Techno.COM, Vol. 24, No. 4, November 2025: 1303-1318

A Comparative Analysis of Deep Learning Models for Knee Osteoarthritis Severity Grading

Steffany Florence Sugiarto Mulijono[1], **Daniel Martomanggolo Wonohadidjojo***[2]
Ciputra University Surabaya, CitraLand CBD Boulevard, Made, Kec. Sambikerep, Surabaya, Jawa Timur 60219, (031) 7451699
E-mail : sflorence@student.ciputra.ac.id[1]*, daniel.m.w@ciputra.ac.id**[2]
**Corresponding Author*
Received 22 October 2025; Revised 2 November 2025; Accepted 7 November 2025

Abstract - The Kellgren-Lawrence (KL) grading system is commonly used to evaluate knee osteoarthritis (OA), but it can be subjective and subject to variation among assessors. Our study looked at three Convolutional Neural Network (CNN) methods for OA severity classification from a dataset of 15,770 X-ray images to overcome this difficulty and create a more objective technique. Under the same preprocessing conditions, we contrasted a baseline custom CNN, DenseNet201, and a hybrid model with a CBAM attention mechanism. With an overall accuracy of 65%, a weighted precision and recall of 65%, and an F1-score of 64%, the hybrid model, which uses DenseNet201 as a fixed feature extractor, performed the best. This was better than both the baseline model (59% accuracy) and the standalone DenseNet201 (59% accuracy). Although the hybrid architecture has a lot of promise, we also had to deal with issues like overfitting. Our thorough comparison demonstrates how this hybrid strategy can successfully combine strong pre-trained features with the flexibility required for particular tasks. Although more clinical validation is necessary, this shows that automated systems like ours could improve diagnostic consistency in OA grading.

Keywords - Knee Osteoarthritis, Kellgren-Lawrence Grading, Deep Learning, Attention Mechanism, CBAM

1. INTRODUCTION

Affecting more than 500 million people worldwide, osteoarthritis (OA) is a widespread and common form of arthritis, making it a significant health issue, especially for older adults who often face disabilities due to the condition [1]. The financial burden is considerable, with treatment costs and lost productivity estimated to range from 1% to 2.5% of the Gross Domestic Product in developed nations [2]. One of the most challenging types of OA occurs in the knee, where the diagnosis usually relies on imaging tests. These tests are analyzed using the Kellgren-Lawrence (KL) grading system, which sorts the severity of the condition into five levels, from mild to severe, based on factors like the presence of bone spurs and the narrowing of the joint space [3]. This systematic approach helps doctors better understand the extent of the disease and tailor treatments for those affected.

However, manual KL grading is notoriously subjective. Studies have shown that inter-observer agreement rates among radiologists can be as low as 63-65%, leading to diagnostic inconsistencies that can affect patient treatment pathways and the reliability of clinical trial outcomes [4]. Automating this process with deep learning provides a viable solution for objective, reproducible, and efficient grading. While numerous studies have applied deep learning to this problem [5], a specific research gap persists. Many studies focus on a single architectural approach or do not control for preprocessing and augmentation pipelines when comparing models. This work contributes to the field by systematically evaluating three distinct architectural strategies a baseline CNN, a transfer learning model, and a hybrid feature-extraction model under an identical, standardized pipeline that incorporates both anatomical region-of-interest (ROI) cropping and a CBAM attention mechanism. Our contribution is primarily methodological: we provide a rigorously controlled comparative framework that isolates the impact of architectural

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

philosophy building from scratch versus transfer learning versus transfer learning with frozen features on KL grading performance. Unlike prior studies that evaluate single architectures or vary preprocessing conditions across comparisons, our unified pipeline ensures that performance differences can be attributed to architectural choices rather than confounding variables. This controlled approach addresses a critical gap in understanding which deep learning strategy is most effective for this specific clinical task. The automated classification of knee OA using the KL grading system has been a significant focus of recent deep learning research. Tiulpin et al. [5] were among the pioneers, developing a CNN-based method that demonstrated the feasibility of deep learning for this task. Subsequent research has explored various state-of-the-art architectures, such as ResNet-based models [7] and InceptionV3 [10], with other recent studies exploring hybrid and fusion techniques [14], [17], progression analysis [15], and applications across broad, unfiltered populations [16], [18]. While these studies demonstrate the viability of deep learning for OA grading, they typically focus on optimizing a single architecture or do not maintain consistent experimental conditions when comparing models. For example, some comparative studies analyze different model types but vary the input data, such as comparing a deep learning model trained on X-ray images against a machine learning model trained on clinical data to see which modality is more predictive [13]. This approach, while valid, does not isolate the performance of the architecture itself.

