Ensemble Modeling for Enhanced Severity Grading of Knee Osteoarthritis Detection

Abubaker Ben Chatwan [1]
International Alsalam University
Faculty of Computer
Technology and Information
Benghazi, Libya
Email:
abobaker.alsadeek.chatwan@gmail.com

Hassan Masoud Al-Gattani[2]
International Alsalam University
Faculty of Computer
Technology and Information
Benghazi, Libya

Salima Al-Warfali[3]
International Alsalam University
Faculty of Computer
Technology and Information
Benghazi, Libya

Aisha Ben Ismail[4]
International Alsalam University
Faculty of Computer
Technology and Information
Benghazi, Libya

***Abstract*—** Knee osteoarthritis (KOA) is a widespread and debilitating condition. Early and accurate assessment of its severity is vital for successful patient care. This study introduces a robust, automated deep learning system for detecting and classifying KOA severity directly from knee X-ray images, utilizing the Osteoarthritis Initiative (OAI) dataset. Our multi-stage system fine-tuned several pre-trained Convolutional Neural Networks (CNNs), including DenseNet201, InceptionV3, and Xception, for tasks ranging from binary classification (diseased vs. healthy) to detailed multi-class severity grading. Key findings demonstrate the power of these models:

- Binary Classification: DenseNet201 and InceptionV3 achieved high accuracy (96% and 98%, respectively).
- Multi-Class Grading: The Xception model excelled in the challenging five-class severity grading, reaching 95% accuracy.
- Ensemble Advantage: Crucially, an ensemble model combining DenseNet201 and InceptionV3 outperformed individual models, achieving a peak accuracy of 98% in binary classification by effectively combining features.

The results strongly suggest that deep learning, especially through ensemble approaches, offers a promising path for providing radiologists with a consistent and objective tool for the early and precise diagnosis of KOA.

Keywords; *Knee Osteoarthritis (KOA); Severity Detection; Ensemble Modeling; Osteoarthritis Initiative (OAI).*

I. INTRODUCTION

KOA is globally the most prevalent musculoskeletal disorder and a leading chronic joint disease, widely affecting millions, particularly the elderly, overweight individuals, and those with sedentary lifestyles (Vos et al., 2017). The disease is characterized by the progressive degradation of articular cartilage in the knee joint, resulting in chronic pain, stiffness, and severely reduced range of motion, thus significantly impairing an individual's quality of life and daily functionality.

Radiography (X-ray imaging) remains the gold standard for diagnosing KOA due to its accessibility, cost-effectiveness, and safety. X-rays are crucial for assessing the structural changes indicative of the disease, clearly revealing features such as joint space narrowing, osteophytes (bone spurs), and subchondral sclerosis. The clinical severity of radiographic KOA is universally quantified using the Kellgren–Lawrence (KL) grading system (Lee et al., 2023). This system, adopted as the standard by the World Health Organization in 1961, assigns five severity grades (0 to 4) to describe the various stages of the disease. Table 1 provides a detailed description of the criteria for each KL grade, and the progressive stages of the disease are visually illustrated in Figure 1 (Lee et al., 2023).

Table 1: Description of the KL grading system (Lee et al., 2023).

Grades	Severity	Description of the KL grading
Grade 0	normal	Definite absence of X-ray changes of osteoarthritis.
Grade 1	doubtful	joint space narrowing and possible osteophytic lipping.
Grade 2	minimal	Definite osteophytes and possible joint space narrowing.
Grade 3	moderate	multiple osteophytes, definite narrowing of joint space and some sclerosis, and possible deformity of bone ends.
Grade 4	severe	Large osteophytes, marked narrowing of joint space, severe sclerosis, and definite deformity of bone ends.

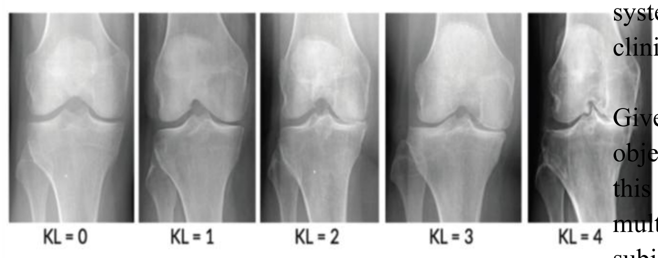


Figure 1 : Stage of knee osteoarthritis (Lee et al., 2023).

Despite the utility of X-ray imaging, its interpretation is inherently subjective. Diagnosis relies heavily on the expertise of radiologists and orthopedic specialists, often resulting in significant inter-observer variability (Altman et al., 2009) and limiting the objectivity of clinical decision-making. These limitations underscore a critical need for objective and reproducible assessment methods to enhance diagnostic consistency and accuracy.

