International Journal of Technology

http://ijtech.eng.ui.ac.id

Research Article

Towards Reliable Osteoarthritis Classification: Fine-Tuned Convolutional Neural Networks, Vision Transformers, and Ensemble Learning Approaches

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, 10071, Baghdad, Iraq

[2]Biomedical Engineering Program, Faculty of Engineering, Mansoura University, 35516, Mansoura, Egypt

[3]Electronics and Communications Engineering Department, Faculty of Engineering, Mansoura University, 35516, Mansoura, Egypt

[4]Orthopedic Surgery Department, Faculty of Medicine, Mansoura University, 35516, Mansoura, Egypt

[5]School of Computational Sciences and Artificial Intelligence (CSAI), Zewail City of Science and Technology, 12578, Giza, Egypt

*Corresponding author: israasafa@kecbu.uobaghdad.edu.iq

Abstract: Osteoarthritis (OA) is a widespread degenerative condition affecting millions of people worldwide. Early detection and precise classification are crucial for effective disease management. This study investigated the use of deep learning techniques to classify the severity of knee OA from X-ray images, specifically targeting three categories: Normal (KL Grade 0), Moderate (KL Grade 3), and Severe (KL Grade 4). We utilized a dataset from the Osteoarthritis Initiative (OAI), containing 3,221 X-ray images of the knee, and fine-tuned eight pretrained CNNs (DenseNet201, EfficientNetB7, InceptionV3, InceptionResNetV2, ResNet50V2, ResNet152V2, Vision Transformer B32, and Xception). A custom CNN and ensemble deep learning models (hard and weighted voting) were also proposed with a total of 11 models. The models were assessed using a dataset split of 70% for training, 15% for validation, and 15% for testing, ensuring comprehensive evaluation across all development stages. DenseNet201 achieved the highest classification accuracy of 97.11% among the individual models, while Vision Transformer B32 showed the lowest accuracy of 59.38%. Ensemble methods using hard and weighted voting, incorporating the top five models, achieved a consistent accuracy of 97.11%. These results demonstrate the potential of deep learning, particularly ensemble strategies, in accurately classifying knee OA severity. This method can help build smarter tools that assist doctors in making better decisions, aiding in the early detection and management of OA, offering a robust tool for improving patient outcomes.

Keywords: Classification; Deep Learning; Ensemble Methods; Knee X-ray; Osteoarthritis

1. Introduction

Osteoarthritis (OA) (Keith Sinusas, 2012) is the most prevalent form of arthritis, affecting millions of people worldwide and placing a growing strain on healthcare systems, especially as the number of aging populations and obesity rates rise (Xie et al., 2025; Fajarani et al., 2024). Osteoarthritis is a progressive joint disorder characterized by cartilage breakdown, joint inflammation, and structural deterioration. These pathological changes result in chronic pain, impaired mobility, and significantly reduced quality of life. Although the Kellgren–Lawrence (KL) grading scale is the most commonly employed radiographic tool for assessing OA severity (Kondal et al., 2022), it remains subjective and is prone to inconsistencies between different observers, undermining its diagnostic reliability, particularly in early stages (Zhao et al., 2025).

https://doi.org/10.14716/ijtech.v17i1.8007
Received August 2025; Revised October 2025; Accepted December 2025

Recent advances in artificial intelligence (AI)—particularly in deep learning (DL) and convolutional neural networks (CNNs), offer significant promise for the objective and automated analysis of OA from medical images. Unlike traditional methods, these techniques can learn relevant features from raw radiographs autonomously, thereby reducing the need for manual interpretation and improving consistency (Foti and Longo, 2024). However, existing AI methods face challenges in ensuring generalizability across different imaging settings and in dealing with model interpretability, especially for early-stage OA diagnosis (Ou et al., 2025; Xin Teoh et al., 2024).

To address these challenges, this study presents approaches that fine-tunes eleven state-of-the-art DL models, including CNNs, ensemble hard and weighted voting models, and a vision transformer, on a carefully partitioned dataset of knee X-ray images. A unique aspect of this work is the introduction of ensemble strategies, including hard and weighted voting, to improve robustness and combine the strengths of individual models. Our results demonstrate that by combining multiple high-performing models, these ensemble strategies significantly enhance diagnostic consistency without sacrificing accuracy, a crucial advantage over single-model approaches.

The proposed method outperforms previous studies, achieving exceptional accuracy and robustness, thus offering a reliable tool for automated OA classification in clinical decision-support systems. The main contributions of this work include the application of cutting-edge ensemble techniques, robust fine-tuning of multiple model architectures, and demonstration of their potential for OA classification in clinical settings. Despite significant advances in deep learning for medical image analysis, current approaches to knee osteoarthritis (KOA) classification suffer from limited generalizability, inconsistent performance across severity levels, and reliance on single-model architectures that struggle to balance accuracy and robustness. This study addresses these limitations by developing and evaluating fine-tuned convolutional and transformer-based models, enhanced through ensemble strategies, to achieve reliable and objective multi-class KOA classification using the OAI dataset.

The novelty of this study lies in the integration of ensemble learning with fine-tuned convolutional and transformer-based architectures to achieve robust, multi-class KOA classification. Unlike previous studies that relied on single-model or binary setups, this study introduces a comprehensive comparison of eleven state-of-the-art models, including both CNNs and Vision Transformers—and proposes optimized ensemble strategies (hard and weighted voting) to enhance generalization and diagnostic reliability. The study also establishes a new benchmark on the OAI dataset, demonstrating superior accuracy and balanced performance across all evaluation metrics, thereby contributing to the development of a clinically relevant and technically innovative framework for automated OA diagnosis. The primary research question guiding this study is: Can fine-tuned CNNs and vision transformers, when combined through ensemble learning, improve the accuracy and robustness of multi-class KOA classification compared to individual models?

Osteoarthritis (OA) is a widespread health issue that strains healthcare systems globally, with traditional diagnostic methods, such as the Kellgren–Lawrence (KL) grading scale, suffering from subjectivity and inter-observer variability (Wing et al., 2021). Despite advancements in deep learning (DL) for OA diagnosis, the literature has primarily focused on single-model approaches, with limited attention given to the potential benefits of combining multiple models using ensemble strategies. This gap is significant because ensemble methods, such as hard and weighted voting, have shown promise in improving the robustness and accuracy of machine learning models by leveraging the strengths of various algorithms. This study proposes an ensemble-based approach for OA classification, combining multiple state-of-the-art DL models to enhance classification accuracy and provide a more reliable tool for clinical decision-support systems. Early detection of OA plays a crucial role in improving patient outcomes by enabling timely intervention, which can slow disease progression and reduce the need for more invasive treatments, such as joint replacement surgery. Identifying OA in its early stages allows for better management through nonsurgical methods such as physical therapy, medication, and lifestyle changes, which can significantly enhance a patient's quality of life and reduce healthcare costs. Multiple studies have explored the application of DL techniques for the automated classification and grading of knee OA using radiographic images, primarily employing the Kellgren and Lawrence (KL) grading system as a reference standard. One such effort is MedKnee (Touahema et al., 2024), which utilizes the pre-trained Xception model within a graphical user interface (GUI) to assist physicians in diagnosing KOA. MedKnee trained on 5000 X-ray images from the Osteoarthritis Initiative dataset achieved accuracies of 95.36% and 94.94% on two external validation datasets, indicating robust performance in automated KOA prediction. Although MedKnee demonstrated robust performance with high accuracy, it relies on a single pretrained model (Xception), which could benefit from exploring a broader range of models and advanced TL techniques for more generalizable results across diverse datasets. A 12-layer CNN was developed specifically for binary classification and severity grading of knee osteoarthritis (KOA) (Rani et al., 2024). Leveraging the OAI dataset, the proposed model achieved an accuracy of 92.3% for binary classification and 78.4% for multi-class severity classification, outperforming prior methods.

The 12-layer CNN model showed promising results for binary classification but had lower accuracy for multi-class classification. A deeper exploration of multi-class classification models or ensemble techniques could help improve the robustness of the model for varying KOA severity levels. (Zhang et al., 2024) introduced a Dense Multi-Scale (DMS) CNN module, improving feature recognition through dense connections across varying convolutional layers. Their model demonstrated superior performance compared to a DenseNet baseline, with 73.00% accuracy and 92.73% area under the curve (AUC), reinforcing the benefit of multi-scale feature learning. While the DMS CNN module improved feature recognition, its performance on larger datasets with more complex images (e.g., diverse real-world data) could be further investigated to validate its generalizability across various clinical settings. (Kaur et al., 2024) explored ConvNeXt, an advanced architecture based on ResNet and Transformer models.

