A Deep Learning and Ensemble Approach for Osteoporosis Detection Using Conventional X-Ray Imaging

Sharmila Rathod[1], Adhya Jain[2], Kajal Gupta[3], Yashraj Gavale[4]
[1234] Department of Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Andheri (West), Mumbai – 400053, Maharashtra, India
[1] Associate Professor, MCT's Rajiv Gandhi Institute of Technology, Andheri (West), Mumbai – 400053, Maharashtra, India
[234] Student, MCT's Rajiv Gandhi Institute of Technology, Andheri (West), Mumbai – 400053, Maharashtra, India

Abstract - **Osteoporosis is a progressive skeletal disorder characterized by reduced bone mineral density and an increased risk of fractures, yet it remains significantly underdiagnosed worldwide. The current gold standard for diagnosis, Dual-energy X-ray Absorptiometry (DEXA), provides reliable assessment but is limited by high cost, restricted accessibility, and low availability in developing regions such as India, making large-scale screening challenging. In contrast, X-ray imaging is widely available, cost-effective, and routinely used in clinical practice, offering a practical alternative when combined with advanced computational methods.**

This study proposes a weighted ensemble framework based on custom Convolutional Neural Networks (CNNs) for automated osteoporosis detection using knee X-ray images. Two publicly available datasets from Kaggle were utilized to improve model robustness and generalization. An initial custom CNN achieved an accuracy of 95.1%; however, it exhibited comparatively lower recall, which is critical in medical diagnosis to minimize false negatives. To address this, multiple models with identical architectures but different initializations were combined using a weighted ensemble strategy, resulting in enhanced predictive performance.

The proposed approach leverages the accessibility of X-ray imaging and the robustness of ensemble deep learning to provide an efficient and scalable solution for osteoporosis detection.

Index Terms—**Osteoporosis, Deep Learning, CNN, Ensemble Learning, X-ray Imaging, Medical AI**

I. INTRODUCTION

Osteoporosis is a chronic skeletal disorder characterized by reduced bone mineral density (BMD) and deterioration of bone microarchitecture, leading to increased bone fragility and susceptibility to fractures. Common fracture sites include the hip, spine, and wrist, which are often associated with significant morbidity, long-term disability, and increased mortality, particularly among the elderly population. Due to its asymptomatic progression, osteoporosis is frequently referred to as a "silent disease," as it often remains undetected until a fracture occurs. According to the World Health Organization, approximately 200 million individuals worldwide are affected by osteoporosis, with nearly one in three women and one in five men over the age of 50 at risk of osteoporotic fractures. The increasing aging population further exacerbates the global burden of the disease.

A major challenge in osteoporosis management is its underdiagnosis and late detection. In clinical practice, the disease is often identified only after the occurrence of fragility fractures, by which point substantial bone loss has already occurred. Such delayed diagnosis leads to increased healthcare costs, reduced quality of life, and higher risk of recurrent fractures. Therefore, the development of early, reliable, and scalable screening methodologies is essential for effective disease management and prevention.

The current clinical gold standard for osteoporosis diagnosis is Dual-energy X-ray Absorptiometry (DEXA), which provides quantitative measurement of bone mineral density with high precision. Despite its effectiveness, DEXA has several limitations that hinder its widespread adoption, particularly in developing countries such as India. The procedure is relatively expensive, with costs typically ranging from ₹1500 to ₹5000 per scan, and requires specialized equipment as well as trained personnel for operation. Furthermore, the availability of DEXA systems per million population in India remains limited, with a higher concentration in urban healthcare facilities, thereby restricting access for rural and semi-urban populations. Consequently, a significant portion of the population remains unscreened, making DEXA unsuitable for large-scale or routine screening programs.

In contrast, conventional radiographic imaging (X-ray) is widely available, cost-effective, and routinely used across healthcare systems, including primary and rural healthcare centers. Although X-ray imaging is not traditionally used for direct measurement of bone mineral density, it contains structural information that can be leveraged for osteoporosis detection. Recent advancements in artificial intelligence have enabled the extraction of such latent features from medical images, thereby facilitating automated diagnosis using existing imaging infrastructure.

Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have demonstrated significant success in medical image analysis due to their ability to automatically learn hierarchical and discriminative feature representations. CNN-based approaches have achieved high performance in various diagnostic tasks, including bone abnormality detection and radiographic classification. However, in medical diagnosis, achieving high recall (sensitivity) is of paramount importance, as false negatives can lead to missed diagnoses and severe clinical consequences. Many existing approaches emphasize overall accuracy while overlooking recall, thereby limiting their clinical applicability.

To address these challenges, this study proposes a weighted ensemble framework built upon custom Convolutional Neural Network (CNN) architectures for osteoporosis detection using knee X-ray images. The approach involves training multiple models with identical architectures but different random initializations, and integrating their predictions through a weighted aggregation strategy to enhance model robustness and improve recall performance, which is critical in medical diagnosis.

The proposed work focuses on developing a custom CNN architecture specifically tailored for extracting discriminative features from radiographic images, while incorporating weighted ensemble learning to achieve more reliable and consistent predictions. By leveraging multiple models, the framework reduces the risk of misclassification and improves sensitivity toward osteoporotic cases. Furthermore, the study emphasizes the use of widely available X-ray imaging as a cost-effective alternative to conventional diagnostic methods, enabling scalable and accessible screening.

Overall, the proposed framework aims to provide an efficient, accurate, and practical solution for early osteoporosis detection, particularly suited for deployment in resource-limited settings where access to advanced diagnostic infrastructure is restricted.

II. RELATED WORK

Recent advancements in medical imaging and artificial intelligence have led to significant progress in automated osteoporosis detection. In this study, a comprehensive analysis of 11 research articles from reputed publishers such as IEEE and Springer was conducted to understand existing approaches, methodologies, and limitations.