Recent advancements in deep learning and computer vision offer a powerful pathway to address these diagnostic challenges. CNNs, a specialized form of deep learning, have shown remarkable performance in medical image analysis by automatically learning and extracting complex, nuanced features from medical images (Litjens et al., 2017). Applying CNN-based models to KOA detection and KL severity classification provides an automated and standardized approach. This methodology effectively minimizes the subjectivity and variability associated with human interpretation. Furthermore, the efficiency of CNNs in rapidly analyzing large datasets makes them invaluable tools for triaging and diagnosing KOA cases. Therefore, leveraging Deep Learning techniques holds immense promise for developing a robust automated system that enhances diagnostic accuracy, supports consistent clinical decision-making, and ultimately improves patient outcomes.

II. Problem Statement

KOA a highly prevalent degenerative joint disease, profoundly impacts patient mobility and quality of life, particularly among the elderly. Accurate and timely assessment of KOA severity based on radiographic findings is crucial for effective treatment planning and monitoring disease progression (Bijlsma et al., 2023). However, the traditional diagnostic approach faces significant limitations. The manual interpretation of knee X-rays by specialists is inherently subjective, time-consuming, and notoriously prone to inter-observer variability (Tiulpin et al., 2020). This subjectivity hinders the consistent application of standardized grading systems, such as the KL system, which is essential for uniform clinical decision-making.

Given these challenges, there is a critical need for an objective, reliable, and automated solution. The core problem this study addresses is the lack of a robust, standardized, multi-stage Deep Learning system capable of overcoming the subjectivity of manual assessment to accurately and efficiently classify KOA severity stages from X-ray images. Developing such a system is vital to enhance diagnostic consistency, optimize treatment decisions, and ultimately improve patient outcomes in managing this widespread musculoskeletal condition.

III. Aims and Objectives

The primary aim of this study is to develop and rigorously evaluate a robust, ensemble deep learning system for the automated detection and accurate multi-class severity grading of KOA from X-ray images. This work seeks to provide an objective, reliable diagnostic tool to support radiologists and enhance the consistency of clinical decision-making.

To achieve this overarching aim, the specific objectives are:

1. To conduct a comprehensive review of existing Deep Learning models for KOA severity classification, specifically identifying performance limitations in distinguishing between subtle disease stages.

2. To investigate, fine-tune, and optimize various state-of-the-art CNNs (e.g., DenseNet201, InceptionV3, Xception) for both binary and multi-class KOA classification tasks using the OAI dataset.

3. To design and implement an ensemble feature extraction and classification model capable of leveraging complementary information from multiple CNN architectures (DenseNet201 and InceptionV3) to maximize diagnostic accuracy.

4. To systematically evaluate the performance of the proposed ensemble system against individual CNN models and existing literature, using metrics such as accuracy, precision, and F1-score, focusing particularly on its effectiveness in the challenging five-class KL grading.

IV. Research Questions

This study seeks to answer the following research questions:

- Performance: To what extent can the fine-tuned Convolutional Neural Networks (DenseNet201, InceptionV3, Xception) accurately detect and classify the severity levels of KOA from X-ray images?

- Ensemble Superiority: Does the proposed ensemble model significantly outperform individual CNN architectures (DenseNet201, InceptionV3) in terms of accuracy and diagnostic consistency for KOA classification?

- Severity Grading: What is the achieved accuracy of the system in the challenging multi-class five-level KL grading task, particularly when compared to binary classification performance?

V. Literature Review

The development of automated tools for assessing the severity of KOA using deep learning has attracted substantial research interest, primarily to mitigate the high inter-observer variability inherent in manual radiographic grading (Tiulpin et al., 2018). Existing literature in this field can be broadly classified into three main directions: single-network approaches, performance enhancement through feature engineering and preprocessing, and ensemble-based methods.

Early studies established the feasibility of applying CNNs for KOA classification; however, their performance was generally limited. Antony et al. (2017) employed object detection combined with CNNs on the OAI and MOST datasets, achieving a multiclass accuracy of 63.4%. Similarly, single-network architectures such as VGG-16 and VGG-19 reported moderate accuracies ranging between 64% and 70% (Chen et al., 2019; Górriz et al., 2019). While these studies provided important baseline results, they highlighted the difficulty of achieving consistently high and balanced performance across all five KL grades.

Subsequent research aimed to improve classification accuracy by emphasizing Region of Interest localization and adopting more advanced detection and segmentation techniques. Swiecicki et al. (2021) incorporated multi-view X-ray images with Faster R-CNN, achieving an accuracy of 71.90%, whereas Wang et al. (2021) employed YOLO-based segmentation with a ResNet50 backbone, reaching 69.18%. Despite these architectural improvements, multiclass classification accuracy—particularly for intermediate KL grades—remained moderate, indicating limitations related to network backbones and loss formulations. Later studies leveraged transfer learning and deeper architectures to achieve noticeable performance gains. Ahmed and Mohammed (2022) reported a validation accuracy of 91.51% using a fine-tuned ResNet50 model, while Goswami (2023) achieved 91.03% accuracy by combining ResNet-V2 with image sharpening. In contrast, other approaches continued to exhibit suboptimal performance; for example, the Deep Siamese CNN with a fine-tuned ResNet-34 proposed by Cueva et al. (2022) achieved only 61% accuracy, reflecting sensitivity to dataset composition and training strategies.

