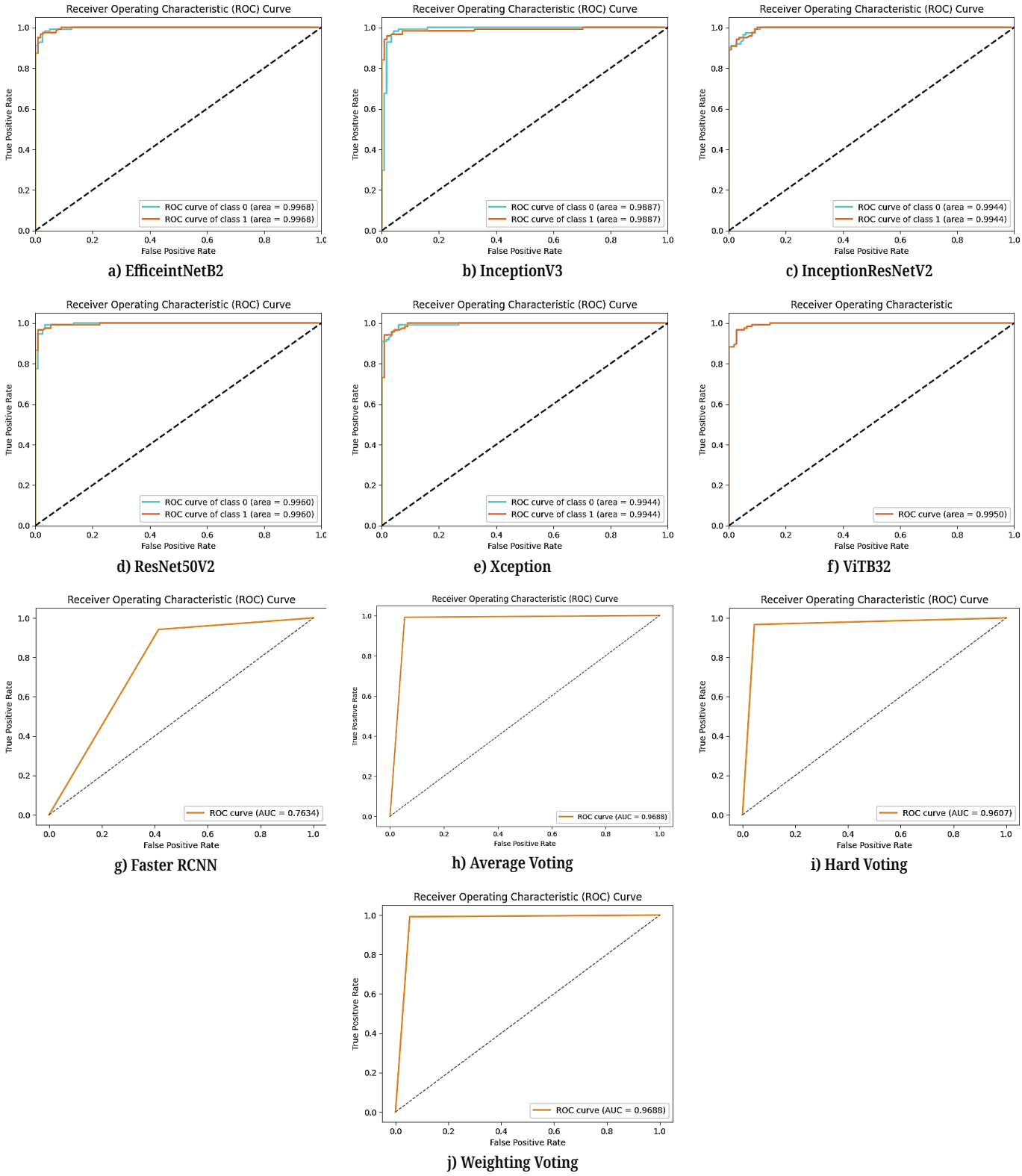

Fig. 4. ROC AUC scores for the proposed models

As seen in Figure 4, the AUC values provide a strong indication of each model's ability to distinguish between osteoporotic and non-osteoporotic cases. EfficientNetB2 achieved the highest AUC at 0.9968, closely followed by ResNet50V2 (0.9960), ViT_B32