A majority of existing works focus on Convolutional Neural Network (CNN)-based techniques for osteoporosis detection using medical images, particularly X-rays and CT scans. Several studies have also leveraged transfer learning using pre-trained architectures such as VGG, ResNet, InceptionNet, and XceptionNet to improve classification performance. While these approaches demonstrate high accuracy, they often rely on limited datasets and primarily optimize for accuracy rather than recall, which is critical in medical diagnosis. Additionally, many studies continue to depend on Dual-energy X-ray Absorptiometry (DEXA) as the reference standard, despite its limitations in terms of accessibility and cost.

Rasool, Ahmad, Sabina, and Whangbo (2024) introduced KONet, a weighted ensemble learning model for knee osteoporosis classification using X-ray images. Their method combined EfficientNetB0 and DenseNet121 through weighted prediction averaging (0.6 and 0.4 respectively), supported by data augmentation to address dataset limitations. KONet achieved higher accuracy, robustness, and smoother training dynamics compared to standalone CNNs, with excellent F1-scores and recall for osteoporotic cases. However, increased computational cost and model complexity limit its clinical scalability [1].

Thrivikrama Rao, Ramesh, Ghali, and Venugopala Rao (2022) proposed a U-Net based framework integrated with thermal wave imaging and clinical variables for osteoporosis diagnosis. Using over 1100 DXA-correlated radiographs, their ensemble achieved 99.23% accuracy, 98.76% sensitivity, and near-perfect AUC. Thermal imaging enhanced localization of disease by capturing bone density and heat features. Despite outstanding results, reliance on large annotated datasets and specialized imaging reduces immediate clinical adoption [2].

Sarhan et al. (2024) evaluated multiple CNN architectures (AlexNet, VGG-16, VGG-19, ResNet-50, InceptionNet, XceptionNet, and a custom CNN) on 1947 knee X-rays. Combining transfer learning with augmentation and normalization, they found VGG-19 to be the most effective, achieving 97.5% accuracy in binary classification and 92.0% in multiclass diagnosis. While performance was strong, challenges included subtle class differentiation and limited external validation [3].

Si, Zhang, Wang, and Zheng (2025) applied machine learning on NHANES data (demographics, labs, and questionnaires) for osteoporosis risk prediction in 8766 participants. Using advanced feature selection and data balancing, LightGBM emerged as the best model with AUC ≈ 0.97, F1 ≈ 0.91, and recall ≈ 0.92. SHAP-based interpretability confirmed critical predictors like age, sex, and medical history. However, reliance on self-reported data and geographical limitations require external validation [4].

Liu et al. (2025) used radiomics from abdominal CT scans for opportunistic osteoporosis screening. Texture features (GLCM, GLRLM, GLSZM, etc.) were extracted from lumbar trabeculae, and multiple ML models were tested on a multicenter dataset of 509 patients. Logistic Regression achieved the best internal AUC of 0.96, though external performance was lower due to scanner variability. While promising, radiomics is limited by time-intensive manual segmentation [5].

Chen et al. (2021) developed a multi-channel CNN for osteoporosis diagnosis using raw ultrasound RF signals of the distal radius, combined with clinical features. Processing four channels with sliding window augmentation, their model boosted accuracy from ~70% (traditional methods) to 84%. t-SNE visualization confirmed effective feature separation. Limitations included small sample size and focus on a single anatomical site [6].

Kim, Yoo, Oh, and Kim (2013) compared classical ML methods (SVM, Random Forest, ANN, Logistic Regression) against the Osteoporosis Self-Assessment Tool in 1674 Korean postmenopausal women. SVM with Gaussian kernel achieved AUC = 0.827 with balanced sensitivity and specificity (~77–78%), outperforming traditional assessments. However, results lacked generalizability across diverse populations and excluded imaging/biochemical markers [7].

Muzaffar, Riaz, and Tahir (2025) proposed OsteoNet, a hybrid CNN integrating Local Phase Quantization (LPQ) and attention mechanisms (channel and spatial). Evaluated on 174 calcaneus radiographs, it achieved 74.1% accuracy and reduced false negatives compared to baseline CNNs. Despite limited dataset size and sensitivity to image quality, the study demonstrated cost-effective screening potential [8].

Hwang et al. (2023) developed MVCTNet, a deep learning model using multi-view CT sagittal slices for osteoporosis and osteopenia detection. With 2883 DXA-labeled patients, MVCTNet achieved AUC = 0.964, sensitivity = 81.3%, and specificity = 90.7%, outperforming CNN baselines. Grad-CAM confirmed relevant vertebral focus. However, manual slice selection and absence of full 3D analysis were noted limitations [9].

Adams et al. (2021) investigated radiofrequency-based wrist wave propagation for osteoporosis detection. Using spectra between 30 kHz–2 GHz and a neural network classifier on 67 subjects, their approach achieved ~83% sensitivity and ~94% specificity. While promising for low-cost office-based screening, limitations included small sample size and anatomical restriction [10].

Despite significant advancements, several critical gaps remain in existing literature. First, many studies rely on relatively small or single-source datasets, limiting model generalizability. Second, a majority of approaches prioritize overall accuracy while overlooking recall (sensitivity), which is crucial in medical diagnosis to avoid false negatives. Third, although ensemble learning has shown promising results, it is either underexplored or associated with high computational complexity, reducing its practicality for real-world deployment. Additionally, dependence on specialized imaging modalities such as CT, thermal imaging, or ultrasound restricts scalability, particularly in resource-limited settings.

To address these limitations, the present study proposes a weighted ensemble framework based on custom CNN architectures using widely available knee X-ray images. By combining multiple models with different initializations, the proposed approach aims to enhance recall while maintaining high accuracy. Furthermore, the use of combined publicly available datasets improves model robustness and generalization. Unlike prior works, this study focuses on developing a cost-effective, scalable, and clinically applicable solution for early osteoporosis detection, particularly suitable for resource-constrained environments.