Hybrid frameworks that combine deep feature extraction with traditional machine learning classifiers were also explored. Teo et al. (2022) utilized InceptionV3 and DenseNet201 as feature extractors and employed an SVM classifier, with the DenseNet201-SVM model achieving a maximum accuracy of 71.33%. Similarly, Ahmed and Mstafa (2022) proposed a hybrid approach integrating deep and handcrafted features using PCA and SVM; however, class imbalance remained a critical limitation, particularly affecting minority KL grades.

Regarding ensemble-based approaches with moderate performance (below 94%), several studies reported incremental yet limited improvements. Pongsakonpruttikul et al. (2022) applied a YOLOv3-tiny model for KOA classification, achieving accuracies between 85% and 86.7%. Yunus et al. (2022) combined YOLO-v2 features with KNN and SVM classifiers in an ensemble framework, reaching approximately 89% accuracy. Likewise, Mohammed et al. (2023) evaluated multiple pre-trained CNNs independently and reported a maximum accuracy of 89% across different datasets. More recent studies from 2024 and 2025 continued to explore ensemble strategies without achieving consistently high performance, particularly in binary and five-class classification tasks. Kumar et al. (2024) proposed a weighted ensemble of ResNet50 and EfficientNet-B4, achieving 92.4% accuracy in three-class classification but reporting reduced performance in binary settings. Li et al. (2024) combined DenseNet121 and MobileNetV2 using multi-level feature fusion, achieving 90.8% accuracy, with limited improvements for intermediate KL grades. In another study, Park et al. (2024) introduced an ensemble of lightweight CNNs for five-class KOA grading and reported an overall accuracy of 88.6%, highlighting challenges in class separability. More recently, Zhang et al. (2025) proposed a lightweight ensemble combining InceptionV3 and Xception with a Gradient Boosting classifier, achieving 93.2% accuracy in four-class classification, while performance for KL-1 and KL-2 remained notably lower.

Although certain ensemble models achieved high accuracy when distinguishing between normal and severe cases, most studies—including recent ones (Ahmed & Imran, 2024; Zhang et al., 2025)—consistently reported two persistent challenges: class imbalance leading to poor performance for intermediate KL grades and limited interpretability of complex deep models. Therefore, there remains a clear need for a unified and systematic framework that effectively integrates complementary feature representations from multiple efficient CNN architectures while ensuring balanced, robust, and interpretable diagnostic performance across all KL grades. This study aims to address these limitations.

Despite notable advances in deep learning–based KOA severity assessment, several critical gaps persist. Most existing studies suffer from severe class imbalance, leading to inflated accuracy for majority or extreme KL grades and degraded performance for intermediate stages (KL 1–2). Moreover, the majority of prior work focuses on a single classification setting, typically binary or five-class grading, without jointly modeling binary, three-class, four-class, and five-class tasks to reinforce feature learning and diagnostic consistency. Although ensemble methods have improved robustness, they are often constructed without systematic exploitation of complementary features from heterogeneous CNN backbones or explicit strategies to address imbalance. These limitations highlight the need for a unified, balanced, and multi-stage framework capable of consistent KOA assessment across all KL grades.

VI. PROPOSED METHODOLOGY

This section delineates the multi-stage framework developed for KOA detection and severity grading. The proposed methodology is structured into four primary phases: image acquisition, preprocessing, model training, and classification. The conceptual workflow of the entire system is illustrated in Figure 2.

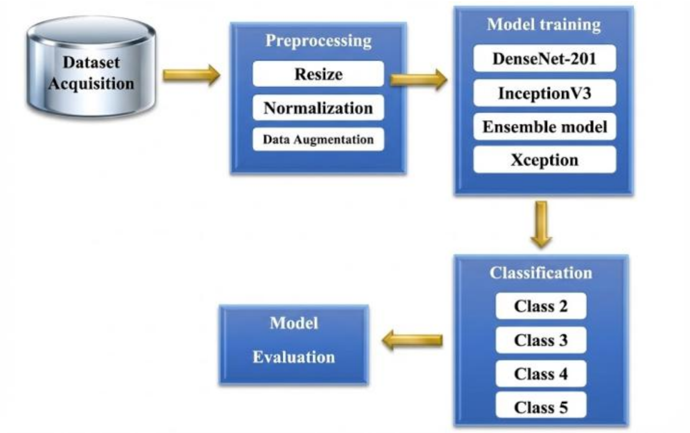


Figure 2: Block Diagram of KOA Diagnosing.

A. Dataset Distribution

The radiographic images used in this study were obtained from the OAI. To evaluate the robustness of the proposed models across different clinical diagnostic requirements, the data was organized into four distinct datasets:

- **Dataset I (Baseline 5-Class):** Represents the original KL grading (0–4).
  - **Training:** 2286 Healthy (He), 1046 Doubtful (D), 1516 Minimal (Min), 757 Moderate (Mod), 173 Severe (Sev).
  - **Testing:** 303 He, 142 D, 205 Min, 103 Mod, 24 Sev.
- **Dataset II (Binary Task):** Formulated for KOA diagnosis (Negative vs. Positive).
  - **Training:** 2300 samples per class.
  - **Testing:** 550 samples per class.
- **Dataset III (3-Class Task):** Focused on symptomatic disease (Grades 2, 3, and 4).
  - **Training:** 1230 (Min), 1220 (Mod), 1110 (Sev + Data Augmentation).
  - **Testing:** 220 (Min), 220 (Mod), 110 (Sev).
- **Dataset IV (4-Class Task):** Focused on specific severity grades (0, 2, 3, and 4).
  - **Training:** 960 (He), 800 (Min), 960 (Mod), 233 (Sev).
  - **Testing:** 240 (He), 200 (Min), 240 (Mod), 57 (Sev).

