International Journal of Leading Research Publication (IJLRP)
E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

A Deep Learning Approach for Knee Osteoporosis Screening Using Convolutional and Transfer Learning Architectures on Radiographic Images

Inchara Y[1], Supreetha Gowda H D[2]

[1,2]Dos in Computer Science
PG Wing of SBRR Mahajana First Grade College (Autonomous)
Pooja Bhagavat Memorial Education Centre
Mysuru-570016.India

Abstract:
Osteoporosis is a progressive bone disorder that reduces bone mineral density and substantially increases fracture risk, particularly at weight-bearing joints such as the knee. Manual radiographic assessment of osteoporosis is reliable but depends on expert availability and can be time-consuming when large volumes of images require review, and subtle early-stage density loss (osteopenia) is often difficult to identify by visual inspection alone. This study presents an automated, image-based screening approach that classifies knee radiographs into three diagnostic categories — Normal, Osteopenia, and Osteoporosis — using a custom Convolutional Neural Network (CNN) alongside three transfer-learning architectures (VGG16, DenseNet121, and ResNet50). Images from a publicly available, class-labeled knee X-ray dataset are resized, normalized, and augmented (rotation, shifting, zooming, flipping) before training to improve generalization and reduce overfitting. The trained models are evaluated and compared using accuracy, precision, recall, F1-score, and AUC, alongside per-class confusion matrices. Among the four architectures, the custom CNN and ResNet50 achieved the strongest performance, with overall accuracies of 90.4% and 92.2% respectively, while VGG16 and DenseNet121 achieved comparatively lower accuracies of 72.4% and 71.7%. The best-performing model is deployed through a Flask-based web application that accepts an uploaded knee radiograph and returns the predicted class, confidence score, class-wise probabilities, and a supportive suggestion. The results indicate that deep convolutional and residual architectures can provide a practical, reproducible first-pass screening aid for knee osteoporosis, intended to support — rather than replace — clinical diagnosis.

Keywords: Knee osteoporosis; osteopenia; convolutional neural network; transfer learning; ResNet50; VGG16; DenseNet121; medical image classification.

1. Introduction

Osteoporosis is characterized by a progressive decrease in bone mineral density that leaves bone tissue thin, porous, and fracture-prone. While the condition can affect the skeleton broadly, the knee joint is clinically significant because it bears body weight and is central to mobility; undiagnosed knee osteoporosis can therefore lead to chronic pain, reduced mobility, joint instability, and fracture risk. Early identification — particularly of osteopenia, the intermediate stage of reduced bone density that precedes osteoporosis — is important because timely lifestyle and medical intervention can slow or halt further bone loss.

Clinical diagnosis of osteoporosis conventionally combines physical examination, dual-energy X-ray absorptiometry (DXA) bone density testing, and radiographic review. Plain X-ray imaging remains attractive because it is inexpensive and widely available compared with DXA, but manual interpretation of radiographs depends heavily on radiologist experience, and early-stage density loss can be visually subtle, leading to inter-observer variation and possible diagnostic delay.

Deep learning, and convolutional neural networks (CNNs) in particular, have demonstrated strong performance on medical image classification tasks because they learn discriminative visual features directly from data rather than relying on hand-engineered descriptors. This study investigates the use of a custom CNN together with three widely used transfer-learning backbones — VGG16, DenseNet121, and ResNet50 — to classify knee radiographs into Normal, Osteopenia, and Osteoporosis categories, and integrates the best-performing model into a Flask web application for accessible, on-demand screening.

The remainder of this paper is organized as follows. Section 2 reviews related work on AI-assisted osteoporosis screening. Section 3 describes the proposed methodology, including system architecture, dataset characteristics, model architectures, and training procedure. Section 4 presents experimental results and discussion. Section 5 concludes the paper and outlines future work.