III. MATERIAL AND METHODOLOGY

3.1 Dataset

The dataset used in this study was obtained from publicly available sources on Kaggle, ensuring accessibility, reproducibility, and relevance to real-world clinical applications. Two knee X-ray datasets were selected and combined to create a comprehensive dataset for osteoporosis classification.

The datasets consist of radiographic images of the knee joint, a clinically significant region for assessing bone density loss and structural degradation associated with osteoporosis. The classification task was defined as a binary problem with two classes: Normal and Osteoporotic.

The first dataset, *Osteoporosis Knee X-ray Dataset (StevePython)* [11], contains 372 images (186 Normal and 186 Osteoporotic). It is well-balanced and consists of high-quality grayscale images with consistent resolution, making it suitable for initial experimentation, preprocessing validation, and model tuning.

The second dataset, *Multi-Class Knee Osteoporosis X-Ray Dataset (Mohamed Gobara)* [12], includes 1,573 images (780 Normal and 793 Osteoporotic). This dataset introduces variability in imaging conditions and patient demographics, enhancing model generalization and robustness.

After data cleaning, duplicate removal, and label harmonization, the datasets were merged to form a combined dataset of 1,945 images, with 966 Normal and 979 Osteoporotic cases. This near-balanced distribution helps reduce bias and ensures reliable model training.

Table 1: Composition of Osteoporosis Datasets Used in the Study

Dataset Name	Total Images	Normal	Osteoporotic
StevePython Dataset [11]	372	186	186
Mohamed Gobara Dataset [12]	1573	780	793
Combined Dataset	1945	966	979

The combination of datasets increases data diversity and volume, which improves generalization and reduces overfitting. It also introduces real-world variability, making the model more adaptable to clinical scenarios. The use of X-ray imaging provides a cost-effective alternative to traditional methods such as Dual-energy X-ray Absorptiometry.

3.2 Preprocessing

All input radiographic images were standardized through a preprocessing pipeline to ensure uniformity and improve model performance. Images were resized to a fixed resolution of 224 × 224 pixels to match the input requirements of the convolutional neural network. Pixel intensity values were normalized to the range [0, 1] to facilitate faster convergence during training.

To enhance model generalization and reduce overfitting, data augmentation techniques were applied, including random rotations, horizontal flipping, and zoom transformations. These augmentations simulate real-world variability in medical imaging conditions.

3.3 Data Splitting

The dataset was divided into three subsets to enable effective training and evaluation of the model. The training set comprised 50% of the total data and was used for model learning. The validation set accounted for 25%, assisting in hyperparameter tuning and performance monitoring. The remaining 25% was reserved as the test set for unbiased evaluation of the final model

3.4 Proposed Custom CNN Architecture

The architecture of the proposed convolutional neural network is designed to effectively extract hierarchical features from knee X-ray images for osteoporosis detection. The model progressively learns low-level spatial features such as edges and contours, followed by high-level semantic representations such as bone texture and structural degradation patterns. The overall architecture is illustrated in Figure 3.1.

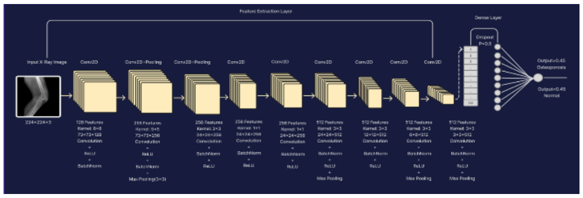
Feature Extraction Layer
Dense Layer
Dropout P=0.5
Input X-Ray Image
224×224×3
Conv2D
128 Features Kernel: 8×8 73×73×128 Convolution + ReLU + BatchNorm
Conv2D+Pooling
256 Features Kernel: 5×5 73×73×256 Convolution + ReLU + BatchNorm + Max Pooling(3×3)
Conv2D+Pooling
256 Features Kernel: 3×3 24×24×256 Convolution + BatchNorm + ReLU
Conv2D
256 Features Kernel: 1×1 24×24×256 Convolution + BatchNorm + ReLU
Conv2D
256 Features Kernel: 1×1 24×24×256 Convolution + BatchNorm + ReLU
Conv2D
512 Features Kernel: 3×3 24×24×512 Convolution + BatchNorm + ReLU + Max Pooling
Conv2D
512 Features Kernel: 3×3 12×12×512 Convolution + BatchNorm + ReLU
Conv2D
512 Features Kernel: 3×3 6×6×512 Convolution + BatchNorm + ReLU + Max Pooling
Conv2D
512 Features Kernel: 3×3 3×3×512 Convolution + BatchNorm + ReLU + Max Pooling
Output>0.45 Osteoporosis
Output<0.45 Normal

Figure 3.1: Proposed Custom CNN Architecture

3.4.1 Input Representation and Preprocessing

The proposed model operates on knee X-ray images, which are resized to a fixed resolution of (224 X 224) pixels to ensure uniformity across the dataset. Since the original radiographs are grayscale, they are converted into a three-channel format by channel replication, resulting in an input tensor of dimension (224 X 224 X 3). This transformation enables compatibility with convolutional operations while preserving the structural and intensity information of the original image. Furthermore, pixel values are normalized to the range ([0,1]) to improve numerical stability and accelerate convergence during training.

3.4.2 Convolutional Feature Extraction

The architecture is primarily composed of stacked convolutional layers that perform hierarchical feature extraction. A convolutional layer applies learnable filters across the input to detect spatial patterns such as edges, textures, and structural variations. Mathematically, the convolution operation can be expressed as:

$$Y(i,j) = \sum_{m} . \sum_{n} X(i+m, j+n) \cdot K(m,n) \quad \dots (1)$$

where (X) represents the input feature map, (K) is the kernel, and (Y) is the resulting feature map.