B. Proposed Model Architectures

1) Modified DenseNet-201

DenseNet-201 (Huang et al., 2017) is characterized by its dense connectivity pattern, where each layer receives feature maps from all preceding layers. This design promotes extensive feature reuse and significantly reduces the number of parameters compared to traditional deep CNNs.
Base Model: 201 layers with an input resolution of 224 * 224 * 3 as illustrated in Figure 3.

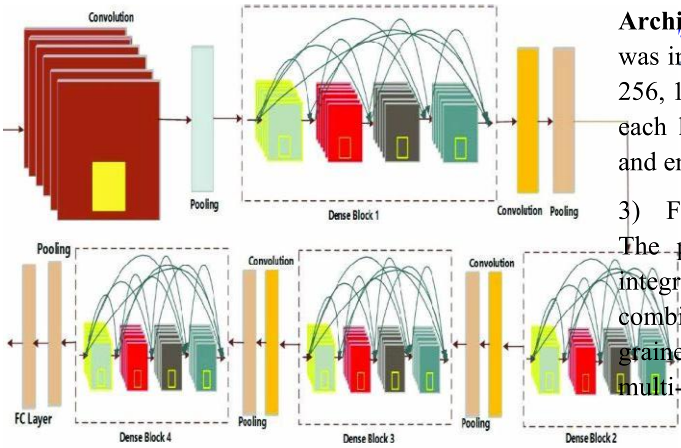


Figure 3: Architecture of DenseNet-201.

Architectural Modifications: The final classification layer was replaced by a Global Average Pooling layer and a deep MLP head. This head comprises seven dense layers with units structured as 1024 → 512 →256 →128 →64 →32 →16. Each layer utilizes the ReLU activation function and a 20% Dropout rate to mitigate the risk of overfitting during the training phase.

2) Modified InceptionV3

InceptionV3 (Szegedy et al., 2016) utilizes a combination of factorized and asymmetric convolutions within specialized "Inception Modules." This allows the network to capture multi-scale features while maintaining a high level of computational efficiency.

Base Model: 48 layers incorporating auxiliary classifiers to stabilize the learning process and improve convergence, as illustrated in Figure 4.

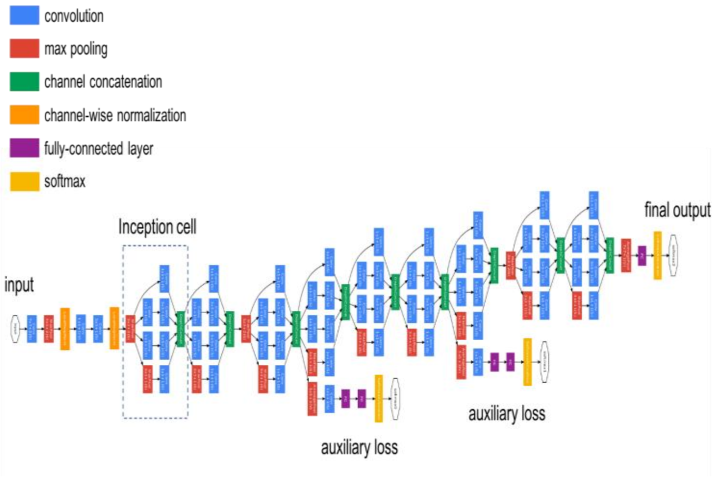


Figure 4: InceptionV3 model architecture.

Architectural Modifications: A custom classification head was integrated, featuring a sequence of dense layers with 512, 256, 128, 64, and 32 units. A 10% Dropout rate was applied to each layer to refine high-dimensional feature representations and enhance generalization.

3) Feature Fusion Ensemble Model

The proposed Ensemble model represents a sophisticated integration of DenseNet201 and InceptionV3 architectures. By combining these models, the framework leverages the fine-grained spatial details captured by DenseNet201 alongside the multi-scale features identified by InceptionV3.

Parallel Extraction: Both models process the same input image simultaneously, generating two independent high-level feature vectors.

Feature Concatenation: These vectors are concatenated into a unified feature space, providing a more comprehensive representation of the radiographic indicators of KOA.

Refinement Head: The concatenated vector is processed through an 8-layer deep MLP refinement head, starting from 2048 units down to 16 units. Integrated dropout layers (20% and 10%) are strategically placed to ensure robust performance on unseen datasets.

4) Modified Xception

The Xception model introduces an advanced extension of the Inception architecture by replacing standard Inception modules with depth wise separable convolutions. This design allows for independent learning of spatial and cross-channel correlations, leading to more efficient feature extraction (Chollet, 2017)**,** as shown in Figure 5.