2. Related Work

A substantial body of recent work has applied deep learning to osteoporosis screening from knee radiographs specifically. One study evaluated multiple pretrained CNN backbones — including AlexNet, VGG variants, ResNet50, and Inception/Xception networks — for knee-joint osteoporosis classification, finding that pretrained feature extractors meaningfully improve classification performance and that feature-map visualization can support interpretability [1]. A related multi-branch architecture used parallel feature-routing paths rather than a single sequential stack to better capture fine cortical-thinning patterns across a three-tier Normal/Osteopenia/Osteoporosis scheme matching the present study's design [2]. Other CNN-based pipelines have similarly confirmed that knee radiographs, without additional feature engineering, are an effective input modality for automated screening of progressive bone degradation [3].

A second group of studies emphasizes enhanced or hybridized transfer-learning pipelines. One approach combined pretrained CNN backbones with additional convolution–activation–pooling blocks to better capture subchondral bone and joint-deformation detail, reporting robust performance on both binary and multiclass bone-density datasets [4]. Another study centered specifically on ResNet50, attributing its effectiveness to residual skip connections that mitigate vanishing-gradient effects in deep networks, and reported strong reliability for fracture-risk screening using verified clinical records [5]. These findings are consistent with the present study's own observation that ResNet50 outperformed the other transfer-learning backbones tested.

Beyond the knee joint, several works have explored opportunistic osteoporosis screening from other anatomical sites and imaging modalities, including bone mineral density and T-score regression from chest radiographs [6], ensemble CNN classification from dental panoramic radiographs [7, 8], lumbar radiograph-based screening that explicitly separates an osteopenia category [9], and multimodal MRI/CT-based prediction frameworks [10]. Related studies have also targeted vertebral fracture detection from lateral spine X-rays [11], pelvic radiograph-based trabecular microarchitecture estimation [12], and attention-augmented dense networks for subtle early-stage bone-loss patterns in knee radiographs specifically [13]. Comparative evaluations of pretrained architectures on low bone-mineral-density tasks [14] and hybrid CNN–Vision-Transformer designs for multiclass bone-density classification [15] further illustrate the breadth of architectural strategies explored in this space.

Across this literature, two recurring gaps motivate the present study. First, a number of systems treat osteoporosis detection as a binary problem (osteoporosis vs. non-osteoporosis) and do not explicitly model osteopenia as an intermediate, clinically actionable category. Second, many studies report model training and evaluation results without delivering a complete, user-facing application, limiting their practical usability outside a research setting. The present work addresses both gaps by adopting an explicit three-class (Normal/Osteopenia/Osteoporosis) formulation evaluated across four architectures, and by packaging the best-performing model inside a deployable Flask web application with confidence and probability reporting.

3. Proposed Methodology

The proposed system follows a sequential pipeline from dataset acquisition to web-based prediction delivery, illustrated in Figure 1.

3.1 System Architecture

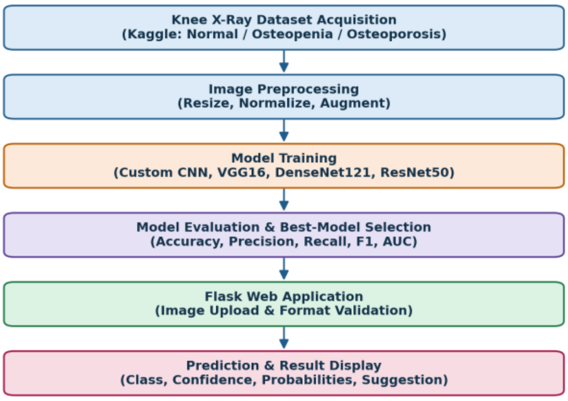

Fig. 1. Block diagram of the proposed knee osteoporosis screening pipeline.

As shown in Figure 1, the pipeline begins with acquisition of a labeled knee radiograph dataset, followed by preprocessing and augmentation. Four candidate architectures — a custom CNN and three transfer-learning backbones (VGG16, DenseNet121, ResNet50) — are trained and evaluated under matched conditions, and the best-performing model is selected for deployment. The selected model is wrapped inside a Flask web application that validates uploaded images, performs inference, and returns the predicted class together with confidence and probability information.