In the proposed model, the initial layers utilize larger kernel sizes (e.g., (8 X 8), (5 X 5)) to capture global anatomical structures such as bone contours and joint boundaries. As the network deepens, smaller kernels ((3 X 3), (1 X 1)) are employed to capture fine-grained features, including trabecular bone patterns that are critical for osteoporosis detection. The number of filters progressively increases from 128 to 512, allowing the network to learn increasingly complex and abstract representations.

3.4.3 Activation Function (ReLU)

Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function, defined as:

$$f(x)= \max(0, x) \quad \dots(2)$$

This non-linear activation enables the network to learn complex mappings between input images and output classes. Additionally, ReLU helps mitigate the vanishing gradient problem and improves computational efficiency by allowing sparse activation.

3.4.4 Batch Normalization

Batch normalization is applied after convolutional layers to stabilize the training process. It normalizes intermediate feature maps by adjusting their mean and variance within each mini-batch, thereby reducing internal covariate shift. This allows the model to converge faster, supports higher learning rates, and introduces a mild regularization effect that helps reduce overfitting.

3.4.5 Pooling Operations

Max-pooling layers are incorporated to reduce the spatial dimensions of feature maps while preserving the most significant features. This operation selects the maximum value within a local neighbourhood, effectively retaining dominant activations corresponding to important structures such as edges and high-density regions. Pooling also reduces computational complexity and introduces translation invariance, which is beneficial for medical image analysis.

3.4.6 Deep Feature Representation

As the network progresses to deeper layers, feature maps become more abstract and semantically meaningful. The gradual increase in depth up to 512 channels enables the model to capture high-level features such as bone texture irregularities, cortical thinning, and structural discontinuities. Simultaneously, spatial dimensions are reduced (from (73 X 73) to (3 X 3)), ensuring that only the most discriminative features are retained.

3.4.7 Global Average Pooling

Instead of using traditional flattening, the architecture employs a Global Average Pooling (GAP) layer, which computes the average value of each feature map to produce a compact feature vector. This significantly reduces the number of trainable parameters and minimizes overfitting. Additionally, GAP maintains a direct correspondence between feature maps and learned representations, improving interpretability in medical diagnosis tasks.

3.4.8 Fully Connected Layer

The extracted feature vector is passed through a dense layer consisting of 512 neurons. This layer performs high-level reasoning by learning complex combinations of features derived from earlier convolutional layers. It enhances the model's discriminative capability for distinguishing between normal and osteoporotic cases.

3.4.9 Dropout Regularization

To further improve generalization, dropout is applied with a rate of 0.5 after the dense layer. This technique randomly deactivates a subset of neurons during training, preventing the network from becoming overly dependent on specific activations and thereby reducing the risk of overfitting.

3.4.10 Output Layer and Classification

The final layer consists of a single neuron with a sigmoid activation function, which outputs a probability value between 0 and 1 representing the likelihood of osteoporosis. The model is trained using Binary Cross-Entropy loss, defined as:

$$\{L\} = -\left[y \backslash log\left(\widehat{\{y\}}\right) + (1 - y)\backslash log\left(1 - \widehat{\{y\}}\right)\right] \quad \ldots(3)$$

where y is the ground truth label and ŷ is the predicted probability. This loss function is well-suited for binary classification problems and ensures effective optimization of the model.

3.4.11 Weighted Ensemble Strategy

To enhance robustness and improve diagnostic performance, particularly recall, multiple instances of the proposed CNN model are trained with different initializations. The final prediction is obtained by combining individual model outputs using a weighted averaging strategy:

$$\widehat{\{y\}} \quad {}_{\{ensemble\}} = \sum_{\{i=1\}}^{N} w_{\hat{\imath}\{y\}_i} \quad \ldots.(4)$$

where w_i represents the weight assigned to the i^{th} model and $\hat{y}_i$ is its prediction.

This ensemble approach reduces model variance, improves generalization, and minimizes false negatives, which is critical in medical diagnosis scenarios such as osteoporosis screening.

3.5 Initial Model Performance

The proposed Custom CNN architecture is designed as a lightweight yet effective framework for medical image classification, with a particular focus on osteoporosis detection from knee X-ray images. The model was evaluated on the testing dataset using standard performance metrics, including accuracy, precision, recall, and F1-score, to ensure a comprehensive assessment of its diagnostic capability.

The model achieves an overall testing accuracy of 93.6%, demonstrating strong discriminative performance. However, in clinical applications, accuracy alone is not a sufficient indicator of model reliability. In particular, recall (sensitivity) plays a critical role, as false negatives correspond to missed disease cases, which may lead to delayed diagnosis and treatment.

To achieve an appropriate balance between precision and recall, a classification threshold of 0.65 is adopted. The detailed classification report is presented in Table 2.

Table 2: Classification Report of the Proposed CNN Model

Class	Precision	Recall	F1-Score	Support
Normal (0)	0.93	0.88	0.90	234
Osteoporotic (1)	0.89	0.94	0.92	253
Overall Accuracy	-	-	0.91	487
Macro Average	0.91	0.91	0.91	487
Weighted Average	0.91	0.91	0.91	487

The comparatively higher recall for the osteoporotic class indicates that the model is effective in identifying positive cases, which is a desirable property in medical screening systems.