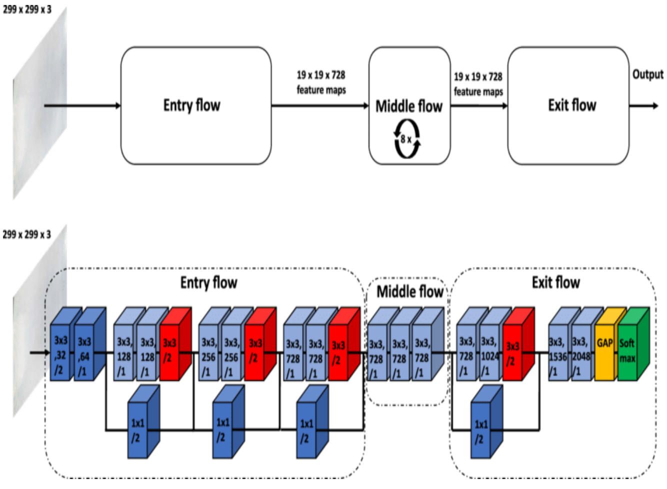


Figure 5: Xception model architecture

Task Optimization: This model was specifically fine-tuned for the complex 5-class classification task (Dataset I).

Architectural Modifications: The architecture features a series of Conv2D reduction layers designed to progressively decrease spatial dimensions. This is followed by a Global Average Pooling layer and a condensed MLP head (64 → 32 →16 units) to provide precise predictions across the full KL severity spectrum.

C. Training Methodology

All models were trained using a consistent optimization pipeline to ensure scientific comparability:

- **Loss Function:** Categorical Cross-Entropy was employed for multi-class tasks, while Binary Cross-Entropy was used for the diagnosis task in Dataset II.
- **Optimizer:** The Adam optimizer was utilized with a dynamic learning rate.
- **Final Layer:** A **SoftMax** activation function was applied to produce the final classification probabilities for the respective KOA grades.

VII. Experimental Results and Discussion

D. . Implementation Environment

The experimental phase was conducted using Google Collaboratory, leveraging its cloud-based environment. To handle the high computational demands of deep CNNs, a Tesla K80 GPU was utilized, which provided the necessary power for training and fine-tuning the models within a research-efficient timeframe.

E. *Experiment 1: Binary Classification (Dataset II)*

This experiment focused on the primary detection of KOA (Healthy vs. Positive). All models demonstrated an exceptional ability to distinguish between these two broad categories. Figure 6 below illustrates the confusion matrix of the ensemble model.

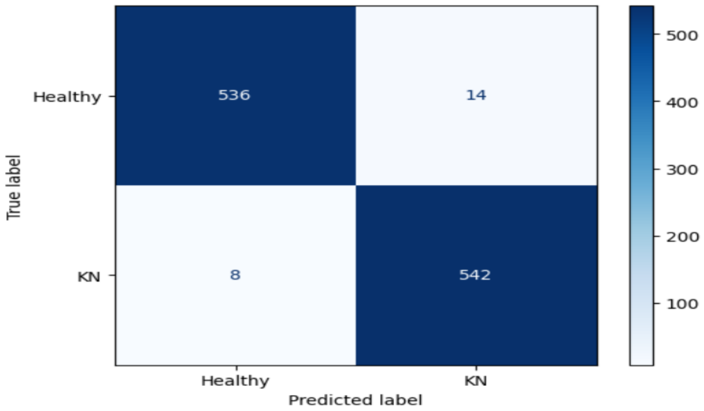


Figure 6: Confusion matrix of the Ensemble model.

Performance Analysis: Both the Ensemble Model and InceptionV3 achieved a top accuracy of 98%, while DenseNet201 followed closely at 96**%**.

Classification Report Summary: The Ensemble model showed a Precision of 97% and a Recall of 98%. The F1-score was consistently 98% for the Positive (Severe) category. These metrics confirm that combining multi-scale and dense features effectively eliminates most false positives in primary screening.

F. *Experiment 2: Three-Class (Dataset III)*

This experiment focused on categorizing symptomatic KOA stages into three levels of severity: Minimal, Moderate, and Severe. The results demonstrated a high degree of precision in distinguishing between these radiographic stages.

Performance Analysis: The Ensemble Model achieved the highest overall accuracy of 97%. DenseNet201 followed with an accuracy of 96%, while InceptionV3 recorded 95% accuracy. The training and testing accuracy of the ensemble model are illustrated in Figure 7.

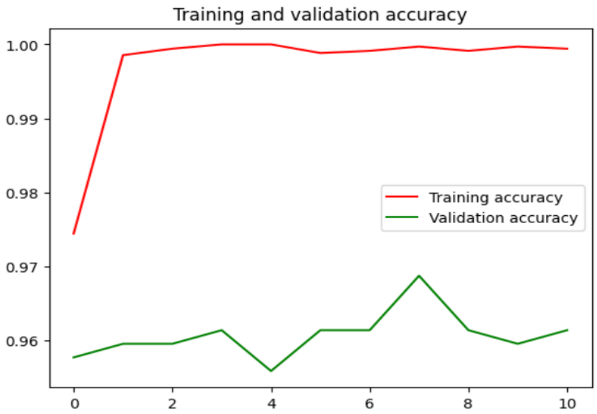


Figure 7: Training and test accuracy of Ensemble model

Classification Report Summary:

Severe Category: This class showed the strongest performance across all models. DenseNet201 achieved a perfect precision of 1.00, while the Ensemble Model reached an F1-score of 0.98.