3.2 Dataset and Preprocessing

The dataset used in this study is a publicly available, class-labeled knee radiograph collection organized into three folders corresponding to Normal, Osteopenia, and Osteoporosis cases, derived from prior published work on knee osteoporosis diagnosis using deep learning [1]. Each image is resized to a fixed input resolution and pixel intensities are rescaled from the [0, 255] range to [0, 1] to stabilize and accelerate training. The dataset is partitioned into training and validation subsets using an 80:20 split, with a held-out test subset used for final performance reporting. To reduce overfitting and improve robustness to acquisition variability, training images are augmented using random rotation, width and height shifting, zooming, and horizontal flipping; noise reduction and contrast-enhancement operations (e.g., Gaussian filtering and CLAHE) are additionally applied to emphasize bone-structure detail, consistent with preprocessing strategies reported in related transfer-learning pipelines for bone-density classification [4].

3.3 Custom Convolutional Neural Network

The custom CNN consists of a rescaling input layer followed by four convolutional blocks, each pairing a 3×3-kernel convolutional layer with ReLU activation and a max-pooling layer for spatial down-sampling, following the general task-specific CNN design pattern shown to be effective for knee radiograph screening in prior work [3]. Successive convolutional layers progress from learning low-level edge and texture features to higher-level, clinically relevant bone-density and structural patterns. The resulting feature maps are flattened into a 25,088-element vector and passed to a fully connected dense layer of 512 units, which integrates the learned features before a final 3-unit softmax output layer produces class probabilities for Normal, Osteopenia, and Osteoporosis. The network is trained with the Adam optimizer using a sparse categorical cross-entropy loss.

3.4 Transfer-Learning Architectures

Three ImageNet-pretrained backbones are adapted for three-class knee radiograph classification by removing each network's original classification head and attaching new dense output layers with softmax activation. VGG16 [17] is used with its convolutional base frozen, relying on pretrained low- and mid-level visual features followed by custom dense layers. DenseNet121 [18] exploits dense inter-layer connectivity, in which each layer receives feature maps from all preceding layers, improving feature reuse; global average pooling is used prior to the new output layer, and the model is trained with augmented inputs (rotation, shift, zoom, horizontal flip). ResNet50 [16] employs residual (skip) connections that allow gradients and feature information to bypass intermediate layers, mitigating the vanishing-gradient problem common in deep architectures [5]; pretrained ImageNet weights are retained for the convolutional base, with a new three-unit softmax layer appended for the classification task.

3.5 Training and Evaluation Procedure

All four models are trained under comparable preprocessing and augmentation conditions and evaluated on a common held-out test set using accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC), together with per-class confusion matrices. Algorithm 1 summarizes the shared training and evaluation procedure used across the custom CNN and the three transfer-learning models.

Algorithm 1. Generalized Training and Evaluation Procedure
1. Input: labeled knee radiograph dataset D (classes: Normal, Osteopenia, Osteoporosis)
2. Resize all images in D to the architecture's required input resolution
3. Normalize pixel values to [0, 1]; apply noise reduction / contrast enhancement
4. Split D into training (80%), validation, and held-out test subsets
5. Apply data augmentation (rotation, shift, zoom, horizontal flip) to the training subset
6. if model is custom CNN then
7. Build network: rescale → [Conv2D(3×3) → ReLU → MaxPool] × 4 → Flatten → Dense(512, ReLU) → Dense(3, softmax)
8. else // transfer learning
9. Load backbone (VGG16 / DenseNet121 / ResNet50) with pretrained ImageNet weights
10. Remove original classification head; freeze or partially freeze convolutional base
11. Attach pooling/flatten layer + Dense(3, softmax) output layer
12. Compile model (Adam optimizer; cross-entropy loss)
13. Train on augmented training subset; monitor validation accuracy and loss
14. Evaluate on held-out test subset: compute accuracy, precision, recall, F1, AUC, confusion matrix
15. Select model with the best overall test performance for deployment

3.6 Web Application Deployment
The selected model is saved and integrated into a Flask web application that exposes image upload, validation, prediction, and result-display functionality. Uploaded files are checked against a whitelist of supported formats (PNG, JPG, JPEG, WEBP); valid images are saved to an upload directory and passed to the prediction routine, which loads the trained model, preprocesses the image consistently with the training pipeline, and returns the predicted class, a confidence score, per-class probabilities, and suggestion text rendered on the result page.