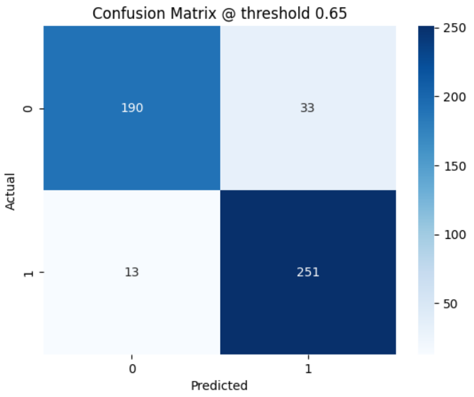

Figure 3.2: Confusion matrix of Initial Model for the testing dataset evaluated at a decision threshold of 0.65

As observed from Fig. 3.2, the model correctly classifies 225 normal samples (true negatives) and 222 osteoporotic samples (true positives). The number of false negatives is limited to 12, indicating that only a small fraction of osteoporotic cases is misclassified as normal. This is a crucial outcome, as minimizing false negatives is essential in medical diagnosis to avoid missed detections.

On the other hand, the model produces 28 false positives, where normal samples are incorrectly classified as osteoporotic. While this may lead to additional clinical verification, such errors are generally more acceptable compared to false negatives, as they do not directly compromise patient safety.

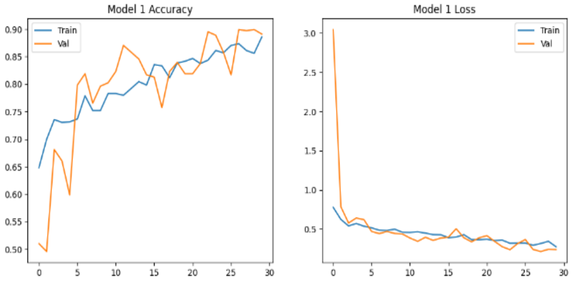

Figure 3.3: Training dynamics of the initial model showing accuracy and loss curves for both training and validation sets.

In Figure 3.3, the accuracy curves exhibit a consistent upward trend, while the loss curves show a steady decline, indicating effective learning behaviour. Furthermore, the close agreement between training and validation performance suggests that the model generalizes well to unseen data, with no significant signs of overfitting.

The experimental results demonstrate that the proposed Custom CNN provides a strong baseline for osteoporosis classification. The model effectively captures discriminative features from knee X-ray images, leading to high overall accuracy and robust recall for the osteoporotic class.

Despite these promising results, the presence of false positives and the need for further enhancement in sensitivity indicate potential areas for improvement. These limitations motivate the exploration of advanced strategies to further refine model performance, particularly in reducing misclassification errors while maintaining high recall.

3.6 Weighted Ensemble Learning (Core Contribution)

To further enhance the diagnostic reliability of the proposed system, a weighted ensemble learning framework is developed on top of the Custom CNN architecture. While a single CNN model achieves competitive performance, its predictions remain sensitive to stochastic factors such as weight initialization and training dynamics. To mitigate this limitation and improve robustness, multiple instances of the same architecture are trained and aggregated.

3.6.1 Multi-Model Ensemble Design

The proposed ensemble consists of five independent CNN models, each sharing an identical architecture but initialized with different random seed values. These models are denoted as (M_0, M_1, M_2, M_3, M_4), corresponding to seed values 100, 101, 102, 103, and 104, respectively.

Although the architecture and training data remain unchanged, variations in initialization and training order lead to diverse learned representations. This diversity is essential for effective ensemble learning, as it ensures that individual models capture complementary features related to bone texture, density variation, and structural abnormalities.

3.6.2 Model-Specific Threshold Optimization

Instead of using a uniform classification threshold, each model is assigned a custom threshold determined through empirical threshold tuning on validation data. The selected thresholds are:

$$T = \{0.65, 0.35, 0.60, 0.50, 0.45\}$$

These thresholds are chosen to balance precision and recall for each individual model. Since medical diagnosis prioritizes sensitivity, threshold selection is biased toward improving recall while maintaining acceptable precision.

3.6.3 Threshold-Based Confidence Normalization

To ensure consistency across models with different thresholds, each model's output probability P_i is transformed into a normalized confidence score C_i. This calibration aligns predictions onto a common scale before aggregation.

$$C_i = \left\{\begin{matrix} \{P_i - T_i\}\{1 - T_i\}, P_i > T_i \\ frac\{P_i\}\{T_i\}, otherwise \end{matrix}\right\} \quad \ldots(5)$$

This step is critical, as it preserves the relative confidence of each model while accounting for its individual decision boundary.

3.6.4 Weighted Ensemble Aggregation

The final prediction is obtained through a weighted combination of all five models, where each model contributes according to its assigned importance. Based on validation performance, the weights are defined as:
w = {1.0, 1.2, 1.0, 1.0, 1.4}

The ensemble prediction is computed as:

$$\widehat{\{y\}} \quad {}_{\{ensemble\}} = \sum_{i=0}^{4} w_i C_i \quad \ldots(6)$$

This formulation assigns higher influence to models demonstrating superior diagnostic capability, particularly in terms of recall and F1-score. The weighting mechanism ensures that stronger models contribute more significantly to the final decision, thereby improving overall performance.

3.6.5 Framework Overview

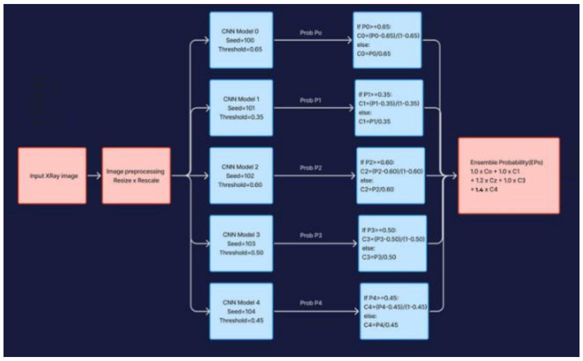

Figure 3.4: The complete workflow of the proposed ensemble system

As shown in Fig. 3.4, the input X-ray image undergoes preprocessing before being simultaneously passed through all five CNN models. Each model generates a probability score, which is then calibrated using its respective threshold. The normalized outputs are subsequently combined using predefined weights to produce the final ensemble prediction.