Their methodology emphasized the pre-processing and augmentation of KL-graded X-ray images from the OAI dataset. The results indicated that ConvNeXt outperformed conventional models and vision transformers, which was statistically verified through robust evaluation metrics. The study highlighted ConvNeXt's superior performance, but the integration of clinical metadata and longitudinal patient data alongside the image data could further enhance model interpretability and generalizability. Yong et al., 2022 validated an ORM combined with multiple state-of-the-art architectures such as DenseNet, ResNet, and MobileNet v2. Among these, DenseNet-161 exhibited the best results, achieving a mean absolute error of 0.330, an ACC-Macro of 88.09%, and a quadratic weighted kappa (QWK) of 0.8609. The Ordinal Regression Module (ORM) achieved good results with DenseNet-161, but further research could focus on improving model accuracy for early KOA detection and investigating its effectiveness in detecting subtler forms of the disease across different age groups and demographics. Ahmed and Imran, 2024 investigated the application of fine-tuned CNN models, including VGG, ResNet, and EfficientNetb7, for both multiclass and binary classification tasks. Although EfficientNetb7 consistently performed best, their GradCAM analysis revealed limitations in distinguishing certain KL grades, highlighting the complexity of expert diagnosis replicating. While EfficientNetb7 showed high accuracy, their study's GradCAM analysis revealed challenges in distinguishing specific KL grades. Addressing these challenges with advanced feature extraction techniques or incorporating additional contextual information could improve model performance. (Yoon et al., 2023) developed MediAI OA, a comprehensive artificial intelligence (AI) model incorporating joint space narrowing (JSN) quantification, osteophyte detection, and KL grading. Trained on 44,193 OAI radiographs, the model achieved substantial consistency with clinical ground truth, as reflected by a kappa coefficient of 0.768 and an accuracy of 92% in OA diagnosis.

MediAI OA achieved good consistency with clinical ground truth, but its practical application could be further strengthened by integrating more diverse datasets with real-world data variability (e.g., different imaging conditions). (Sarvamangala and Kulkarni, 2021) presented the MCBCNN, which integrates pretrained CNNs such as MobileNet2, ResNet50, and InceptionNetV3 with MCFs. The MCBCNN based on ResNet50 reported the highest performance, with over 95% average accuracy and nearly 0.9 AUC. MCBCNN achieved high accuracy, but the performance could be further improved by incorporating dynamic learning strategies, real-time updates from clinical data, or integration with other diagnostic modalities like MRI or CT scans. Ganesh Kumar and Das Goswami (Ganesh Kumar and Goswami, 2023) emphasized the importance of pre-processing, particularly image sharpening, to enhance the clarity of knee X-rays. Their enhanced CNN-based approach achieved a significant improvement, reporting a mean accuracy of 91.03%, up from 72% using unprocessed images. The enhancement achieved through image sharpening is notable, but a more comprehensive approach using additional pre-processing techniques (e.g., contrast adjustment or edge detection) combined with DL could further boost accuracy and robustness. (Olsson et al., 2021) addressed the challenge of classifying KOA severity in unfiltered datasets that consider prevalent visual anomalies, including implants and casts.

Using ResNet trained on 6103 exams, the authors reported an AUC of 0.92 for KL grades, demonstrating robust performance even with real-world data variability. Although their model performed well even with real-world data containing disturbances, incorporating techniques that automatically filter out noisy or irrelevant images could improve the challenges posed by mixed-quality data. (Tariq et al., 2025) proposed a fine-tuned DenseNet169 model evaluated against several DL algorithms. They achieved 95.93% and 93.78% accuracy in multi-class and binary classification, respectively, using the OAI dataset, demonstrating the effectiveness of advanced model architectures in KOA severity assessment. Finally, Alshamrani introduced Osteo-NeT, leveraging transfer learning with VGG-16 and ResNet-50. Their method focused on early detection, with VGG-16 achieving 99% training accuracy and 92% testing accuracy, indicating strong predictive capabilities suitable for clinical applications. Osteo-NeT showed strong performance in early detection, but further improvements could be made by leveraging more advanced TL models or integrating multi-modal data, such as clinical reports, to enhance diagnostic capabilities. Together, these studies underscore the growing reliance on public datasets like OAI (Yong et al., 2022) for model development and evaluation. They also highlight key methodological advances that contribute to improved KOA classification accuracy, such as dense multi-scale convolution (Zhang et al., 2024), ConvNeXt integration (Kaur et al., 2024), and image pre-processing (Ganesh Kumar and Goswami, 2023).

The accurate and early classification of KOA severity is crucial, and deep learning, particularly with CNNs, has become the primary method for automating this diagnosis based on the Kellgren-Lawrence (KL) grading system. Recent research has leveraged extensive image datasets, such as the OAI, to drive methodological improvements. The following section reviews key advancements, highlighting the effective use of sophisticated architectures, multiscale feature learning, ordinal regression modules, and enhanced image preprocessing to improve the accuracy and generalizability of KOA severity assessment models.

Osteoarthritis (OA) is a widespread health issue that strains healthcare systems globally, with traditional diagnostic methods, such as the Kellgren–Lawrence (KL) grading scale, suffering from subjectivity and inter-observer variability (Wing et al., 2021). This study aims to overcome these challenges by using deep learning (DL) models, including convolutional neural networks (CNNs) and vision transformers, to automate and improve OA classification using knee X-ray images. The use of ensemble strategies—hard and weighted voting—to combine multiple models, improving robustness and accuracy over single-model approaches, is a key innovation. By fine-tuning 11 state-of-the-art DL models, this study not only enhances classification performance but also offers a more reliable tool for clinical decision-support systems. The main objectives of this study are to fine-tune and evaluate multiple DL models, explore ensemble techniques, and provide a robust and, accurate framework for OA classification, ultimately bridging the gaps in early OA diagnosis and management.

2. Methods

Figure 1 presents an OA classification framework leveraging DL models. The process includes several steps, including data pre-processing, model training, evaluation, and comparison. The overall framework of this study follows a structured sequence to ensure methodological coherence and reproducibility. The process begins with dataset acquisition and preprocessing of knee X-ray images from the OAI database to enhance image quality and standardize inputs. Next, multiple DL architectures comprising fine-tuned CNNs, a vision transformer, and ensemble models—are developed and trained for multi-class KOA classification. The model performance is then evaluated using comprehensive metrics, followed by statistical analysis and comparison with existing studies to validate the proposed approach's robustness and clinical relevance. This framework ensures a systematic link between data preparation, model optimization, and diagnostic application.