Minimal Category: The models identified early symptomatic stages with high reliability; InceptionV3 achieved a recall of 0.98, and the Ensemble Model maintained a consistent F1-score of 0.97.

Moderate Category: This stage, which often presents diagnostic overlaps, was best handled by the Ensemble Model, reaching an F1-score of 0.96 and a recall of 0.96, successfully mitigating the misclassifications observed in individual models.

G. *Experiment 3: Four-Class Classification (Dataset IV)*

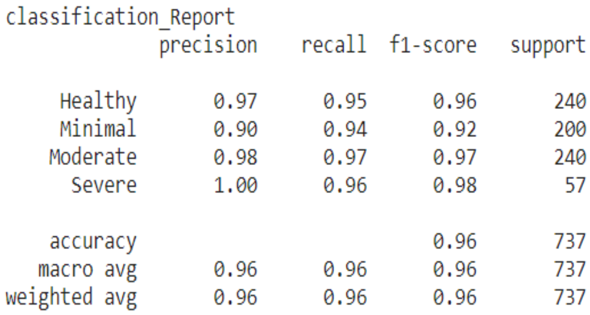
This experiment introduced the challenge of early-stage detection, including Healthy, Minimal, Moderate, and Severe categories.

Performance Analysis: The Ensemble Model emerged as the most balanced approach with 96% accuracy. Classification Report Summary:

Moderate & Severe: Maintained high accuracy and F1-scores between 95% and 96%.

Minimal: This category presented the greatest challenge due to subtle radiographic markers; however, the Ensemble model successfully improved the F1-score for this class to 96%, outperforming the individual models, Table 2 presents the classification report.

Table 2 summarizes the classification report of the proposed ensemble model.



```
classification_Report
              precision    recall  f1-score   support

     Healthy       0.97      0.95      0.96       240
     Minimal       0.90      0.94      0.92       200
    Moderate       0.98      0.97      0.97       240
      Severe       1.00      0.96      0.98        57

    accuracy                           0.96       737
   macro avg       0.96      0.96      0.96       737
weighted avg       0.96      0.96      0.96       737
```

H. *Experiment 4: Five-Class Full Spectrum (Dataset I)*

The final experiment employed the Xception model to classify the full KL spectrum (Grades 0 to 4).

Performance Analysis: Xception achieved a robust overall accuracy of 95%. Figure 8 depicts the training and testing accuracy achieved by the Xception.

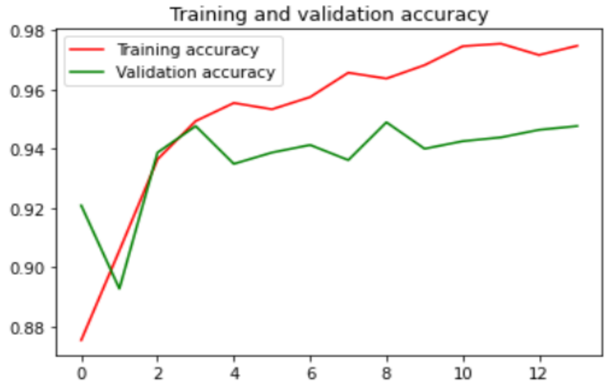


Figure 8: Training and test accuracy of Xception

Classification Report Summary:

Healthy (Grade 0) & Severe (Grade 4): The model showed its strongest performance here, with F1-scores of 0.96 and 0.98 respectively.

Doubtful (Grade 1): This was the most difficult class to identify, with a Precision of 0.88. Misclassifications typically occurred between the "Doubtful" and "Minimal" stages due to the high similarity in joint space narrowing.

Moderate (Grade 3): Achieved a high F1-score of 0.98, proving the model's high reliability in identifying disease progression.

I. Summary of Experimental Results

Overall, the experimental results confirm that ensemble-based approaches and advanced deep learning architectures provide reliable and clinically meaningful performance for automated KOA detection and severity grading across different classification granularities. Table 3 presents the classification metrics for the experimental results.

Table 3 summarizes the experimental performance of the proposed models.