(0.9950), and InceptionResNetV2 and Xception (both 0.9944). These high scores reflect excellent model performance in terms of sensitivity and specificity across all thresholds. InceptionV3 also performed well with an AUC of 0.9887, indicating reliable classification, though slightly below the top-tier models. In contrast, Faster R-CNN had a significantly lower AUC of 0.7634, suggesting it struggled to consistently differentiate between classes and is less suitable for this task. The ensemble models showed strong and consistent performance. Both average voting and weighted voting achieved an AUC of 0.9688, while hard voting followed closely at 0.9607. These results highlight the benefit of combining predictions from multiple models to improve overall classification reliability, even if slightly below the best individual models.

5 COMPARISON AND DISCUSSION

Our study, using ResNet50V2, achieved the highest accuracy of 97.83%, surpassing all previous works in knee X-ray analysis and osteoporosis classification. This superior performance underscores the effectiveness of our approach in leveraging transfer learning with fine-tuning strategies tailored for medical imaging.

Table 4. Comparison of proposed ResNet50V2 model with related works

Authors	Model	Results (%) or Metrics
A. O. M & Mohan Kumar [4] (2024)	ResNet50-GRU	95.65% Accuracy
Rasool et al. [5] (2023)	DenseNet121 + EfficientNetB0 (Ensemble)	95.76–96.52% Accuracy
Sarmadi et al. [6] (2024)	Vision Transformer (ViT)	VGG16 (61.7%), ViT (63.8%)
Amany M. Sarhan et al. [7] (2024)	VGG-19	92.0% (multi-class), 97.5% (binary)
Dodamani and Danti [8] (2023)	VGG16, VGG19, DenseNet121, InceptionV2	81.20% Accuracy
Mohammed & George [9] (2023)	55-layer DCNN (Transfer Learning)	98.91% (Dataset A), 96.61% (Dataset B)
Wani & Arora [1] (2023)	AlexNet	91.1% Accuracy
Naguib et al. [10] (2024)	Superfluity DL (multi-CNN)	85.42% (Dataset1), 79.39% (Dataset2)
Abubakar et al. [11] (2023)	GoogLeNet, VGG-16	90.0% (gray), 87.0% (RGB)
Hwang et al. [12] (2023)	MVCTNet	AUC 0.9640, Sensitivity 81.33%, Specificity 90.67%
Xie et al. [13] (2024)	Few-shot Learning (FSL)	Accuracy 0.728, Sensitivity 0.774
Zhang et al. [14] (2023)	Broad-learning System (BLS)	AUC 0.802, Sensitivity 78.2%, Specificity 82.2%
Peng et al. [15] (2024)	DL on CT Scans	Correlation: 0.996 (train), 0.981 (test), 0.94 (independent)
Tsai et al. [17] (2024)	AI (Chest X-ray)	AUC 0.930 (internal), 0.892 (external)
Mebarkia et al. [18] (2023)	Handcrafted (HOG, LPQ)	89.66%
Ryu et al. [19] (2023)	Semantic Segmentation + MTL	Dice 0.947, Detection Accuracy 96%
S. K. S. Kiran & A. S. Areeckal [20] (2025)	Wavelet + ML (k-NN)	78.24%
A. W. Muzaffar et al. [21] (2025)	OsteoNet (CNN + Attention)	74.1%
M. Je et al. [22] (2025)	XGBoost	70.5%
F. Amani et al. [23] (2020)	Meta-analysis (DL Models)	Sensitivity 86%, Specificity 89%, AUC 0.94
S. A. Inigo et al. [24] (2023)	Meta-analysis (ML on hip DXA)	AUC 0.878, Sensitivity 0.844, Specificity 0.781
Our Study	ResNet50V2 (split 70/15/15)	97.83% Accuracy