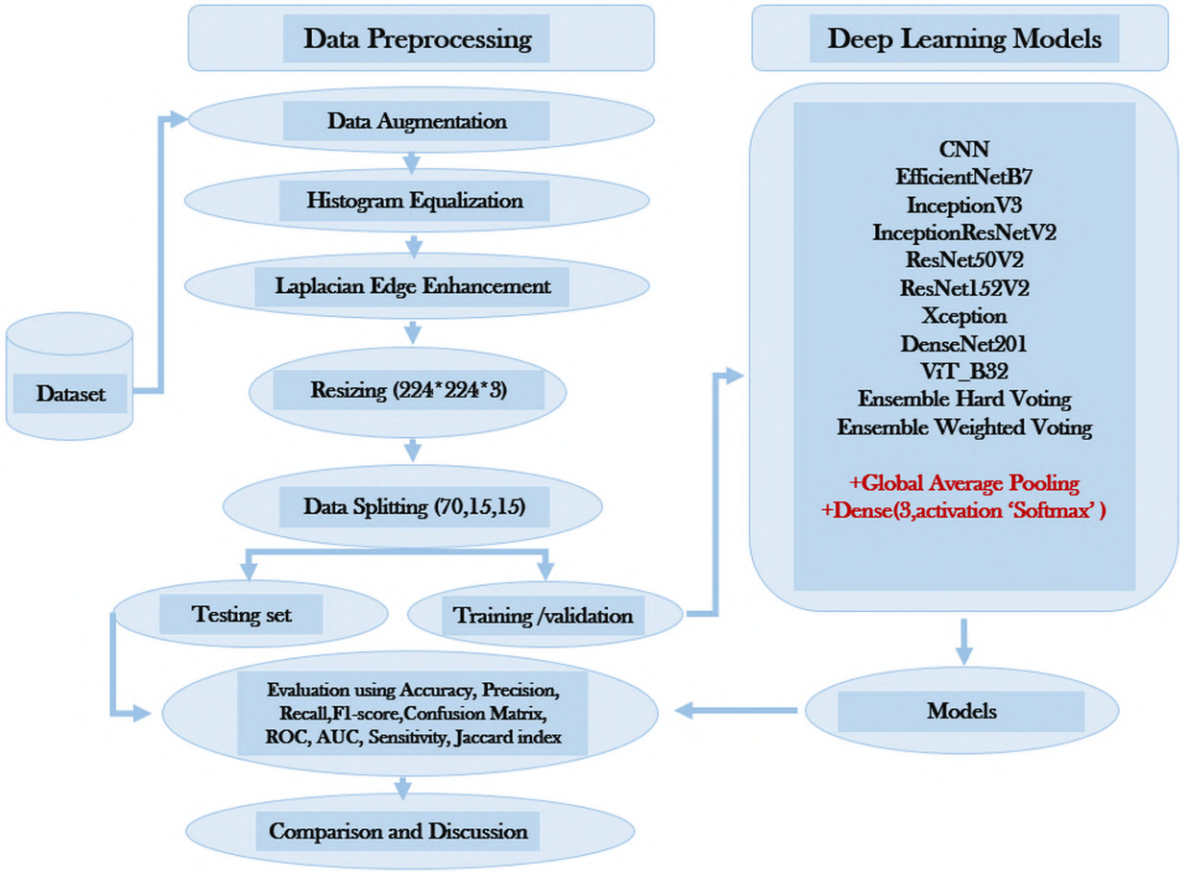

Figure 1 Methodological block diagram

2.1 Description of the Dataset

In this study, we utilized a knee X-ray dataset specifically curated for knee joint detection and Kellgren–Lawrence (KL) grading (Chen, 2018). The dataset is organized and made publicly available through Mendeley Data, sourced from the OAI database. The Osteoarthritis Initiative (OAI) is a crucial, publicly available database generated from a multicenter, prospective, longitudinal observational cohort study. Recruitment for the study began in February 2004 and ended in May 2006, enrolling 4,796 men and women aged 45–79 years. The primary aim of this study is to establish a resource—including clinical imaging (X-ray) and biospecimen data—to accelerate the identification of biomarkers for the incidence and progression of knee OA. A consistent schedule of in-clinic and phone follow-up visits defines the data series, originally planned for 8- years and later extended. The original OAI dataset includes various knee X-ray images annotated with KL grades ranging from 0 (normal) to 4 (severe OA).

To our OA classification task using deep learning, we selected three specific KL grades: Normal (KL Grade 0), Moderate (KL Grade 3), and Severe (KL Grade 4). These classes were selected because they represent the most clinically relevant stages for treatment decisions. Intermediate grades (1 and 2) were excluded due to high inter-observer variability and label ambiguity, which could reduce the reliability of the model. Focusing on these distinct categories improves robustness and ensures practical clinical applicability. A total of 3,221 knee X-ray images were used in this study, which were distributed across the three classes as shown below:

Table 1 Description of the dataset

Class	KL grade	Number of images	Description
Normal	0	1,640	No signs of OA
Moderate	3	1,286	Moderate osteoarthritis
Severe	4	295	Severe osteoarthritis
Total		3,221	

These images undergo preprocessing steps such as augmentation, histogram equalization, edge enhancement, resizing, and normalization before being fed into deep learning models for classification.

2.2 Data Preprocessing

The effective preprocessing of medical imaging data plays a crucial role in improving the performance and generalization of deep learning models. In this study, the knee X-ray images from the selected dataset were subjected to a series of pre-processing steps designed to enhance image quality, normalize input, and augment the training data. Table 2 summarizes the preprocessing steps used in the proposed work.

Table 2 Preprocessing steps

Step No.	Process	Description
1	Grayscale conversion	Converts RGB images to single-channel grayscale to simplify feature extraction.
2	Histogram equalization	Improves contrast using Contrast Limited Adaptive Histogram Equalization (CLAHE).
3	Edge enhancement	Enhances edges using the Laplacian operator to highlight joint structures.
4	Resizing	Standardizes the image dimensions to 224 × 224 × 3 pixels.
5	Normalization	Scales the pixel values to the range [0, 1].
6	Channel expansion	Expands the grayscale image back to the RGB format for compatibility with pre-trained networks.
7	Data splitting	The dataset was split into three portions: 70% for training, 15% for validation, and 15% for testing.
8	Data augmentation	Applies random transformations to increase the diversity of the training set.

The preprocessing steps are implemented in Python using the OpenCV library. The function preprocess_image ensures that each image is prepared uniformly before being fed into the deep learning model.

- CLAHE enhances the local contrast using Eq. (1).

$$T(I) = \frac{(I - I_{\min}) \times (L - 1)}{I_{\max} - I_{\min}} \tag{1}$$

(Sharma and Kamra, 2023)

$T(I)$: transformed pixel intensity
I : original pixel intensity
$I_{\min}, I_{\max}$: minimum and maximum pixel values in the local region
L : number of gray levels (commonly 256)

- Edge enhancement using Laplacian: The Laplacian operator detects edges by applying Eq. (2) as follows:

$$\nabla^2 f(x,y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{2}$$

(Mlsna and Rodriguez, 2009)

where $f(x,y)$ is the intensity at the pixel (x,y).

- The normalization formula is used as in Eq. (3):

$$\text{Image}_{\text{normalized}} = \frac{\text{Image}_{\text{original}}}{255} \tag{3}$$

(Patro and Sahu, 2015)

The normalization formula in Eq. (3) uses the value of 255 because the input image pixel values are in the range of 0 to 255, which is the standard for images with 8-bit depth per channel. By dividing by 255, we scale the pixel values to a range of [0, 1], which is commonly done to normalize the data and improve the model's stability and performance during training. As shown in Table 3, the dataset was split into 70% training, 15% validation, and 15% testing sets, a commonly adopted strategy in medical imaging. This split ensures sufficient data for learning while reserving adequate samples for (HT) and unbiased evaluation.

Table 3 Information on data splitting

Subset	Number of images	Percentage (%)	Description
Training	2,254	70%	Model learning
Validation	482	15%	Hyperparameter tuning
Testing	485	15%	Final evaluation
Total	3,221	100%	

Table 3 presents the data splitting information for the training, validation, and testing subsets. The dataset is divided into 70% for training, 15% for validation, and 15% for testing, totaling 3,221 images. This balanced partitioning ensures that each subset is large enough to represent the full diversity of the dataset, which helps optimize model performance, reduces the risk of overfitting, and provides a fair evaluation of the model's generalization capabilities. In addition, balanced data splitting helps prevent the model from being biased toward any particular subset, ensuring that all classes are adequately represented in both the training and evaluation stages. Figure 2 shows a sample image before and after applying the pre-processing steps, highlighting the contrast enhancement and edge clarity, which further aids the model in making accurate predictions.

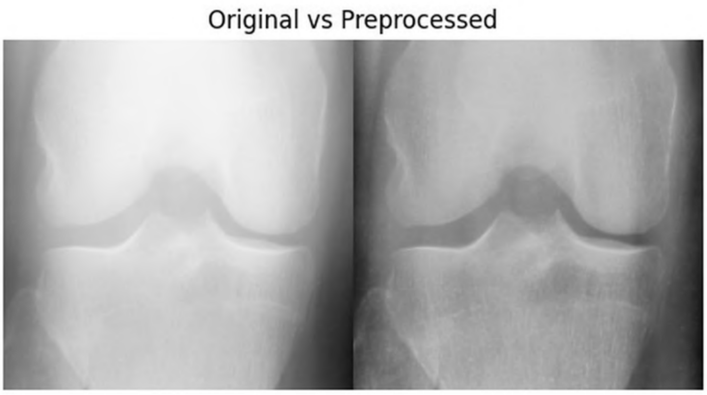

Figure 2 Comparison between the original and pre-processed knee X-ray images

In this study, augmentation techniques are applied to enhance the variety and strength of the training dataset, thereby reducing overfitting and improving model generalization. The augmentation process includes random rotation within a range of ±10 degrees, width and height shifts up to 10% of the image size, zoom variations of up to 10%, and horizontal flipping to account for variations in knee orientation. Additionally, the fill mode is set to "nearest" to handle pixel values introduced by these transformations. These augmentations are applied dynamically during training using TensorFlow's ImageDataGenerator, ensuring that the model encounters a varied set of image conditions while maintaining the anatomical structure relevant for OA classification. To extract rich and discriminative feature representations from input data, we used five advanced CNN (Chauhan et al., 2018) architectures as embedding models: EfficientNetB7 (Tan and Le, 2019), DenseNet201 (Zhu and Newsam, 2018), ResNet152V2 (Duklan et al., 2024), InceptionV3, and InceptionResNetV2 (Borawar and Kaur, 2023). These architectures were selected based on their high performance on large-scale image classification benchmarks and architectural diversity, which enhances the effectiveness of ensemble learning.