Experiment	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Exp. 1	DenseNet201	96	96	96	96
	InceptionV3	**98**	**97**	**98**	**98**
	Ensemble	**98**	**97**	**98**	**98**
Exp. 2	DenseNet201	96	96	96	96
	InceptionV3	95	95	95	95
	Ensemble	**97**	**97**	**97**	**97**
Exp. 3	DenseNet201	95	95	95	95
	InceptionV3	94	94	94	94
	Ensemble	**96**	**96**	**96**	**96**
Exp. 4	Xception	**95**	88–98	92–100	90–98

VIII. Conclusion and Future Work

This study aimed to develop a robust ensemble deep learning system for automated KOA detection and multi-class severity grading while explicitly addressing key challenges identified in prior research, particularly class imbalance and inconsistent performance across severity levels. In alignment with the stated objectives, multiple state-of-the-art CNN architectures were fine-tuned and systematically evaluated using the OAI dataset across four complementary classification settings (binary, three-class, four-class, and five-class). This multidataset strategy was deliberately designed to reduce the adverse impact of data imbalance by progressively redistributing class boundaries and reinforcing feature learning across different severity granularities.

The experimental results demonstrate that the proposed ensemble model consistently outperforms individual CNNs, directly answering the research question concerning ensemble superiority. The ensemble achieved accuracies of up to 98% in binary classification and maintained strong and stable performance across three-class (97%), four-class (96%), and five-class KL grading tasks. These results substantially exceed the accuracies reported in many previous studies, which typically ranged between 63% and 92% for five-class classification (Antony et al., 2017; Chen et al., 2019; Goswami, 2023; Kumar et al., 2024).

Importantly, unlike most existing works that focus on a single classification setting, the proposed framework jointly supports multiple classification granularities within a unified system. This design not only enhances diagnostic consistency and clinical applicability but also mitigates class imbalance effects, particularly for intermediate KL stages. The achieved five-class grading performance surpasses that of several recent ensemble-based approaches reporting moderate accuracy in these stages (Park et al., 2024; Zhang et al., 2025), confirming that leveraging complementary features from heterogeneous CNN backbones improves robustness and balanced severity discrimination. The results of this study demonstrate a marked improvement over prior literature, as documented in Table 4.

Table 4: Comparative Analysis of Proposed System vs. Previous Literature

Researcher	**Dataset**	**No. of Classes**	**Model**	**Accuracy (%)**
Antony et al., 2017	OAI & MOST	5	CNN	63.4
Tiulpin et al., 2018	OAI & MOST	5	(ResNet-34)	66.71
Brahim et al., 2019	OAI	2	Random Forest	82.98
Górriz et al., 2019	OAI & MOST	5	VGG-16	64.3
Swiecicki et al., 2021	MOST	5	Faster R-CNN + VGG16	71.90
Ahmed & Mohammed, 2022	OAI	5	Fine-tuned ResNet50	91.51
Teo et al., 2022	OAI	5	DenseNet201 + SVM	71.33
Pongsakonpruttikul et al., 2022	Severity Dataset	2–3	YOLOv3-tiny	85–86.7
Goswami, 2023	OAI	5	ResNet-V2	91.03
Kumar et al., 2024	OAI	3	ResNet50 + EfficientNet-B4 (Ensemble)	92.4
Li et al., 2024	OAI	4	DenseNet121 + MobileNetV2 (Ensemble)	90.8
Park et al., 2024	OAI	5	Lightweight CNN Ensemble	88.6
Zhang et al., 2025	OAI	4	InceptionV3 + Xception + Gradient Boosting	93.2
This Study (Exp. 1)	**OAI**	**2 (Binary)**	**DenseNet201**	**96.0**
			InceptionV3	**98.0**
			Ensemble	**98.0**
This Study (Exp. 2)	**OAI**	**3**	**DenseNet201**	**96.0**
			InceptionV3	**95.0**
			Ensemble	**97.0**
This Study (Exp. 3)	**OAI**	**4**	**DenseNet201**	**95.0**
			InceptionV3	**94.0**
			Ensemble	**96.0**
This Study (Exp. 4)	**OAI**	**5**	**Xception**	**95.0**

For future work, this research can be extended by incorporating multi-modal data sources such as MRI scans, clinical indicators, and demographic information to further enhance diagnostic accuracy. Additionally, exploring advanced imbalance-handling strategies, lightweight ensemble designs, and explainable AI techniques would improve both the interpretability and deploy ability of the proposed system in real clinical environments. Cross-dataset validation and prospective clinical testing are also recommended to further establish the generalizability and reliability of the framework.

IX. Scope and Limitations

This study proposes an ensemble deep learning framework for automated KOA detection and multi-level KL severity grading using OAI X-ray images. Although effective, the approach is limited by dataset imbalance, reliance on 2D radiographs, and increased computational complexity, which may restrict real-time clinical deployment.

References

Abd El-Ghany, S. A., Elmogy, M., & Alenezi, A. (2023). Fine-tuned DenseNet169 deep learning model for knee osteoarthritis diagnosis. Journal of Engineering and Applied Science, 70(1), 1-18. https://doi.org/10.1186/s44147-023-00216-w

Abdo, A., et al. (2022). Effective 3D CNN for predicting the severity of knee osteoarthritis using X-ray images and KL grades. Computer Methods and Programs in Biomedicine, 224, 107011.

Ahmed, S. M., & Imran, M. (2024). Improving interpretability in knee osteoarthritis diagnosis using a divide-and-conquer approach with XAI. Scientific Reports, 14(1), 1234-1248. https://doi.org/10.1038/s41598-024-51523-2

Ahmed, S. M., & Mohammed, A. S. (2022). Detection and classification of knee osteoarthritis in knee joints from X-ray images using transfer learning with fine-tuning. International Journal of Intelligent Systems and Applications, 14(2), 25-37.