3.6.6 Performance Motivation

The proposed ensemble strategy is designed to address several critical challenges associated with medical image classification, particularly in the context of osteoporosis detection. One of the primary motivations is the reduction of model variance. By aggregating predictions from multiple independently trained convolutional neural networks, the ensemble stabilizes the overall prediction and reduces sensitivity to random initialization and training fluctuations.

Another key objective is the improvement of recall, which is of paramount importance in medical diagnosis. Recall is defined as:

$$Recall = TP/(TP + FN) \quad \ldots(7)$$

A lower number of false negatives ((FN)) directly leads to higher recall. By combining multiple models with diverse decision boundaries, the proposed ensemble significantly reduces false negatives, thereby improving the system's ability to correctly identify positive osteoporosis cases.

Furthermore, the ensemble enhances robustness to data variability. Medical imaging datasets often exhibit heterogeneity due to differences in imaging devices, patient demographics, and acquisition conditions. The integration of multiple models enables the system to capture diverse feature representations, making it more resilient to such variations.

In addition, the ensemble improves generalization performance. Let f_i(x) denote the prediction of the i^{th} model. The ensemble output can be expressed as:

$$f_{\{ensemble\}(x)} = \sum_{i=1}^{N} w_i f_{i(x)} \quad \ldots(8)$$

where w_i represents the weight assigned to each model. This aggregation of multiple hypotheses allows the system to generalize better on unseen data compared to individual models.

3.6.7 Significance of the Proposed Ensemble

Unlike conventional approaches that rely on a single deep learning model, the proposed method integrates model diversity, threshold optimization, and weighted aggregation into a unified framework. The final prediction is obtained using a decision threshold:

$$\widehat{\{y\}} = \begin{cases} 1, f_{(ensemble)(x)} \geq \tau \\ 0, \text{otherwise} \end{cases} \quad \ldots.(9)$$

This formulation allows flexible control over the trade-off between precision and recall, which is crucial in medical applications.

The proposed ensemble demonstrates superior performance compared to individual models, achieving an accuracy of 0.94 along with balanced precision and recall values of 0.94. More importantly, the confusion matrix analysis reveals a significant reduction in false negatives (FN = 2), which directly enhances diagnostic reliability.

This improvement aligns with the primary objective of medical diagnosis, where maximizing detection capability is more critical than minimizing false alarms. Therefore, the proposed ensemble framework not only improves predictive performance but also provides a more reliable and clinically applicable solution for osteoporosis screening.

IV. EXPERIMENTAL RESULTS AND EVALUATION

4.1 Evaluation Metrics

The performance of the proposed model is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics are computed using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and are defined as follows:

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

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

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

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Accuracy reflects the overall correctness of the model, while precision and recall provide insight into classification performance on positive samples. In medical diagnosis, recall is particularly important, as it minimizes false negatives, which correspond to missed disease cases.

4.2 Individual Model Performance

Five convolutional neural network (CNN)-based models were trained and evaluated independently to serve as weak learners. The training process demonstrated stable convergence, as evidenced by consistent training and validation accuracy and loss trends.

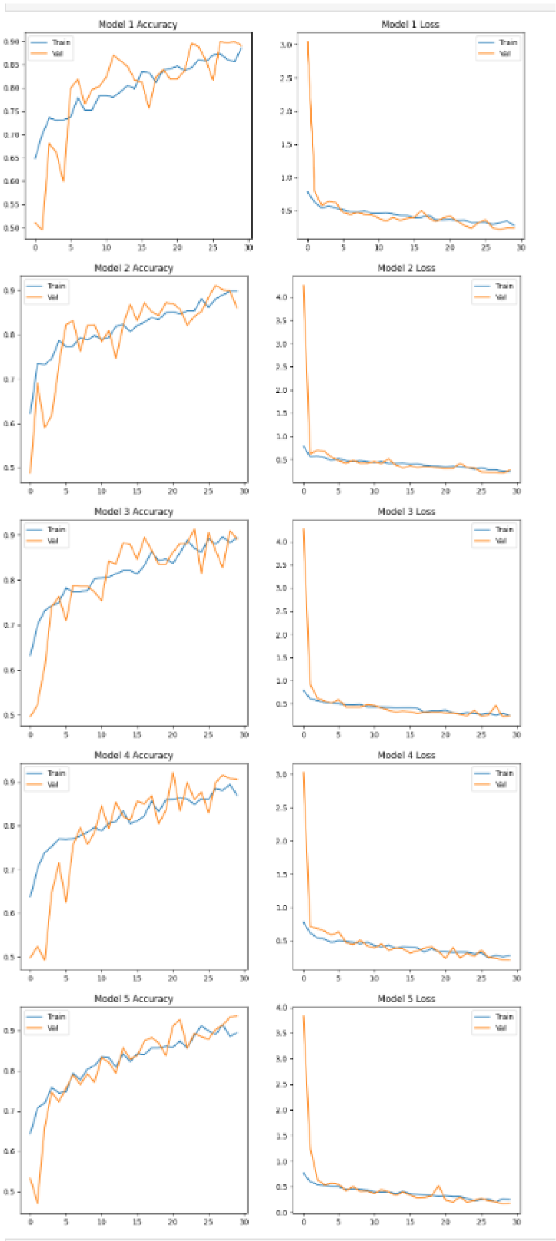

Fig 4.1 Accuracy and Loss curves of 5 models

The models achieved accuracy values ranging from 0.90 to 0.93. Although these results indicate strong baseline performance, certain inconsistencies were observed in balancing precision and recall. This limitation suggests that relying on a single model may not be sufficient for achieving optimal diagnostic performance.

4.3 Confusion Matrix Analysis of Individual Model

Although the model correctly classifies a large number of samples, the presence of false negatives indicates that some positive cases are incorrectly predicted as negative. Mathematically, an increase in FN directly reduces recall:

$$Recall = \frac{TP}{TP+FN} \quad \ldots(10)$$

This limitation is critical in medical applications, motivating the use of ensemble learning to improve detection sensitivity.