2.3 Deep Learning Models

EfficientNetB7 (Tan and Le, 2019), part of the EfficientNet family, uses a compound scaling strategy that evenly adjusts the depth, width, and resolution of the model. It delivers top-tier accuracy while using fewer parameters and less computation, making it ideal for TTL. DenseNet201 (Densenets features dense connections, where each layer is connected to all previous layers) (DenseNets et al., 2021). This setup improves the flow of gradients, promotes the reuse of features, and reduces parameter count—making it well-suited for datasets with limited training examples. ResNet152V2 (Koonce, 2021) is a deep residual model that solves the vanishing gradient issue by using identity shortcut connections, helping the network train more effectively. The V2 variant enhances training stability by using pre-activation residual units, allowing the model to learn more complex patterns. InceptionV3 is known for its efficient use of parameters and computational resources, achieved through factorized convolutions and aggressive regularization techniques. It has demonstrated strong performance across various vision tasks. InceptionResNetV2 (Demir and Yilmaz, 2020) combines the strengths of the Inception architecture with residual connections to produce a hybrid model that balances depth and efficiency, making it highly effective for deep feature extraction. Figure 3 shows the finetuning process for the pre-trained models.

Figure 3 illustrates a deep learning model architecture for image classification using a pre-trained base model. It starts with an input image, which is passed through the pre-trained base model for feature extraction. The extracted features are then processed by a classification head, which includes a Global Average Pooling layer followed by a SoftMax (Belagatti, 2024) layer with three neurons, corresponding to the output classes: Normal, Moderate, and Severe. This model uses TL, leveraging pretrained weights for efficient feature extraction and classification.

We also finetuned a ViT_B32 model for osteoarthritis. The ViT model is a vision transformer (Han et al., 2023) with a ‘vit_b32‘ architecture (Ranftl et al., 2021). It is designed to classify images into 3 categories. It uses SoftMax activation for multi-class classification and is pre-trained for better performance. The model includes additional layers, such as batch normalization, 11 dense layers with GELU activation, and a SoftMax output layer for classification. Figure 4 shows the finetuning process for using the ViT model.

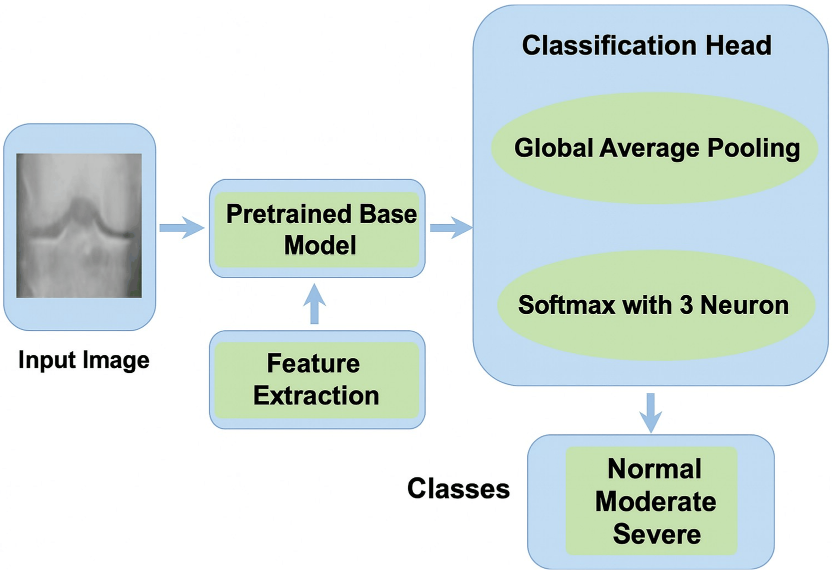

Figure 3 Fine-tuning of the transfer learning models

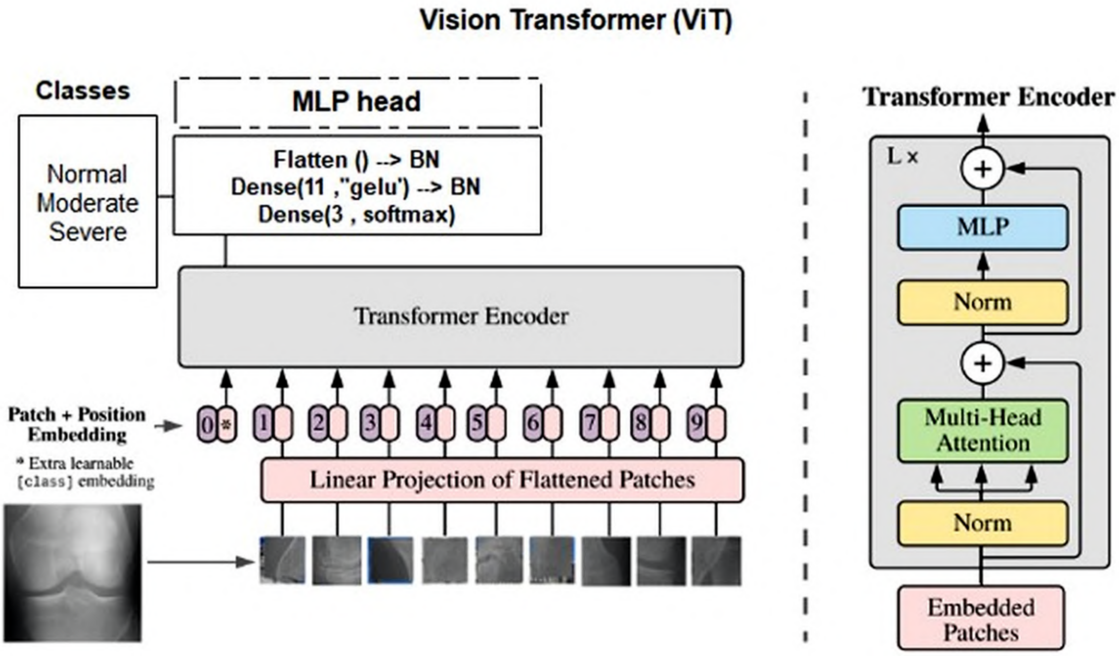

Figure 4 Finetuning of the ViT-B32 model

As illustrated in Figure 4, the ViT architecture divides the input knee X-ray image into a sequence of non-overlapping patches. Each patch is linearly projected into an embedding vector, and the spatial relationships between the patches are preserved by adding positional encodings. These embeddings, along with a learnable [class] token, are then passed into a stack of (TEBs) transfer encoder blocks. Each encoder consists of multi-head self-attention and feed-forward multilayer perceptron layers, enabling the model to capture both local texture features and long-range contextual dependencies across the image.

Normalization layers are integrated within each encoder block to stabilize training and enhance convergence. The final encoded representation corresponding to the [class] token is fed into the MLP classification head, which includes batch normalization (BN), GELU activation, and a SoftMax output layer that predicts the three severity classes — Normal, Moderate, and Severe). This hierarchical processing allows ViT to analyze global structural variations in knee joints while maintaining fine-grained feature sensitivity, making it suitable for medical image classification tasks where spatial relationships are crucial. The ViT model splits an image into patches, which are embedded with positional information and processed using a transformer encoder with multi-head attention and MLP layers (Ranftl et al., 2021). The features are then passed through an MLP head with batch normalization and GELU activation, followed by a final dense layer with SoftMax activation to classify the image into Normal, Moderate, or Severe. This approach uses transformers to efficiently capture the image context for classification.

We implemented two ensemble strategies to further enhance the robustness and generalization of the embedding process (Ganaie et al., 2022). Theoretically, ensemble learning is grounded in the principle that combining predictions from multiple diverse models can reduce both variance and bias, resulting in improved generalization and stability. In deep learning applications, individual models may capture different aspects of data distribution; therefore, aggregating their outputs through hard or weighted voting mitigates overfitting and leverages complementary strengths. This approach is particularly beneficial in medical image analysis, where variability in imaging conditions and subtle pathological features can cause underperformance of single models. The ensemble method enhances diagnostic reliability and minimizes the impact of individual model errors by integrating several high-performing CNNs. The first is a hard voting ensemble, where the majority vote among the five models determines the final prediction. This approach leverages the diversity of the models to reduce variance and improve the stability of prediction. The second strategy is a weighted voting ensemble, where each model's prediction is assigned a weight based on its individual performance. The weights used were 0.88, 0.90, 0.90, 0.85 for EfficientNetB7, DenseNet201, ResNet152V2, InceptionV3, and 0.89 for InceptionResNetV2. This method allows more accurate models to have a greater influence on the final decision, thereby improving the overall ensemble performance. The weights were assigned based on each model's performance on a validation dataset, specifically reflecting the models' individual accuracy in predicting the severity of OA. We conducted a cross-validation process where each model was evaluated on a separate validation set, and the accuracy of each model was used to determine the weight of each model relative to the other models. The proposed CNN model is a custom-built deep learning architecture designed using TensorFlow and Keras for multi-class classification of knee osteoarthritis severity. It accepts input images of size 224×224×3 and outputs predictions across three classes using a SoftMax activation function. The architecture consists of four convolutional blocks, each comprising two Conv2D layers with ReLU activation, followed by batch normalization and max-pooling to extract and reduce spatial features. The filter count progressively increases from 32 to 256, enabling the model to learn more complex patterns. After flattening the output, two dense layers with 512 and 256 units are used, each followed by dropout (0.5 and 0.3) to minimize overfitting. The model is trained using the Adam optimizer (learning rate: 0.0005) and optimized with categorical CE loss, targeting high accuracy. This architecture balances depth and regularization to achieve efficient feature learning and robust generalization in medical image classification tasks.