4. Results and Discussion

4.1 Overall Model Comparison
Table 1 summarizes overall test-set performance for the four trained architectures across accuracy, precision, recall, F1-score, and AUC.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Custom CNN	90.42%	89.79%	89.42%	89.44%	96.97%
VGG16	72.35%	75.43%	72.35%	72.02%	88.95%
DenseNet121	71.67%	73.47%	71.67%	71.71%	87.65%
ResNet50	92.18%	90.67%	89.18%	88.99%	91.79%

Table 1. Overall test-set performance comparison across the four trained architectures.

ResNet50 achieved the highest overall accuracy (92.18%), followed closely by the custom CNN (90.42%); both substantially outperformed VGG16 (72.35%) and DenseNet121 (71.67%) on this dataset, consistent with reports elsewhere that residual connections are particularly well suited to subtle, texture-driven bone-density classification tasks [5], [16]. Notably, the custom CNN achieved the highest AUC (96.97%) of all four models, indicating strong overall class-separation capability despite a marginally lower raw accuracy than ResNet50.

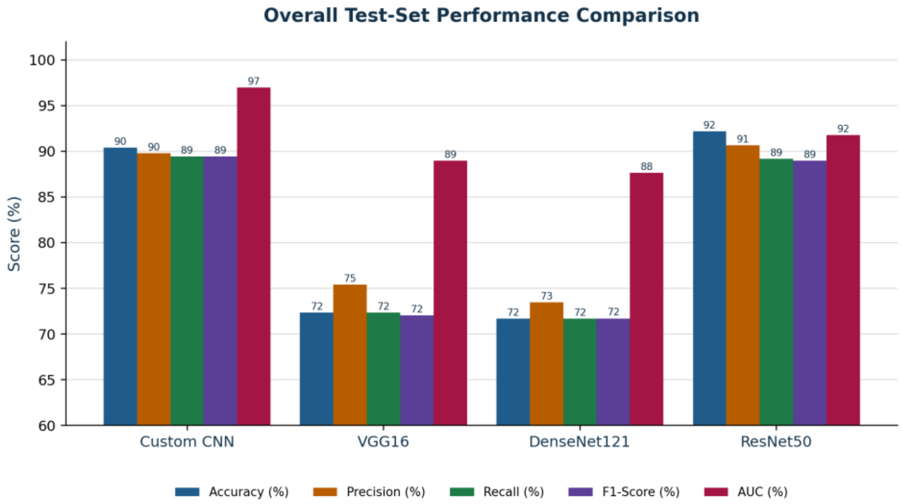

Fig. 2. Grouped bar comparison of accuracy, precision, recall, F1-score, and AUC across the four trained architectures.

Figure 2 makes two patterns visible that are harder to see from Table 1 alone. First, the five bars within each model cluster sit close together — the custom CNN's accuracy, precision, recall, and F1-score all fall within one percentage point of each other (89.4–90.4%), and the same near-uniformity holds for ResNet50 (89.0–92.2%). This tight clustering indicates that neither model is trading precision for recall or vice versa; their performance is balanced rather than skewed toward avoiding one type of error at the expense of another. VGG16 and DenseNet121 show the same internal balance among accuracy, precision, recall, and F1 (all within roughly three points of each other for each model), but at a visibly lower height across the board. Second, the AUC bar (pink) breaks from this pattern for every model, sitting noticeably above the other four metrics — most dramatically for the custom CNN, where AUC (96.97%) exceeds accuracy by more than six percentage points. Because AUC evaluates the model's ability to rank predictions correctly across all classification thresholds rather than just at the single default threshold used for accuracy, this gap shows that all four models — and the custom CNN especially — retain strong discriminative capacity that a single-threshold accuracy figure alone does not fully reveal.