As shown in Table 4, this comparative analysis highlights the wide range of approaches and performance levels in osteoporosis detection using deep learning and machine learning models across different studies. Sarmadi et al. [6] reported relatively modest performance using ViT, achieving 63.8% accuracy, slightly outperforming traditional CNNs such as VGG16 (61.7%). These results, though limited by dataset imbalance and size, suggest the potential of ViTs with sufficient data. In contrast, methods such as ResNet50-GRU (A. O. M and Mohan Kumar [4]) and KONet ensemble (Rasool et al. [5]) achieved significantly higher accuracies of 95.65% and up to 96.52%, respectively demonstrating the strength of hybrid and ensemble architectures. The highest accuracy was achieved by Mohammed & George [9] using a transfer learning-based 55-layer DCNN, reaching 98.91% (Dataset A). Similarly, the current study (our study) using ResNet50V2 achieved 97.83%, placing it among the top-performing models.

Amany M. Sarhan et al. [7] also achieved high results with 92% (multi-class) and 97.5% (binary), underscoring the importance of classification type in model performance. Other studies such as Wani and Arora [1] (AlexNet, 91.1%), Naguib et al. [10] (multi-CNN, 85.42%), and Abubakar et al. [11] (GoogLeNet/VGG-16, 90%) confirm that classical CNN architectures remain competitive when tuned effectively. Meanwhile, more novel techniques such as MVCTNet by Hwang et al. [12] (AUC 0.9640) and few-shot learning by Xie et al. [13] (accuracy 72.8%) show promise in specialized scenarios with limited data. Meta-analyses by Amani et al. [23] and Inigo et al. [24] reported AUCs of 0.94 and 0.878, respectively, affirming that deep learning models, when evaluated across multiple studies, maintain high sensitivity and specificity. On the other hand, simpler handcrafted or traditional ML approaches (e.g., Kiran & Areeckal [20], Muzaffar et al. [21], and Je et al. [22]) lag behind with accuracies ranging from 70% to 78%, reflecting limitations in feature expressiveness and generalization.

To assess performance differences, McNemar's test was applied to test set predictions. As shown in Table 5, ResNet50V2 significantly outperformed InceptionResNetV2 ($p = 0.0159$). Differences with EfficientNetB2, Xception, and InceptionV3 were not statistically significant, though Xception and InceptionV3 approached significance ($p \approx 0.0771$). These results support ResNet50V2 as the most reliable model.

Table 5. McNemar's test results comparing ResNet50V2 with other deep learning models

Compared Models	χ^2 Value	*p*-Value	Significance
ResNet50V2 vs EfficientNetB2	1.231	0.2673	No
ResNet50V2 vs Xception	3.125	0.0771	No (but close)
ResNet50V2 vs InceptionV3	3.125	0.0771	No (but close)
ResNet50V2 vs InceptionResNetV2	5.818	0.0159	**Yes ($p < 0.05$)**

Faster R-CNN demonstrated significantly lower performance across all metrics compared to the other models. This can be attributed to its architecture, which is optimized for object detection rather than global image classification. Its reliance on region proposal networks and bounding box regression makes it less effective for binary classification tasks where fine-grained, holistic interpretation of the entire image is crucial. This architectural mismatch likely contributed to its lower specificity, AUC, and overall accuracy in detecting osteoporotic changes.

In summary, ensemble learning, transfer learning with deep CNNs, and attention-based models consistently outperform basic CNNs and handcrafted methods. ViTs, while not top-performing in this study, hold future promise if trained with larger,

balanced datasets. The findings reinforce that model architecture, data quality, and class balance are decisive factors in osteoporosis detection accuracy. Our study utilizing ResNet50V2 outperformed all prior approaches, achieving 97.83% accuracy, demonstrating its superiority in feature extraction and classification performance.