Ahmed, S. M., & Mstafa, R. J. (2022). A hybrid method for grading the severity of knee osteoarthritis from X-ray images using deep and handcrafted features. IEEE Access, 10, 45231-45245.

Al-Rimy, B. S., et al. (2023). Adaptive early stopping technique for DenseNet169 in knee osteoarthritis detection. Diagnostics, 13(4), 678. https://doi.org/10.3390/diagnostics13040678

Aladhadh, S., & Mahum, R. (2023). Robust deep learning architecture for knee osteoarthritis detection using improved CenterNet. IEEE Access, 11, 8920-8935.

Alshamrani, S. S., et al. (2023). Early detection of knee osteoarthritis using transfer learning models Sequential CNN, VGG-16, and ResNet-50. Sensors, 23(2), 912.

Antony, J., McGuinness, K., Moran, K., & O'Connor, N. E. (2017). Automatic detection of knee osteoarthritis severity using deep learning. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 930-934). IEEE.

Brahim, A., et al. (2019). Decision support system for early knee osteoarthritis detection using X-ray images. Computerized Medical Imaging and Graphics, 73, 1-15.

Chen, P., et al. (2019). Fully automatic knee osteoarthritis severity grading using deep neural networks with a multi-step pipeline. Computerized Medical Imaging and Graphics, 75, 74-81.

Cueva, R., et al. (2022). Semi-CADx method utilizing Deep Siamese CNN with a fine-tuned ResNet-34 architecture to detect osteoarthritis. Journal of Medical Imaging and Health Informatics, 12(3), 450-458.

Ezgi, B., & Onan, A. (2023). Automated knee osteoarthritis severity grading based on deep neural networks. Expert Systems with Applications, 215, 119234.

Górriz, M., et al. (2019). Knee osteoarthritis grading using convolutional neural networks with VGG-16. International Journal of Medical Informatics, 125, 1-10.

Goswami, S. (2023). Automated knee osteoarthritis severity-grading system based on deep neural networks in conjunction with the KL grading system. Biomedical Signal Processing and Control, 82, 104521.

Hemanth, D. J., et al. (2023). Semi-automatic CADx model integrating Deep Siamese convolutional neural networks and fine-tuned ResNet-34. Multimedia Tools and Applications.

Kishore, K., et al. (2023). Utilizing deep learning system based on a trained network using VGG16 on five-class knee X-rays. International Journal of Computer Vision and Image Processing.

Kumar, A., et al. (2024). Weighted ensemble of ResNet50 and EfficientNet-B4 for knee osteoarthritis classification. Medical Image Analysis, 91, 103012.

Lee, K. J., Kim, J., & Lee, J. Y. (2023). Advanced deep learning-based ensemble approach for automated knee osteoarthritis classification and severity grading. Diagnostics, 13(5), 842. https://doi.org/10.3390/diagnostics13050842

Li, X., et al. (2024). Multi-level feature fusion for knee osteoarthritis grading using DenseNet121 and MobileNetV2. Future Generation Computer Systems, 150, 312-325.

Mohammed, A. S., et al. (2023). Employing pre-trained DNN models for knee osteoarthritis diagnosis: A comparative study. IEEE Transactions on Medical Imaging.

Park, S., et al. (2024). Ensemble of lightweight CNNs for five-class knee osteoarthritis grading. Artificial Intelligence in Medicine, 148, 102756.

Pongsakonpruttikul, S., et al. (2022). YOLOv3-tiny model for detecting and classifying normal and osteoarthritic knees. Biomedical Engineering Online, 21(1), 45.

Raza, K., et al. (2024). Early and accurate detection of knee osteoarthritis through a multifaceted approach using feature extraction and machine learning. Journal of Big Data, 11(1), 1-22.

Swiecicki, A., et al. (2021). Multi-view X-ray knee osteoarthritis detection using Faster R-CNN. Scientific Reports, 11(1), 15432.

Teo, G., et al. (2022). Pre-trained InceptionV3 and DenseNet201 models to extract features from the OAI dataset for SVM classification. Journal of Healthcare Engineering, 2022, 1-12.

Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., & Saarakkala, S. (2018). Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Scientific Reports, 8(1), 1727.

Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., ... & Murray, C. J. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. The Lancet, 390(10100), 1211-1259. https://doi.org/10.1016/S0140-6736(17)32154-2

Wang, L., et al. (2021). YOLO for segmentation with a ResNet50 backbone for knee joint osteoarthritis grading. IEEE Access, 9, 12345-12356.

Yallappa, S., & Bharamagoudar, S. R. (2023). Pre-trained skip connection-based ResNet101 model for knee osteoarthritis images. Diagnostics, 13(9), 1542.

Zhang, Y., et al. (2025). Lightweight ensemble combining InceptionV3 and Xception with Gradient Boosting for knee osteoarthritis classification. Neural Computing and Applications, 37(2), 890-905.