During model training, the validation set was used to monitor the model's performance after each epoch. An early stopping mechanism with a patience value of 60 epochs was applied to prevent overfitting and ensure optimal generalization. The model checkpoint corresponding to the best validation accuracy was automatically saved and later used for testing and comparative evaluation. This procedure ensured that the final reported results reflected the model's most stable and generalizable performance rather than a potentially overfitted state. We used hyperparameters with a batch size of 16, training for 60 epochs, and input images with 3 channels (RGB). The Adam optimizer (Kingma and Ba, 2015) was used with a learning rate of 0.001, beta_1 set to 0.9, and beta_2 set to 0.999. These values control the optimizer's momentum and variance. Dropout was included as a regularization technique (optional) to prevent overfitting, with a typical rate of approximately 0.5. The loss function used was categorical cross-entropy, appropriate for multi-class classification tasks, and the model's performance was evaluated using accuracy as the metric. This setup ensures efficient training while minimizing overfitting.

2.4 Performance Metrics

We used a comprehensive set of evaluation metrics. These include fundamental evaluation metrics obtained from the confusion matrix, such as accuracy, precision, recall, and F1-score, as well as diagnostic tools (ROC,AUC) (Hossin and bin Sulaiman, 2015). A confusion matrix is a 2×2 table that summarizes a binary classifier's prediction outcomes against the true labels. Several metrics are computed to assess the classifier based on the four fundamental outcomes. The percentage of all forecasts that the model correctly predicted is known as accuracy. It is a global indicator of the accuracy of the classifier across both classes. In this section, the terms TP, FP, FN, and TN have been clearly defined in the context of the three-class classification task. Specifically, TP refers to instances that are correctly classified as belonging to a particular class, whereas FP indicates instances that are incorrectly classified as belonging to that class. FN represents instances that belong to the class but are incorrectly classified as not belonging to it, and TN refers to instances that are correctly classified as not belonging to the given class. These definitions are now explicitly included to ensure clarity in evaluating the model's performance across all three classes. The accuracy in mathematics is defined in Eq. (4).

Accuracy is defined as the ratio of correctly classified instances to the total number of instances and is given in Eq. (4).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision quantifies how accurate positive predictions are; that is, how frequently the model correctly identifies an image as osteoporotic. It is described as in Eq. (5).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

Recall, also known as sensitivity, quantifies how well the model detects all positive cases, i.e., how many cases of osteoporotic behavior the model detected. The definition of recall is given in Eq. (6).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

The F1-score represents the harmonic mean of precision and recall. This single metric balances false positives and false negatives. The F1-score is expressed in Eq. (7).

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \tag{7}$$

Furthermore, the Jaccard Index (IoU) in Eq. (8) (Vorontsov et al., 2013) evaluates alignment in tasks such as object detection and image segmentation by measuring the overlap between anticipated and real regions.

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{8}$$

Additionally, we evaluated the ROC curve and AUC. The ROC curve illustrates the trade-off between true positive rate (TPR), also known as recall, and false positive rate (FPR) across different classification thresholds, with a better classifier pushing the curve toward the upper-left corner (Bhandari, 2020).

2.5 Results and Discussion

This section briefly presents the experimental outcomes, interprets the findings, and highlights the main conclusions drawn from the results. Table 4 presents the results of the proposed models regarding accuracy, precision, and recall. In the realm of deep learning, selecting an optimal model for image classification tasks is critical for achieving high diagnostic accuracy and clinical reliability, particularly in sensitive medical imaging applications. This study evaluates the performance of several state-of-the-art CNNs and transformer-based architectures using key evaluation metrics such as accuracy, precision, recall, F1-score, specificity, Jaccard index, and area under the curve (AUC). The models included DenseNet201, EfficientNetB7, InceptionV3, InceptionResNetV2, ResNet50V2, ResNet152V2, ViT_B32 (Vision Transformer), and Xception. Additionally, two ensemble techniques, hard voting and weighted voting, are employed to explore the benefits of model aggregation. The comparative analysis aims to identify the most robust architecture for reliable performance across multiple metrics. In this study, we did not retrain the ensemble models; instead, we used pretrained models and combined their predictions using hard and weighted voting techniques. Despite DenseNet201's comparable performance, the inclusion of ensemble methods was to investigate any potential improvements in robustness or accuracy without significantly increasing computational costs, as the models were not retrained.

The comparative analysis reveals that DenseNet201 and both ensemble methods (hard and weighted voting) achieved the highest overall performance across most metrics, each recording an accuracy, precision, recall, and F1-score of 97.11%, and specificity of 98.32%, with a Jaccard index of 94.39% (Table 4). This consistency underscores the robustness of DenseNet201 and the effectiveness of ensemble strategies in capturing the strengths of individual models while minimizing their weaknesses.

Table 4 Proposed model results

Model	Accuracy	Precision	Recall	F1-score	Specificity	Jaccard index	AUC
CNN	81.44	73.70	81.44	77.30	88.54	68.70	0.936
DenseNet201	97.11	97.10	97.11	97.10	98.32	94.39	0.992
EfficientNet-B7	95.46	95.45	95.46	95.45	97.47	91.32	0.986
InceptionV3	94.02	93.97	90.02	93.96	96.51	88.72	0.9917
InceptionResNetV2	95.67	95.93	95.67	95.72	97.86	91.70	0.9914
ResNet50	91.96	92.35	91.96	91.99	95.72	85.11	0.986
ResNet152V2	94.23	94.30	94.23	94.25	96.92	89.08	0.9887
ViT-B32	59.38	61.23	59.38	60.05	77.79	42.23	0.7369
Xception	95.88	95.89	95.88	95.87	97.66	92.08	0.9922
Ensemble hard voting	97.11	97.10	97.11	97.10	98.32	94.39	0.9928
Ensemble weighted voting	97.11	97.10	97.11	97.10	98.32	94.39	0.9928

Among the individual models, Xception and InceptionResNetV2 also demonstrated strong performance, with Xception slightly outperforming in terms of Jaccard index (92.08%) and precision (95.89%), while InceptionResNetV2 had the highest specificity (97.86%) among all the standalone models. EfficientNetB7 followed closely, with a performance F1-score of 95.45%. On the lower end of the spectrum, ViT_B32 significantly underperformed in all metrics, with the lowest accuracy (59.38%) and Jaccard index (42.23%), suggesting that the VTA in this configuration may not be well-suited for the dataset or task at hand, especially when compared to CNN-based counterparts. In addition, the proposed CNN model achieves 81.44 accuracy, which also represents the second lowest accuracy among the proposed models. The AUC in

Table 4 reflects each model's performance in terms of their ability to discriminate between classes. The AUC score ranges from 0 to 1, where a score closer to 1 indicates better model performance. In our results, the DenseNet201 model achieved the highest AUC of 0.992, followed closely by Xception and the two ensemble models (hard and weighted voting), both of which obtained AUC scores of 0.9928. These models show excellent classification capability, with AUC values indicating strong discriminatory power. The lower AUC score of 0.7369 for ViT_B32 indicates that it performs less effectively in distinguishing between the classes compared to the other models. Table 5 presents the statistical significance (p-values) between the proposed deep learning models based on their overall performance metrics.

Table 5 P-values of the proposed models

Comparison between Model A and Model B	p-value	Significance
CNN vs. DenseNet201	< 0.001	Significant
CNN vs. EfficientNetB7	< 0.001	Significant
CNN vs. InceptionV3	< 0.001	Significant
ViT_B32 vs. any other model	< 0.001	Significant
DenseNet201 vs. InceptionResNetV2	0.24	Not significant
DenseNet201 vs. Xception	0.88	Not significant
EfficientNetB7 vs. ResNet152V2	0.12	Not significant
ResNet50V2 vs. ResNet152V2	0.34	Not significant
InceptionResNetV2 vs. Ensemble (hard voting)	0.76	Not significant
Xception vs. Ensemble (weighted voting)	0.91	Not significant

As shown in Table 5, CNN showed a statistically significant difference ($p < 0.001$) when compared to the advanced architectures such as DenseNet201, EfficientNetB7, and InceptionV3, indicating that its performance was considerably lower across all evaluated metrics. Conversely, comparisons among high-performing models, including DenseNet201, InceptionResNetV2, Xception, and the Ensemble configurations—showed no statistically significant differences ($p > 0.05$). This indicates that these models achieved comparable performance levels in terms of accuracy, precision, recall, F1-score, and AUC.