6 CONCLUSION

This study demonstrated the effectiveness of deep learning models in osteoporosis classification, with ResNet50V2 standing out as the most accurate and balanced model, achieving 97.83% accuracy and the highest Jaccard Index. EfficientNetB2 and ViT_B32 also showed strong and consistent performance, emphasizing the impact of architecture and scaling strategies. While traditional models such as Faster R-CNN underperformed due to their detection-focused design, ensemble techniques, particularly weighted and average voting offered reliable, high-performing alternatives by combining model strengths. The results confirm that leveraging ensemble learning and transfer learning with fine-tuning enhances diagnostic accuracy and robustness, making these models highly suitable for clinical decision support systems in osteoporosis detection. Despite the promising results, this study has several limitations. First, the limited dataset size may increase the risk of overfitting. Second, the models were trained on a specific bone type and imaging modality, limiting generalizability. Third, while data augmentation was applied during training, it did not improve test accuracy and occasionally reduced it. Finally, the lack of external validation limits the real-world applicability of the findings. Future work should address these issues using larger, more diverse, and multi-institutional datasets. This study offers practical value by benchmarking the performance of several deep learning architectures for osteoporosis classification, aiding model selection for clinical implementation. It also contributes theoretically by highlighting how different model architectures CNN-based, attention-based, and detection-oriented behave on a specialized medical imaging task. These insights can inform future research aimed at optimizing model design for osteoporosis and related diagnostic challenges.

6.1 Data availability

The datasets used in this study are publicly available from online repositories. Specifically, knee X-ray images were obtained from publicly shared datasets on Kaggle and Mendeley. All datasets are anonymized and used in accordance with their respective terms of use. As the data are publicly available and contain no identifiable patient information, ethical approval was not required for this study.

7 REFERENCES

[1] I. M. Wani and S. Arora, "Osteoporosis diagnosis in knee X-rays by transfer learning based on convolution neural network," *Multimed. Tools Appl.*, vol. 82, pp. 14193–14217, 2023. https://doi.org/10.1007/s11042-022-13911-y

[2] M. Velagapudi and J. Kethar, "MRI detection of osteoporosis bone fractures opposed to DEXA scan," *J. Student Res.*, vol. 12, no. 4, 2023. https://doi.org/10.47611/jsrhs.v12i4.5902

[3] S. Wang *et al.*, "Fully automated deep learning system for osteoporosis screening using chest computed tomography images," *Quant. Imaging Med. Surg.*, vol. 14, no. 4, pp. 2816–2827, 2024. https://doi.org/10.21037/qims-23-1617

[4] A. O. M. and M. M. Kumar, "An effective deep learning based model for the prediction of osteoporosis from knee X-ray images," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 3, pp. 480–489, 2024.

[5] M. J. A. Rasool, S. Ahmad, U. Sabina, and T. K. Whangbo, "KONet: Toward a weighted ensemble learning model for knee osteoporosis classification," *IEEE Access*, vol. 12, pp. 5731–5742, 2024. https://doi.org/10.1109/ACCESS.2023.3348817

[6] A. Sarmadi, Z. S. Razavi, A. J. van Wijnen, and M. Soltani, "Comparative analysis of vision transformers and convolutional neural networks in osteoporosis detection from X-ray images," *Sci. Rep.*, vol. 14, 2024. https://doi.org/10.1038/s41598-024-69119-7

[7] A. M. Sarhan *et al.*, "Knee osteoporosis diagnosis based on deep learning," *Int. J. Comput. Intell. Syst.*, vol. 17, 2024. https://doi.org/10.1007/s44196-024-00615-4

[8] P. S. Dodamani and A. Danti, "Transfer learning-based osteoporosis classification using simple radiographs," *Int. J. Online Biomed. Eng. (iJOE)*, vol. 19, no. 8, pp. 66–87, 2023. https://doi.org/10.3991/ijoe.v19i08.39235

[9] A. Z. Mohammed and L. E. George, "Diagnosis of osteoporosis using transfer learning in the same domain," *Int. J. Online Biomed. Eng. (iJOE)*, vol. 19, no. 14, pp. 142–159, 2023. https://doi.org/10.3991/ijoe.v19i14.42163

[10] S. M. Naguib, M. K. Saleh, H. M. Hamza, K. M. Hosny, and M. A. Kassem, "A new superfluity deep learning model for detecting knee osteoporosis and osteopenia in X-ray images," *Sci. Rep.*, vol. 14, 2024. https://doi.org/10.1038/s41598-024-75549-0

[11] U. B. Abubakar, M. M. Boukar, S. Adeshina, and S. Dane, "Transfer learning model training time comparison for osteoporosis classification on knee radiograph of RGB and grayscale images," *WSEAS Trans. Electron.*, vol. 13, pp. 45–51, 2022. https://doi.org/10.37394/232017.2022.13.7