Table 3: Best Threshold and Accuracy of Individual Models

Model	Best Threshold	Accuracy
Model 1	0.65	91%
Model 2	0.35	90%
Model 3	0.60	92%
Model 4	0.50	90%
Model 5	0.50	94%

The images from Figure 4.2 to Figure 4.6 represent the confusion matrices of Model 1 to Model 5, respectively, evaluated at their optimal threshold values. Each confusion matrix illustrates the distribution of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), providing insight into the classification performance of the models in detecting osteoporosis from knee X-ray images. These matrices help analyze model accuracy, error patterns, and overall predictive reliability.

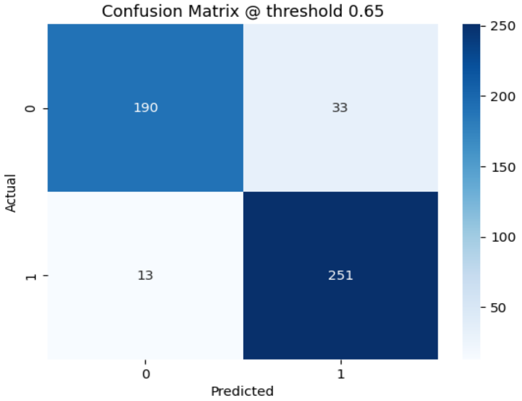

Figure 4.2: Confusion Matrix for Model 1 at Threshold 0.65

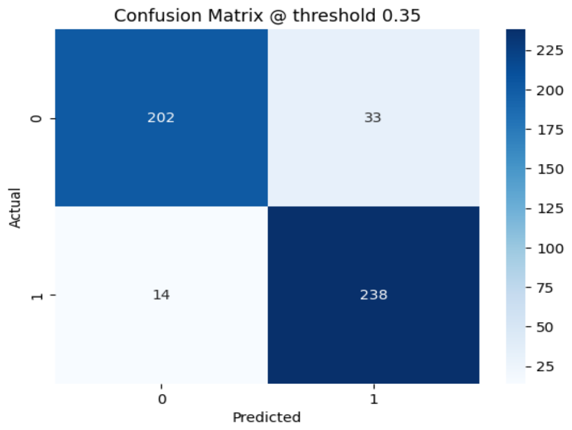

Figure 4.3: Confusion Matrix for Model 2 at Threshold 0.35

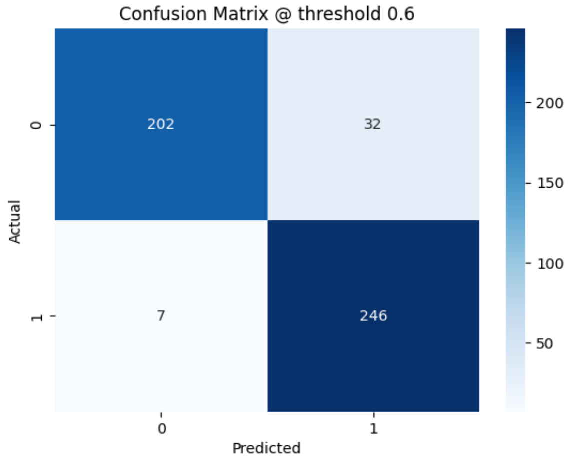

Figure 4.4: Confusion Matrix for Model 3 at Threshold 0.60

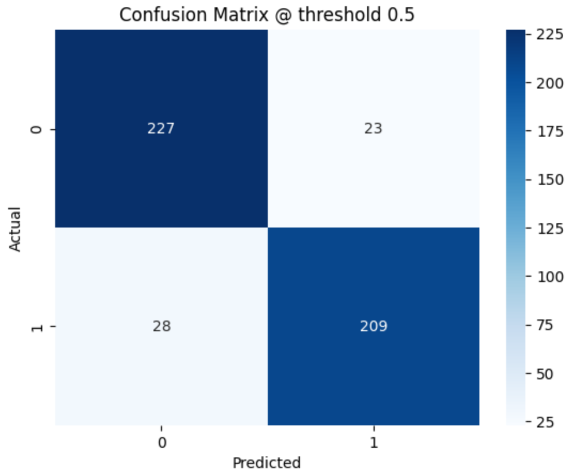

Figure 4.5: Confusion Matrix for Model 4 at Threshold 0.50

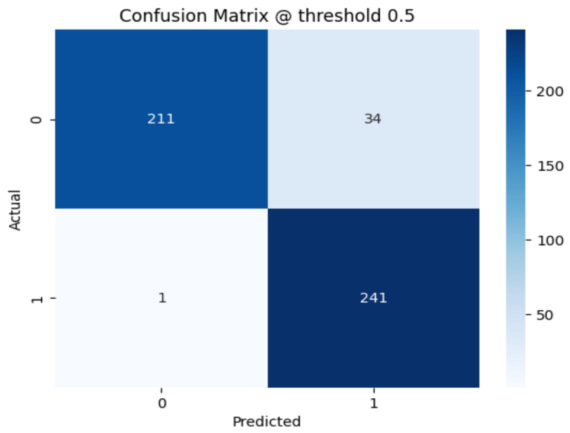

Figure 4.6: Confusion Matrix for Model 5 at Threshold 0.50

4.4 Ensemble Model Evaluation

To overcome the limitations of individual models, a weighted ensemble learning approach is proposed. The ensemble combines predictions from multiple CNN models using a weighted aggregation mechanism.

Let $p_i(x)$denote the predicted probability from the ith CNN model and $w_i$represent its corresponding weight. The final ensemble prediction score $S(x)$is computed as:

$$S(x) = \sum_{i=1}^{N} w_i p_{i(x)} \quad \ldots(11)$$

where $N$is the total number of models in the ensemble.