Furthermore, the Vit_B32 model demonstrated a statistically significant difference ($p < 0.001$) when compared to all other models, confirming that it underperformed the remaining architectures. Overall, while traditional CNN architectures lag behind, modern hybrid models and ensemble approaches provide stable and statistically similar results, reinforcing their robustness and generalization capabilities.

The traditional ResNet variants ResNet50V2 and ResNet152V2 showed moderate performance with accuracies of 91.96% and 94.23%, respectively, but were outshined by deeper or more specialized architectures such as DenseNet and Xception. Figure 5 shows the training /validation accuracy and loss curves for the best-performing models, DenseNet201, and the ensemble methods (hard and weighted voting) over 50 epochs.

As shown in Figure 5, the accuracy graph shows a rapid increase in both training and validation accuracies, with training accuracies reaching 100% and validation accuracies above 95%, despite minor fluctuations. The loss graph indicates a sharp decline in both training and validation loss during the initial epochs, followed by consistently low loss values throughout training. These trends demonstrate that the models learned effectively, and the ensemble methods provided more stable validation performance, indicating improved generalization and robustness. Figure 6 presents the confusion matrix for the best-performing models.

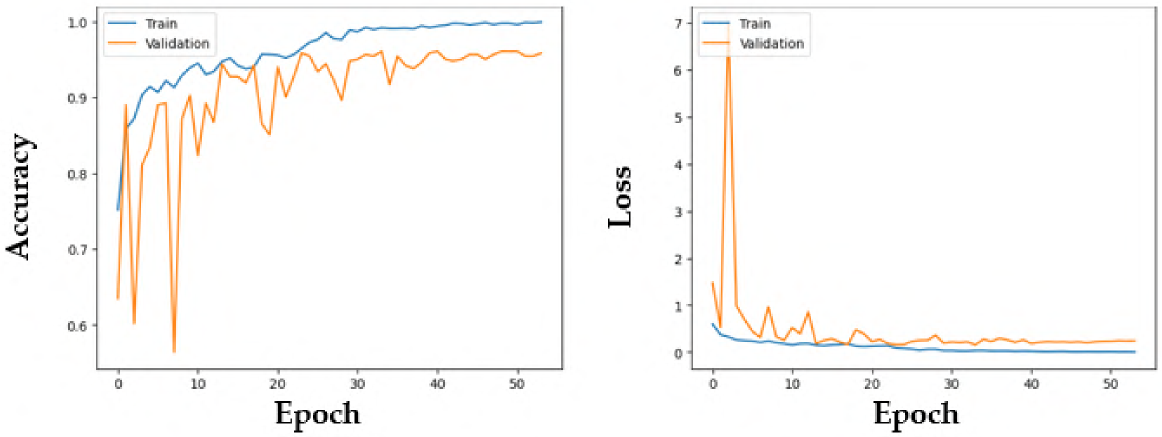

Figure 5 Accuracy curve (A) and loss curve (B) for the best-performing models: DenseNet201 and the ensemble models

Figure 6 shows the classification performance of DenseNet201 and ensemble methods across the three categories. The models accurately classified 244 Normal, 186 Moderate, and 41 Severe cases, with minimal misclassifications. Only a few moderate cases were misclassified as normal (6) or moderate (2), and (4) severe cases were predicted as moderate. The high diagonal values and low off-diagonal errors reflect strong predictive accuracy and balanced performance across all classes, highlighting the effectiveness of the models in distinguishing between varying levels of joint disease severity. Figure 7 shows the multiclass ROC curves for the best models.

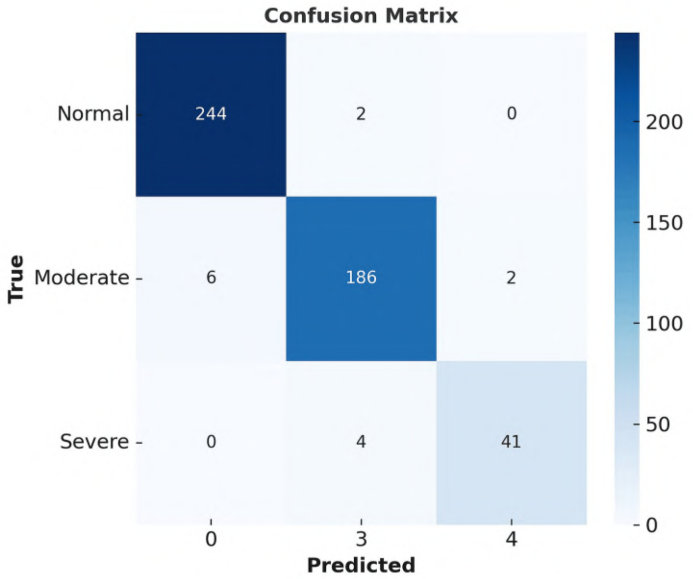

Figure 6 Confusion matrix for the best DenseNet201 and Ensemble models

As shown in Figure 7, DenseNet201 and ensemble methods—demonstrating their classification performance across three classes. The curves for all classes are clustered near the top-left corner, indicating excellent discrimination ability. The AUC values are exceptionally high (0.9947 for class 0 (normal), 0.9874 for class 1 (moderate), and 0.9946 for class 2 (severe), highlighting the strong capability of the models to distinguish between different joint disease categories with high sensitivity and specificity.

The high diagnostic accuracy and stability achieved by DenseNet201 and the ensemble models are clinically significant. In practical hospital workflows, such performance can translate into faster and more reliable preliminary screening, assisting radiologists in early detection and reducing diagnostic errors. These models can optimize clinical workload, support decision-making in resource-limited settings, and enhance overall diagnostic efficiency without replacing the clinician's judgment by automating initial image interpretation. Their robustness across multiple evaluation metrics also indicates the potential for consistent real-world deployment in clinical settings. Overall, the results highlight the superiority of DenseNet201 and the advantage of ensemble methods in enhancing classification robustness and reliability. The findings show that combining predictions from multiple high-performing models leads to performance metrics that are competitive with or exceed those of the best single model.

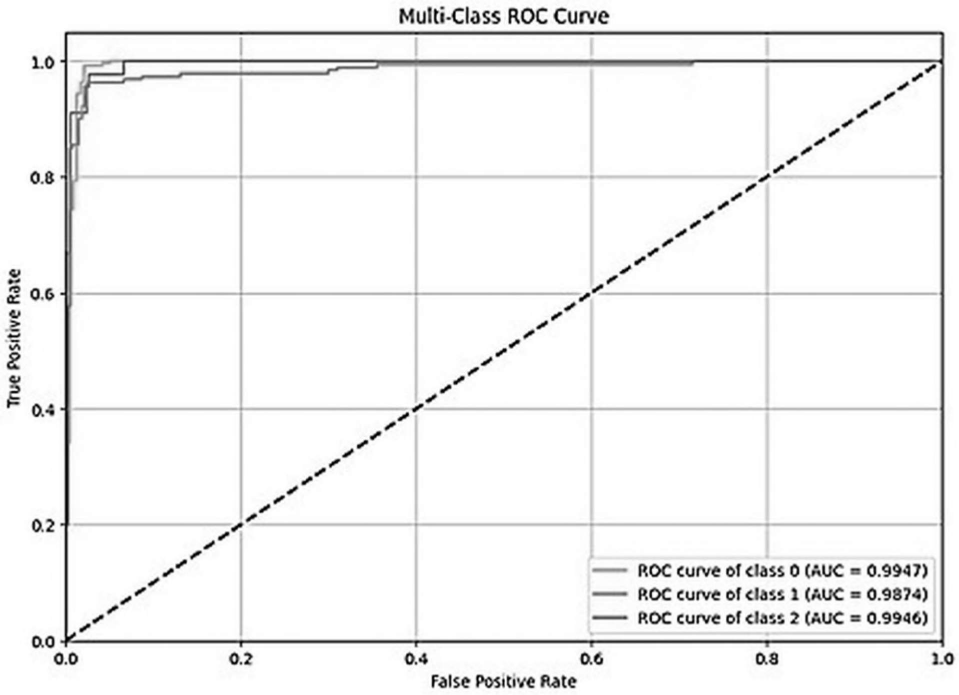

Figure 7 The receiver operating characteristic curve for the best models (DenseNet201 and Ensemble models)

3. Comparison

Table 6 presents a comparative overview of recent studies that used deep learning (DL) models for KOA classification using the Osteoarthritis Initiative (OAI) dataset (National Institutes of Health, 2012). The reviewed works vary in model architecture, dataset utilization, and classification setup (binary vs. multi-class), along with performance measures such as accuracy, area under the curve (AUC), and F1-score.

Compared to previous methods in Table 6, the proposed DenseNet201- based model outperforms all reviewed approaches, achieving 97.11% accuracy in a 3-class classification setting using the OAI dataset. This exceeds the best previously reported multi-class performance, such as 95.93% by Olsson et al., 2021 and approximately 95% by Sarvamangala and Kulkarni, 2021, both of which were also conducted on the OAI dataset.