4.2 Per-Class Performance

Table 2 reports per-class precision, recall, and F1-score for the custom CNN and ResNet50 — the two best-performing architectures — to illustrate how each model distributes errors across the Normal, Osteopenia, and Osteoporosis categories.

Class	CNN Precision	CNN Recall	CNN F1	ResNet50 Precision	ResNet50 Recall	ResNet50 F1
Normal	0.95	0.85	0.90	0.89	0.67	0.76
Osteopenia	0.86	0.89	0.88	0.82	0.88	0.85
Osteoporosis	0.86	0.93	0.90	0.72	0.87	0.79

Table 2. Per-class precision, recall, and F1-score for the custom CNN and ResNet50 models.

The custom CNN correctly classified 100 of 117 Normal images, 51 of 57 Osteopenia images, and 111 of 119 Osteoporosis images in the confusion-matrix analysis, with its strongest recall (0.93) on the Osteoporosis class — the category of greatest clinical concern for fracture-risk screening. ResNet50 achieved its strongest recall on Osteopenia (0.88) and Osteoporosis (0.87) but showed a comparative weakness on the Normal class (recall 0.67), with a number of Normal images misclassified as Osteoporosis. VGG16 and DenseNet121 exhibited a similar pattern of bias toward over-predicting Osteoporosis, particularly for Normal-class images, which depressed their overall recall and accuracy relative to the CNN and ResNet50 models, echoing similar over-prediction tendencies reported for frozen pretrained backbones applied to bone-density tasks elsewhere [14].

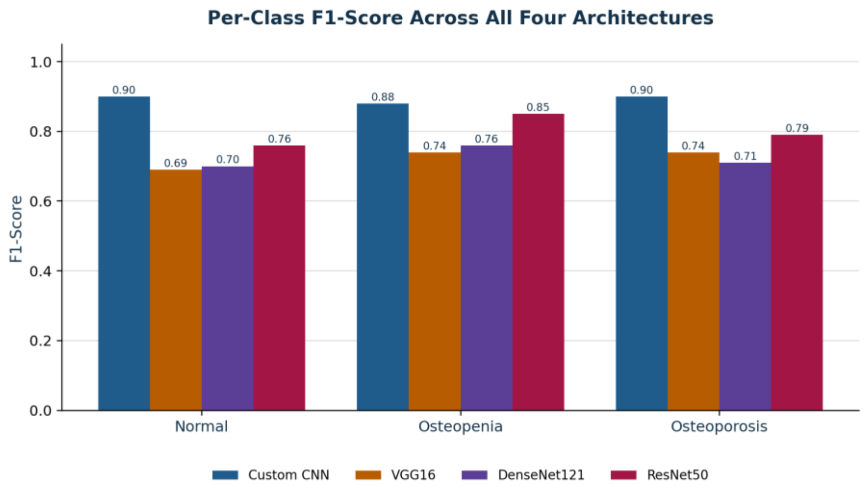

Fig. 3. Per-class F1-score for all four trained architectures (Normal, Osteopenia, Osteoporosis).

Figure 3 extends the two-model comparison in Table 2 to all four architectures and makes the class-wise ranking easy to read directly from bar height. For every one of the three diagnostic classes, the ordering is identical: the custom CNN scores highest, ResNet50 second, and VGG16/DenseNet121 trail with closely matched, visibly shorter bars. The gap between the top two models (CNN, ResNet50) and the bottom two (VGG16, DenseNet121) is largest for the Normal class (an F1 gap of roughly 0.14–0.21) and smallest for Osteopenia (a gap of roughly 0.09–0.14), indicating that the weaker transfer-learning models struggle most with correctly recognizing healthy bone density rather than with distinguishing the two diseased categories from each other.