[12] D. H. Hwang, S. H. Bak, T. J. Ha, Y. Kim, W. J. Kim, and H. S. Choi, "Multi-view computed tomography network for osteoporosis classification," *IEEE Access*, vol. 11, pp. 22297–22306, 2023. https://doi.org/10.1109/ACCESS.2023.3252361

[13] H. Xie *et al.*, "A few-shot learning framework for the diagnosis of osteopenia and osteoporosis using knee X-ray images," *J. Int. Med. Res.*, vol. 52, no. 9, 2024. https://doi.org/10.1177/03000605241274576

[14] B. Zhang *et al.*, "Development and validation of a feature-based broad-learning system for opportunistic osteoporosis screening using lumbar spine radiographs," *Acad. Radiol.*, vol. 31, no. 1, pp. 84–92, 2023. https://doi.org/10.1016/j.acra.2023.07.002

[15] T. Peng *et al.*, "A study on whether deep learning models based on CT images for bone density classification and prediction can be used for opportunistic osteoporosis screening," *Osteoporos. Int.*, vol. 35, pp. 117–128, 2024. https://doi.org/10.1007/s00198-023-06900-w

[16] S. Oh *et al.*, "Evaluation of deep learning-based quantitative computed tomography for opportunistic osteoporosis screening," *Sci. Rep.*, vol. 14, 2024. https://doi.org/10.1038/s41598-023-45824-7

[17] D. J. Tsai, C. Lin, C. S. Lin, C. C. Lee, C. H. Wang, and W. H. Fang, "Artificial intelligence-enabled chest X-ray classifies osteoporosis and identifies mortality risk," *J. Med. Syst.*, vol. 48, 2024. https://doi.org/10.1007/s10916-023-02030-2

[18] M. Mebarkia, A. Meraoumia, L. Houam, and S. Khemaissia, "X-ray image analysis for osteoporosis diagnosis: From shallow to deep analysis," *Displays*, vol. 76, p. 102343, 2023. https://doi.org/10.1016/j.displa.2022.102343

[19] S. M. Ryu *et al.*, "Diagnosis of osteoporotic vertebral compression fractures and fracture level detection using multitask learning with U-Net in lumbar spine lateral radiographs," *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 3452–3458, 2023. https://doi.org/10.1016/j.csbj.2023.06.017

[20] S. K. S. Kiran and A. S. Areeckal, "Classification of osteoporotic X-ray images using wavelet texture analysis and machine learning," *Int. J. Comput. Digit. Syst.*, vol. 17, no. 1, 2025. https://doi.org/10.12785/ijcds/1570996365

[21] A. W. Muzaffar, F. Riaz, and M. Tahir, "OsteoNet – A framework for identifying osteoporosis in bone radiograph images using attention based VGG network," *IEEE Access*, vol. 13, pp. 25175–25185, 2025. https://doi.org/10.1109/ACCESS.2025.3538828

[22] M. Je, S. Hwang, S. Lee, and Y. Kim, "Development and evaluation of a machine learning model for osteoporosis risk prediction in Korean women," *BMC Women's Health*, vol. 25, 2025. https://doi.org/10.1186/s12905-025-03669-4

[23] S. Yang, B. Yin, W. Cao, C. Feng, G. Fan, and S. He, "Diagnostic accuracy of deep learning in orthopaedic fractures: A systematic review and meta-analysis," *Clin. Radiol.*, vol. 75, no. 9, pp. 713.e17–713.e28, 2020. https://doi.org/10.1016/j.crad.2020.05.021

[24] S. A. Inigo, R. Tamilselvi, and M. P. Beham, "A review on imaging techniques and artificial intelligence models for osteoporosis prediction," *Curr. Med. Imaging*, vol. 20, pp. 1–18, 2023. https://doi.org/10.2174/1573405620666230608091911

[25] https://www.kaggle.com/datasets/stevepython/osteoporosis-knee-xray-dataset

[26] https://www.kaggle.com/datasets/sachinkumar413/Osteoporosis-knee-dataset-preprocessed128x256