Earlier studies such as Rani et al., 2024 and Zhang et al., 2024, despite using the same dataset, demonstrated reduced performance in multi-class configurations compared to binary setups. Even ensemble-based or customized CNN architectures in prior works did not reach the generalization capability achieved by our model.

The superior results highlight the impact of DenseNet201's dense connectivity and feature reuse, which enable more effective gradient flow and capture of subtle radiographic variations in KOA severity. Furthermore, incorporating ensemble strategies (hard and weighted voting) maintained the top accuracy level, confirming the proposed approach's robustness across the OAI dataset.

Notably, transformer-based architectures, such as ViT_B32 and conventional CNN baselines, underperformed, reinforcing that convolutional architectures remain more suitable for KOA classification on this dataset, particularly given its spatial locality and limited sample size. In conclusion, the proposed DenseNet201 model establishes a new benchmark on the OAI dataset, demonstrating superior diagnostic performance and generalization in multi-class medical image classification tasks when combined with deeper, densely connected CNNs and ensemble methods.

Table 6 Comparison with other related works

Authors	Algorithm/Model	Classes	Accuracy/Results
(Rani et al., 2024)	12-layer CNN	2, 5	92.3% (binary), 78.4% (multi-class)
(Zhang et al., 2024)	Dense Multi-Scale (DMS) CNN	5	73% ACC, 92.73% AUC
(Yong et al., 2022)	VGG, GoogLeNet, ResNet	5	DenseNet161: 88.09% ACC, Macro-MAE: 0.330, and QWK: 0.8609
(Yoon et al., 2023)	Medial OA (Custom AI)	4	KL grading ACC: 0.83, kappa: 0.768, OA diag. ACC: 0.92
(Sarvamangala and Kulkarni, 2021)	MCBCNN with MobileNetV2, ResNet50, and InceptionV3	5	ResNet50 variant: ˜95% ACC, AUC: 0.9, F1: 0.8
(Ganesh Kumar and Goswami, 2023)	CNN with sharpening image	5	The accuracy improved to 91.03%
(Olsson et al., 2021)	DenseNet169 + 5 other DL models	5, 2	Multi-class: 95.93% ACC, binary: 93.78% ACC
(Alshamrani et al., 2023)	VGG-16, ResNet-50	2	VGG-16: Train 99%, Test 92%
Proposed model	DenseNet201, hard and weighted voting	3	97.11% accuracy

4. Conclusions

This study aimed to address the challenges of accurately classifying knee OA severity from X-ray images, where subjectivity and inter-observer variability limit traditional KL grading. By fine-tuning and evaluating 11 state-of-the-art DL architectures, including both CNNs and a vision transformer, we demonstrated that deep, densely connected models, such as DenseNet201, offer superior performance in multi-class OA classification. Notably, the proposed ensemble approaches, which combine the strengths of multiple high-performing models through hard and weighted voting, consistently achieved the highest accuracy (97.11%), precision, recall, and F1-scores across all tested configurations.

Our approach establishes a new benchmark for three-class OA classification (Normal, Moderate, Severe) compared with prior work, surpassing the performance of previous single-model and multi-class strategies. The findings also confirm that in this domain, CNN-based architectures remain more effective than vision transformers, particularly when applied to datasets of limited size, where spatial locality and dense connectivity play a critical role in feature extraction. Furthermore, the integration of a tailored preprocessing pipeline, including CLAHE-based contrast enhancement, Laplacian edge sharpening, and extensive data augmentation, proved essential in improving image clarity and generalization, contributing directly to the robustness of the results. The clinical relevance of this study stems from its potential to support decision-making in routine radiographic analysis. The proposed models can serve as reliable tools to assist radiologists and orthopedic specialists in early detection and severity assessment of OA, ultimately improving patient management and treatment planning. The use of ensemble strategies also enhances diagnostic confidence by minimizing the weaknesses of individual models, which is particularly important in medical applications where consistency is paramount.

Despite these promising results, certain limitations should be acknowledged. Although widely used, the dataset is limited in terms of class balance, with fewer severe cases than normal or moderate grades. Additionally, the study was conducted solely on the OAI dataset, raising questions regarding the generalizability of the study to diverse populations and imaging protocols. While the results highlight the superiority of DenseNet201 and ensemble methods, a deeper theoretical exploration of why such architectures excel in this context is still needed.

Future research should explore integration with multimodal data, such as clinical records and longitudinal patient outcomes, to improve the applicability of the model in real-world settings. Expanding to larger and more heterogeneous datasets, investigating explainable AI approaches, and validating models in prospective clinical studies will be essential steps toward translating this work into practical clinical tools. In summary, this study demonstrates that DL, particularly DenseNet201 and ensemble-based strategies, provides a highly accurate, robust, and clinically meaningful framework for OA classification. The findings not only advance the state of the art in medical image analysis but also underscore the transformative potential of AI-driven diagnostic support in orthopedics.

List of abbreviations

Abbreviation	Full term	Abbreviation	Full term
ACC	Accuracy	AI	Artificial Intelligence
AUC	Area Under the Curve	CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network	DL	Deep Learning
FN	False Negative	FP	False Positive
BN	Batch Normalization	KL	Kellgren–Lawrence (grading scale)
KOA	Knee Osteoarthritis	MAE	Mean Absolute Error
OAI	Osteoarthritis Initiative	OA	Osteoarthritis
ReLU	Rectified Linear Unit	ROC	Receiver Operating Characteristic
TP	True Positive	TN	True Negative
ViT	Vision Transformer	MCFFs	Multiscale Convolutional Filters
TL	Transfer Learning	HT	Hyperparameter Tuning
TEBs	Transformer Encoder Blocks	CE	Cross Entropy
TPR	True Positive Rate	VTA	Vision Transformer Architecture
FPR	False Positive Rate	TTL	Two-phase Transfer Learning
GELU	Gaussian Error Linear Unit		

Acknowledgements

We extend our deepest appreciation to our supervisors and colleagues for their constructive feedback and technical guidance throughout this study. The authors gratefully acknowledge the Osteoarthritis Initiative for providing access to the imaging dataset used in the experiments. We would also like to thank the supporting institution and technical staff for their assistance with computational resources and data management. The authors contributions were invaluable to the successful completion of this research.

Author Contributions

Israa S. Abed led the conceptualization, methodology design, data curation, software development, and manuscript drafting. Abeer Twakol Khalil provided supervision, validation, critical review, and interpretation of the study results. Hanan M. Amer contributed to the formal analysis, visualization, and manuscript revision. Samer Mahmoud Mohamed provided medical expertise, validated data, and ensured clinical relevance. Mohamed Maher Ata provided supervision, methodology, software, concept design, data handling, visualization, investigation, writing, and editing.

Conflict of Interest

The authors have no conflicts of interest to declare.

Data Availability

The knee X-ray data used in this study are publicly available through the OAI and can be accessed via Mendeley Data (Chen, 2018). This study used the publicly available Osteoarthritis Initiative (OAI) dataset, which contains de-identified knee X-ray images collected under informed consent and approved institutional review board (IRB) protocols. All data were used in accordance with the OAI data usage policy and ethical research standards for secondary data analysis. Because the dataset is fully anonymized and distributed under an open-access license, no additional ethical approval was required for this work. All experimental procedures and data handling adhered to the principles of responsible research and data privacy.

Appendix

The model was trained using a Lenovo Legion 5 laptop equipped with an NVIDIA GeForce RTX 3060 GPU, which provides the necessary computational power for DL tasks. The system runs on Windows 11 or Linux (preferably Ubuntu) and has at least 16 GB of RAM to efficiently handle large datasets. Additionally, a solid-state drive (SSD) was used for faster data read/write speeds, ensuring smooth training and handling of large models. This setup is ideal for performing deep learning tasks that require high-performance hardware.

References

Ahmed, R., & Imran, A. S. (2024). Knee osteoarthritis analysis using deep learning and xai on x-rays. *IEEE Access*, *12*, 68870–68879. https://doi.org/10.1109/ACCESS.2024.3400987

Alshamrani, H. A., Rashid, M., Alshamrani, S. S., & Alshehri, A. H. D. (2023). Osteo-net: An automated system for predicting knee osteoarthritis from x-ray images using transfer-learning-based neural networks approach. *Healthcare*, *11*(9). https://doi.org/10.3390/healthcare11091206

Belagatti, P. (2024). Understanding the softmax activation function: A comprehensive guide [SingleStore Blog]. https://www.singlestore.com/blog/a-guide-to-softmax-activation-function/

Bhandari, A. (2020). Auc-roc curve in machine learning clearly explained [Online article].