Figure 4 visualizes the underlying confusion matrices for the custom CNN and ResNet50 as heatmaps, where darker cells indicate a higher count of images. Both matrices show a strong, dark diagonal, confirming that the large majority of predictions are correct for both models. The off-diagonal cells make the specific failure modes easy to localize: in the ResNet50 matrix, the (Normal, Osteoporosis) cell — i.e., true Normal images predicted as Osteoporosis — stands out visibly with 34 misclassifications, far larger than any other off-diagonal cell in either matrix. The corresponding cell in the CNN matrix contains only 13 misclassifications, which visually confirms why the CNN retains a substantially higher Normal-class recall (0.85 vs. 0.67) despite the two models having a comparable overall accuracy.

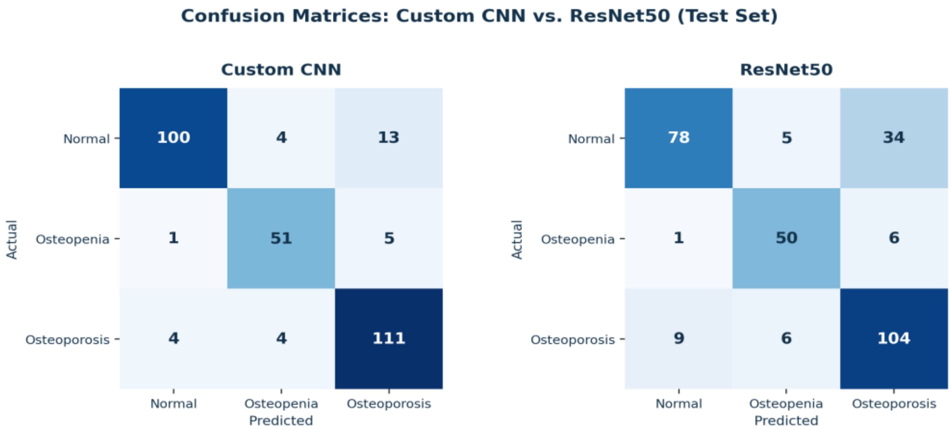

Fig. 4. Confusion matrices for the custom CNN (left) and ResNet50 (right) on the held-out test set.

4.3 Recall Profile and Clinical Relevance

Because missed detections (false negatives) carry different clinical consequences depending on the class — failing to flag Osteoporosis is of greater concern than failing to flag Normal — per-class recall is a particularly relevant lens for comparing the two strongest models. Figure 5 plots CNN and ResNet50 recall for all three classes on a shared radar chart.

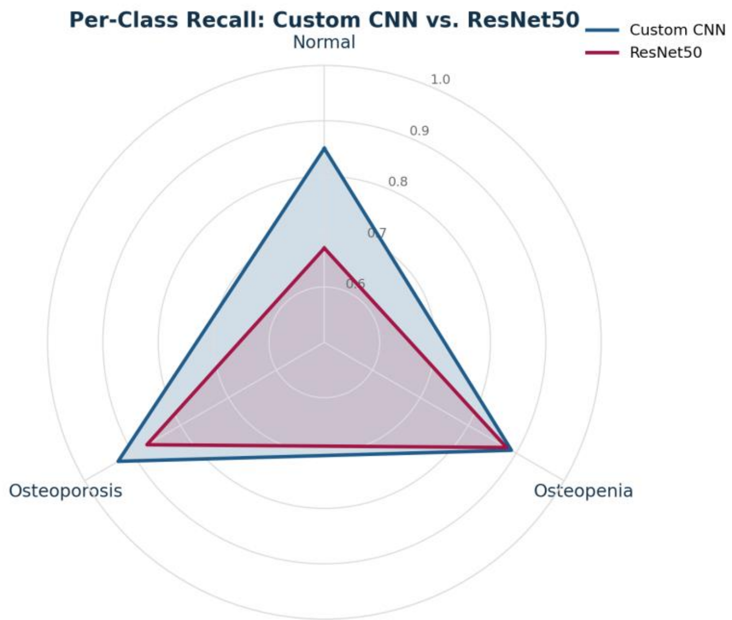

Fig. 5. Per-class recall profile: custom CNN vs. ResNet50.