[27] https://www.kaggle.com/datasets/mrmann007/Osteoporosis

[28] I. Majeed Wani, "Knee X-ray osteoporosis database," Mendeley Data, vol. VI, 2021. https://doi.org/10.17632/fxjm8fb6mw.1

[29] A. Rosebrock, "OpenCV histogram equalization and adaptive histogram equalization (CLAHE)," PyImageSearch, 2021. https://pyimagesearch.com/2021/02/01/opencv-histogram-equalization-and-adaptive-histogram-equalization-clahe/

[30] V. Agarwal, "Complete architectural details of all EfficientNet models," medium, 2020. https://medium.com/data-science/complete-architectural-details-of-all-efficientnet-models-5fd5b736142

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *Proceedings of the AAAI Conf. Artif. Intell.*, vol. 31, no. 1, pp. 4278–4284, 2017. https://doi.org/10.1609/aaai.v31i1.11231

[32] A. Demir and F. Yilmaz, "Inception-ResNet-v2 with Leakyrelu and averagepooling for more reliable and accurate classification of chest X-ray images," in *2020 Med. Technol. Congr. (TIPTEKNO)*, 2020, pp. 1–4. https://doi.org/10.1109/TIPTEKNO50054.2020.9299232

[33] D. Gupta, "Transfer learning pretrained models in deep learning," Analyticsvidya, 2017. https://www.analyticsvidhya.com/blog/2017/06/transfer-learning-the-art-of-fine-tuning-a-pre-trained-model/

[34] S. Cuenat and R. Couturier, "Convolutional neural network (CNN) vs vision transformer (ViT) for digital holography," in *2022 2nd Int. Conf. Comput. Control Robot. (ICCCR)*, 2022, pp. 235–240. https://doi.org/10.1109/ICCCR54399.2022.9790134

[35] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Min. Knowl. Manag. Process (IJDKP)*, vol. 5, no. 2, pp. 1–11, 2015. https://doi.org/10.5121/ijdkp.2015.5201

[36] K. H. Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Casp. J. Intern. Med.*, vol. 4, no. 2, pp. 627–635, 2013.

8 AUTHORS

Israa S. Abed is an Assistant Lecturer at Al-Khwarizmi College of Engineering, Biomedical Engineering Department, University of Baghdad, Iraq (E-mail: israasafa@kecbu.uobaghdad.edu.iq).

Abeer Twakol Khalil is currently working as an Associate Professor at the Electronics and Communications Department, Faculty of Engineering, Mansoura University. As well as she is Former the Executive manager of the Postgraduate Medical

Engineering Program in Mansoura University, as well as she is the Executive manager of Artificial Intelligence Engineering program, Egypt (E-mail: abeer.twakol@mans.edu.eg).

Hanan M. Amer is currently working as an Associate Professor at the Electronics and Communications Department, Faculty of Engineering, Mansoura University. As well as she is the Director of the Graduate Biomedical Engineering Program for postgraduate studies at Mansoura University, Egypt (E-mail: eng_hanan_2007@mans.edu.eg).

Samer Mahmoud Mohamed Ali is an Associate Professor at Orthopedic Surgery Department, Faculty of Medicine, Mansoura University, Egypt (E-mail: samermahmoud@mans.edu.eg).

Mohamed Maher Ata is an Associate Professor at the School of Computational Sciences and Artificial Intelligence (CSAI), Zewail City of Science and Technology, Egypt. He specializes in Data Science and Artificial Intelligence (DSAI). His research interests span a wide spectrum of AI-driven technologies, including signal and image processing, machine learning, deep learning, computer vision, multimedia analysis, and video understanding. Dr. Ata has authored more than 50 peer-reviewed research articles published in high-impact ISI and SJR-indexed journals. His interdisciplinary contributions extend across artificial intelligence, deep learning, machine learning, computer vision, natural language processing (NLP), interpretable and explainable AI, nature inspired computation, biomedical engineering, astrophysics, cryptography, electrical communications, bioinformatics, software-defined networking (SDN), optimization, and intelligent transportation systems (ITS) (E-mail: momaher@zewailcity.edu.eg).