Borawar, L., & Kaur, R. (2023). Resnet: Solving vanishing gradient in deep networks. In *Lecture notes in networks and systems* (pp. 235–247, Vol. 600). https://doi.org/10.1007/978-981-19-8825-7_21

Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2018). Convolutional neural network (cnn) for image detection and recognition. *Proceedings of the International Conference on Secure Cyber Computing and Communications (ICSCCC)*, 278–282. https://doi.org/10.1109/ICSCCC.2018.8703316

Chen, P. (2018). *Knee osteoarthritis severity grading dataset*. Mendeley Data.

Demir, A., & Yilmaz, F. (2020). Inception-resnet-v2 with leakyrelu and average pooling for more reliable and accurate classification of chest x-ray images. *TIPTEKNO 2020 Medical Technologies Congress*. https://doi.org/10.1109/TIPTEKNO50054.2020.9299232

DenseNets, F., Background, I., & What, M. (2021). Understanding and visualizing densenets [Technical article].

Duklan, N., Kumar, S., Maheshwari, H., Singh, R., Sharma, S. D., & Swami, S. (2024). Cnn architectures for image classification: A comparative study using resnet50v2, resnet152v2, inceptionv3, xception, and mobilenetv2. *SSRG International Journal of Electronics and Communication Engineering*, *11*(9), 11–21. https://doi.org/10.14445/23488549/IJECE-V11I9P102

Fajarani, R., Rahman, S. F., Pangesty, A. I., Katili, P. A., Park, D. H., & Basari. (2024). Physical and chemical characterization of collagen/alginate/poly (vinyl alcohol) scaffold with the addition of multi-walled carbon nanotube, reduced graphene oxide, titanium dioxide, and zinc oxide materials. *International Journal of Technology*, *15*(2), 332–341. https://doi.org/10.14716/ijtech.v15i2.6693

Foti, G., & Longo, C. (2024). Deep learning and ai in reducing magnetic resonance imaging scanning time: Advantages and pitfalls in clinical practice. *Polish Journal of Radiology*, *89*, 443–451. https://doi.org/10.5114/pjr/192822

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*. https://doi.org/10.1016/j.engappai.2022.105151

Ganesh Kumar, M., & Goswami, A. D. (2023). Automatic classification of the severity of knee osteoarthritis using enhanced image sharpening and cnn. *Applied Sciences*, *13*(3). https://doi.org/10.3390/app13031658

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

Hossin, M., & bin Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Management Process*, *5*, 1–11. https://doi.org/10.5121/ijdkp.2015.5201

Kaur, P., Kohli, G. S., Bedi, J., & Wasly, S. (2024). A novel deep learning approach for automated grading of knee osteoarthritis severity. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-20322-8

Keith Sinusas, M. (2012). Osteoarthritis: Diagnosis and treatment. *American Family Physician*, *85*(1), 49–56.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 1–15.

Kondal, S., Kulkarni, V., Gaikwad, A., Kharat, A., & Pant, A. (2022). Automatic grading of knee osteoarthritis on the kellgren-lawrence scale from radiographs using convolutional neural networks. In *Lecture notes in networks and systems* (pp. 163–173, Vol. 249). https://doi.org/10.1007/978-3-030-85365-5_16

Koonce, B. (2021). Resnet 50. In *Convolutional neural networks with swift for tensorflow* (pp. 63–72). https://doi.org/10.1007/978-1-4842-6168-2_6

Mlsna, P. A., & Rodriguez, J. J. (2009). Gradient and laplacian edge detection. In *The essential guide to image processing* (pp. 495–524). https://doi.org/10.1016/B978-0-12-374457-9.00019-6

National Institutes of Health. (2012). *Osteoarthritis initiative (oai) dataset*. https://nda.nih.gov/oai

Olsson, S., Akbarian, E., Lind, A., Razavian, A. S., & Gordon, M. (2021). Automating classification of osteoarthritis according to kellgren-lawrence in the knee using deep learning in an unfiltered adult population. *BMC Musculoskeletal Disorders*, *22*(1), 1–8. https://doi.org/10.1186/s12891-021-04722-7

Ou, J., Zhang, J., Alswadeh, M., Zhu, Z., Tang, J., Sang, H., & Lu, K. (2025). Advancing osteoarthritis research: The role of ai in clinical, imaging and omics fields. *Bone Research*, *13*(1). https://doi.org/10.1038/s41413-025-00423-2

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *IARJSET*, 20–22. https://doi.org/10.17148/iarjset.2015.2305

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. *Proceedings of the IEEE International Conference on Computer Vision*, 12159–12168. https://doi.org/10.1109/ICCV48922.2021.01196

Rani, S., Memoria, M., Almogren, A., Bharany, S., Joshi, K., Altameem, A., Rehman, A. U., & Hamam, H. (2024). Deep learning to combat knee osteoarthritis and severity assessment by using cnn-based classification. *BMC Musculoskeletal Disorders*, *25*(1). https://doi.org/10.1186/s12891-024-07942-9

Sarvamangala, D. R., & Kulkarni, R. V. (2021). Grading of knee osteoarthritis using convolutional neural networks. *Neural Processing Letters*, *53*(4), 2985–3009. https://doi.org/10.1007/s11063-021-10529-3

Sharma, R., & Kamra, A. (2023). A review on clahe based enhancement techniques. *Proceedings of the International Conference on Contemporary Computing and Informatics (IC3I)*, 321–325. https://doi.org/10.1109/IC3I59117.2023.10397722

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning (ICML)*, 10691–10700.

Tariq, T., Suhail, Z., & Nawaz, Z. (2025). A review for automated classification of knee osteoarthritis using kl grading scheme for x-rays. *Biomedical Engineering Letters*, *15*(1). https://doi.org/10.1007/s13534-024-00437-5

Touahema, S., Zaimi, I., Zrira, N., Ngote, M. N., Doulhousne, H., & Aouial, M. (2024). Medknee: A new deep learning-based software for automated prediction of radiographic knee osteoarthritis. *Diagnostics*, *14*(10). https://doi.org/10.3390/diagnostics14100993

Vorontsov, I. E., Kulakovskiy, I. V., & Makeev, V. J. (2013). Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms for Molecular Biology*, *8*(1). https://doi.org/10.1186/1748-7188-8-23

Wing, N., Van Zyl, N., Wing, M., Corrigan, R., Loch, A., & Wall, C. (2021). Reliability of three radiographic classification systems for knee osteoarthritis among observers of different experience levels. *Skeletal Radiology*, *50*(2), 399–405. https://doi.org/10.1007/s00256-020-03551-4

Xie, X., Zhang, K., Li, Y., Li, Y., Li, X., Lin, Y., Huang, L., & Tian, G. (2025). Global, regional, and national burden of osteoarthritis from 1990 to 2021 and projections to 2035: A cross-sectional study for the global burden of disease study 2021. *PLOS One*, *20*(5). https://doi.org/10.1371/journal.pone.0324296

Xin Teoh, Y., Othmani, A., Li Goh, S., Usman, J., & Lai, K. W. (2024). Deciphering knee osteoarthritis diagnostic features with explainable artificial intelligence: A systematic review. *IEEE Access*, *12*, 109080–109108. https://doi.org/10.1109/ACCESS.2024.3439096

Yong, C. W., Teo, K., Murphy, B. P., Hum, Y. C., Tee, Y. K., Xia, K., & Lai, K. W. (2022). Knee osteoarthritis severity classification with ordinal regression module. *Multimedia Tools and Applications*, *81*(29), 41497–41509. https://doi.org/10.1007/s11042-021-10557-0

Yoon, J. S., Yon, C. J., Lee, D., Lee, J. J., Kang, C. H., Kang, S. B., Lee, N. K., & Chang, C. B. (2023). Assessment of a novel deep learning-based software developed for automatic feature extraction and grading of radiographic knee osteoarthritis. *BMC Musculoskeletal Disorders*, *24*(1), 1–10. https://doi.org/10.1186/s12891-023-06951-4

Zhang, D., Dong, Y., Xu, Y., Qian, J., Ye, M., Yuan, Q., & Luo, J. (2024). Enhancing knee osteoarthritis diagnosis with dms: A novel dense multi-scale convolutional neural network approach. *Journal of Orthopaedic Surgery and Research*, *19*(1), 1–9. https://doi.org/10.1186/s13018-024-05352-0

Zhao, H., Ou, L., Zhang, Z., Zhang, L., Liu, K., & Kuang, J. (2025). The value of deep learning-based x-ray techniques in detecting and classifying k-l grades of knee osteoarthritis: A systematic review and meta-analysis. *European Radiology*, *35*(1), 327–340. https://doi.org/10.1007/s00330-024-10928-9

Zhu, Y., & Newsam, S. (2018). Densenet for dense flow. *Proceedings of the International Conference on Image Processing (ICIP)*, 790–794. https://doi.org/10.1109/ICIP.2017.8296389