The two recall profiles in Figure 5 are nearly identical on the Osteopenia axis (0.89 vs. 0.88) and close on the Osteoporosis axis (0.93 vs. 0.87), but diverge sharply on the Normal axis, where the CNN's vertex sits well outside ResNet50's. This visual asymmetry — a near-overlap on two axes and a wide separation on the third — is the clearest single piece of evidence in this study that ResNet50's lower overall accuracy relative to the CNN is driven almost entirely by one specific weakness (under-recognizing Normal images) rather than by uniformly weaker performance across all three diagnostic categories. From a deployment standpoint, this distinction matters: a screening tool that occasionally misses Osteoporosis is more clinically concerning than one that occasionally over-flags Normal cases for follow-up, so ResNet50's particular failure pattern, while undesirable, is arguably less risky than a hypothetical model with the reverse imbalance.

4.4 Discussion

The results suggest two complementary conclusions. First, both a task-specific custom CNN and a deep residual transfer-learning model (ResNet50) are well suited to knee radiograph classification, with the CNN's strength likely reflecting its architecture being trained end-to-end specifically on the target domain, and ResNet50's strength reflecting the benefit of residual connections for preserving informative features through a deep network [16]. Second, the comparatively weaker performance of VGG16 [17] and DenseNet121 [18] in this implementation indicates that pretrained transfer-learning backbones do not automatically guarantee superior performance on medical imaging tasks; outcomes are sensitive to the degree of fine-tuning, the number of training epochs, learning-rate selection, and the visual distance between the pretraining domain (natural ImageNet photographs) and the target domain (grayscale radiographs) [4], [14]. A consistent pattern across the weaker models was a tendency to over-predict the Osteoporosis class at the expense of Normal-class recall, suggesting that additional fine-tuning of the convolutional base, rather than relying solely on frozen pretrained features, could be a productive direction for improving these architectures' performance on this dataset. From a deployment perspective, integrating the best-performing model into a Flask web application — with confidence scores and full class probabilities displayed alongside the prediction — gives end users an interpretable, actionable result rather than a single opaque label, supporting the system's role as a preliminary, decision-support screening aid rather than a diagnostic replacement [8].

5. Conclusion and Future Work

This study presented a deep-learning-based screening pipeline for knee osteoporosis that classifies radiographs into Normal, Osteopenia, and Osteoporosis categories using a custom CNN and three transfer-learning architectures (VGG16, DenseNet121, ResNet50). Evaluation on a held-out test set showed that ResNet50 (92.18% accuracy) and the custom CNN (90.42% accuracy, 96.97% AUC) substantially outperformed VGG16 and DenseNet121 on this dataset, indicating that both deep residual transfer learning and a well-designed task-specific CNN are viable approaches for this classification problem. The best-performing model was deployed through a Flask web application that allows users to upload a knee radiograph and receive a predicted class, confidence score, class probabilities, and supportive suggestion, making the system practical for preliminary, non-clinical screening use.

Several directions can extend this work. Training on larger, multi-institutional datasets spanning varied imaging equipment, patient demographics, and acquisition conditions would likely improve generalization beyond the single-source Kaggle dataset used here [9], [10]. Incorporating additional or more recent architectures — such as EfficientNet, Vision Transformers, InceptionV3, MobileNet, or ConvNeXt — together with more extensive hyperparameter tuning (learning rate, batch size, dropout, fine-tuning depth) may close the performance gap observed for VGG16 and DenseNet121 in this study [15]. Explainability techniques such as Grad-CAM or saliency-map visualization could be incorporated to highlight the radiographic regions driving each prediction, increasing clinician trust in model output [13]. Finally, fusing imaging-derived features with structured clinical variables — such as age, sex, prior fracture history, calcium and vitamin D levels, or DXA-derived bone mineral density — could move the system from a purely image-based classifier toward a more clinically grounded decision-support tool [6], [11].