The final class label $\hat{y}$is determined using a decision threshold τ:

$$\widehat{\{y\}} = \begin{cases} 1, S(x) \geq \tau \\ 0, \text{otherwise} \end{cases} \quad \ldots(12)$$

The performance comparison between individual CNN models and the proposed ensemble model is presented in Table 2. The ensemble achieves an accuracy of 0.94, with precision, recall, and F1-score all equal to 0.94, outperforming all individual models.

The improvement can be attributed to the aggregation of diverse feature representations learned by different models, which reduces variance and enhances generalization.

4.5 Confusion Matrix Analysis of Ensemble Model

The confusion matrix of the proposed ensemble model is shown in Fig. 5.3 and is given by:

$$\begin{bmatrix}214 & 28\\ 2 & 243\end{bmatrix}$$

$$Accuracy = \frac{214 + 243}{487} = 0.94$$

$$Precision = \frac{243}{243 + 28\backslash approx 0.90}$$

$$Recall = \frac{243}{243 + 2\backslash approx 0.99}$$

$$F1\ score = approx 0.94$$

The results indicate that the model achieves a very low number of false negatives (FN = 2), which significantly improves recall. This is particularly desirable in medical diagnosis, where detecting positive cases is critical.

Although the number of false positives is relatively higher, this trade-off is acceptable, as it prioritizes minimizing missed detections over avoiding false alarms.

4.6 Threshold Optimization Analysis

The effect of varying the decision threshold was analyzed over the range $0.3 \leq \tau \leq 0.7$. The threshold directly influences the classification decision:

- Lower values of τincrease recall but may introduce more false positives.
- Higher values of τreduce false positives but may increase false negatives.

The optimal threshold was found to be:

$$\tau = 0.5$$

At this value, the model achieves the best balance between precision and recall, both reaching 0.94. This demonstrates that threshold tuning plays a crucial role in optimizing ensemble performance.

4.7 Discussion

The experimental results demonstrate that the proposed ensemble model consistently outperforms individual CNN models. The most significant improvement is observed in recall, which indicates a substantial reduction in false negatives.

The weighted aggregation strategy enables the model to combine complementary strengths of individual learners, resulting in improved robustness and reliability. Furthermore, threshold optimization enhances the balance between precision and recall, leading to superior overall performance.

These results confirm that the proposed approach is well-suited for medical image classification tasks, where accuracy and reliability are of paramount importance.

V. CONCLUSION

This study proposes a weighted ensemble framework based on custom Convolutional Neural Networks (CNNs) for automated osteoporosis detection using knee X-ray images. Unlike conventional DEXA-based diagnosis, the proposed approach leverages widely available radiographic imaging to enable a cost-effective and scalable screening solution, particularly suitable for resource-limited settings.

While the standalone CNN model demonstrated strong baseline performance, the integration of multiple independently trained models through weighted ensemble learning significantly improved recall and reduced false negative predictions. This enhancement is critical in medical diagnosis, where missed detections can lead to delayed treatment and increased fracture risk.

Experimental results indicate that the proposed ensemble framework achieves improved diagnostic reliability while maintaining high classification accuracy. Therefore, the method presents a practical and accessible solution for early osteoporosis screening using existing X-ray infrastructure. Future work will focus on clinical validation and model interpretability for real-world deployment.

ACKNOWLEDGMENT

The authors thank the Department of Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Andheri (West), Mumbai, for providing the necessary facilities and academic support to carry out this research. The authors also acknowledge the availability of publicly accessible datasets and resources that contributed to the development and evaluation of the proposed system.

REFERENCES

[1] R. M. J. Rasool, S. Ahmad, U. Sabina, and T. K. Whangbo, "KONet: Toward a weighted ensemble learning model for knee osteoporosis classification," *IEEE Access*, 2024.

[2] D. T. Rao, K. S. Ramesh, V. S. Ghali, and M. V. Rao, “The osteoporosis disease diagnosis and classification using U-Net deep learning process,” *IEEE Access*, 2022.

[3] Z. Si, D. Zhang, H. Wang, and X. Zheng, “PrOsteoporosis: Predicting osteoporosis risk using NHANES data and machine learning approach,” *BMC Research Notes*, 2025.

[4] Z. Liu, Y. Li, C. Zhang, H. Xu, J. Zhao, C. Huang, X. Chen, and Q. Ren, “Radiomics and machine learning for osteoporosis detection using abdominal computed tomography: A retrospective multicenter study,” *BMC Medical Imaging*, 2025.

[5] A. M. Sarhan, M. Gobara, S. Yasser, Z. Elsayed, G. Sherif, N. Moataz, Y. Yasir, E. Moustafa, S. Ibrahim, and H. A. Ali, “Knee osteoporosis diagnosis based on deep learning,” *Springer Journal*, 2024.

[6] Z. Chen, W. Luo, Q. Zhang, B. Lei, T. Wang, and J. Liu, “Osteoporosis diagnosis based on ultrasound radio frequency signal via multi-channel convolutional neural network,” *IEEE Conference Proceedings*, 2021.

[7] K. Kim, A. Yoo, E. Oh, and D. Kim, “Osteoporosis risk prediction using machine learning and conventional methods,” *IEEE Conference Proceedings*, 2013.

[8] 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*, 2025.

[9] D. Hwang, S. Bak, T. Ha, Y. Kim, and H. Choi, “Multi-view computed tomography network for osteoporosis classification,” *IEEE Access*, 2023.

[10] J. W. Adams, Z. Zhang, G. Noetscher, A. Nazari, and S. Makarov, “Application of a neural network classifier to radiofrequency-based osteopenia/osteoporosis screening,” *IEEE Journal of Translational Engineering in Health and Medicine*, 2021.