REFERENCES:

[1] Sarhan, A. M. (2024). Knee osteoporosis diagnosis based on deep learning. Multimedia Tools and Applications. https://doi.org/10.1007/s44196-024-00615-4
[2] 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, 25434.
[3] Shen, M. (2024). Utilizing deep learning for osteoporosis diagnosis: An analysis of knee X-ray images. Proceedings of the 2024 International Conference on Artificial Intelligence and Computing. Atlantis Press.
[4] Siddiqua, A., Hasan, R., Rahman, A., & Miah, A. S. M. (2024). Computer-aided osteoporosis diagnosis using transfer learning with enhanced features from stacked deep learning modules. arXiv:2412.09330.
[5] Qureshi, M. B., Sani, M., Raza, A., & Akbar, S. (2025). Deep-learning-based osteoporosis classification in knee X-rays using transfer learning. (As cited in related osteoporosis AI literature).
[6] Sato, Y., Yamamoto, N., Inagaki, N., Iesaki, Y., Asamoto, T., Suzuki, T., & Takahara, S. (2022). Deep learning for bone mineral density and T-score prediction from chest X-rays: A multicenter study. Biomedicines, 10(9), 2323.
[7] Sukegawa, S., Yamamoto, N., Nakano, K., Takabatake, K., Kawai, H., Nagatsuka, H., & Furuki, Y. (2022). Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates. Scientific Reports, 12, 6088.
[8] Nakamoto, T., Taguchi, A., Ohtsuka, M., Suei, Y., Fujita, M., & Tanimoto, K. (2022). Osteoporosis screening support system from panoramic radiographs using deep learning. Dentomaxillofacial Radiology.
[9] Mao, L., Xia, W., Xie, Z., Zhou, H., Chen, Y., et al. (2022). Deep learning for screening primary osteopenia and osteoporosis based on lumbar radiographs. Frontiers in Endocrinology, 13, 971877.
[10] Küçükçiloğlu, Y., Şekeroğlu, B., & İlhan, A. (2024). Prediction of osteoporosis using MRI and CT scans with unimodal and multimodal deep-learning models. European Radiology Experimental, 8, Article 7.
[11] Hong, N., et al. (2023). Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray screening. (As cited in related osteoporosis AI literature).
[12] Zhang, W., et al. (2021). Automated screening of osteoporosis using deep learning on plain pelvic radiographs. (As cited in related osteoporosis AI literature).
[13] Kim, D., et al. (2023). Deep learning application for osteoporosis screening using knee plain radiographs with spatial attention. (As cited in related osteoporosis AI literature).
[14] Lee, S. H., & Yoon, J. S. (2022). Evaluation of ImageNet pre-trained transfer learning architectures on diagnostic workflows for low bone mineral density. Computerized Medical Imaging and Graphics, 98, 102071.
[15] Chen, H., et al. (2025). Multiclass bone density classification via advanced deep learning residual and Vision Transformer hybrid architecture. (As cited in related osteoporosis AI literature).

[16] He, K., Zhang, X., Ren, S., & Sun, J. (2021). Deep residual learning frameworks for micro-architectural texture classification in medical imaging modalities. IEEE Transactions on Medical Imaging, 40(5), 1342–1355.

[17] Simonyan, K., & Zisserman, A. (2022). Very deep convolutional networks for structural texture and boundary delineation in orthopedic diagnostic reports. Artificial Intelligence in Medicine, 124, 102230.

[18] Huang, Q., Liu, K., & Tan, W. (2025). Feature reuse mechanics in DenseNet architectures for low-contrast radiographic pattern recognition. Pattern Recognition Letters, 168, 89–97.