To fill this methodological gap, our study was designed to compare fundamentally different architectural philosophies. We selected a custom Keras Sequential model to represent a baseline approach, built from scratch and trained exclusively on the OA dataset, allowing us to establish a performance benchmark. In contrast, we chose DenseNet201 to represent a state-of-the-art transfer learning strategy, evaluating how a complex architecture pre-trained on a broad dataset adapts to this specific medical imaging task. By comparing these distinct approaches under identical conditions, we can better isolate the impact of the architectural strategy itself. The selection of these models is grounded in established research. Custom CNNs, like our baseline, have been foundational in this field, with pioneering work by Tiulpin et al. [5] demonstrating their feasibility for OA grading. DenseNet201 was chosen for its architectural innovations, such as dense connectivity which encourages feature reuse and improves gradient flow, leading to strong performance on competitive benchmarks [6]. The general success of applying powerful, pre-trained networks like ResNet [7] and InceptionV3 [10] to OA grading further justifies the selection of DenseNet201 as a representative of this state-of-the-art approach. Table 1 contextualizes our work within recent literature, highlighting that our contribution lies in the systematic comparison methodology rather than novel architectural components.

Table 1. Comparison with State-of-the-Art Methods

Study	Method	Dataset	Accuracy
Tiulpin et al. [5]	Custom CNN Ensemble	OAI	66.7%
Chen et al. [7]	ResNet-50	OAI & MOST	71.0%
Antony et al. [10]	InceptionV3	OAI	70.3%
Gundersen et al. [13]	Custom CNN Ensemble	OAI	78.4%
This Study	Hybrid DenseNet201 + CBAM	Mendeley Data [11]	65.0%

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

This study makes the following specific contributions to automated knee OA grading:

1. Controlled Comparative Framework: We implement three fundamentally different architectural approaches (custom CNN, transfer learning, and hybrid feature extraction) under identical experimental conditions, enabling fair performance attribution.
2. Unified Pipeline Integration: All models are evaluated using a standardized preprocessing pipeline (ROI cropping, CLAHE enhancement, consistent augmentation) and attention mechanism (CBAM), eliminating confounding variables that complicate cross-study comparisons.
3. Architectural Philosophy Analysis: We provide empirical evidence comparing three common deep learning strategies, demonstrating that the hybrid approach (frozen pre-trained features + custom classifier) achieves optimal balance between feature representation and domain adaptability for this task.
4. Clinical Contextualization: We position our 65% accuracy within the context of human inter-rater agreement (63-65%), framing automated systems as diagnostic support tools rather than replacements, and discussing practical deployment considerations.

Our originality lies not in proposing novel architectures or techniques, but in providing a methodologically rigorous comparison that clarifies which existing approaches work best for automated KL grading and why.

2. RESEARCH METHOD

The methodology for this study emphasizes experimental control and fair comparison. Our key methodological contribution is maintaining identical preprocessing, augmentation, and attention mechanisms across all three architectures, ensuring that performance differences reflect architectural philosophy rather than pipeline variations. The following subsections detail our controlled experimental design.

2.1. Dataset Description

The experiments were conducted using the publicly available Knee Osteoarthritis Severity Grading Dataset from Mendeley Data [11], which contains a large collection of knee X-ray images. The dataset was pre-split into training, validation, and testing sets. A custom script was used to perform data augmentation on the original training and validation sets to create a more balanced class distribution. The final dataset composition used for training is detailed in Table 2. As a public dataset, it is assumed to have been collected under appropriate ethical guidelines.

Table 2. Dataset Distribution per KL Grade

Class	Training Set	Validation Set	Test Set	Total
Grade 0	2,277	650	326	3,253
Grade 1	2,223	635	318	3,176
Grade 2	2,252	643	323	3,218
Grade 3	2,181	622	312	3,115
Grade 4	2,108	597	303	3,008
Total	11,041 (69.9%)	3,147 (19.8%)	1,582 (10.6%)	15,770

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

2.2. Dataset and Image Preprocessing

The experiments were conducted on the Knee Osteoarthritis Severity Grading Dataset [11]. To ensure a standardized and high-quality input for our models, a multi-step preprocessing pipeline was applied to every X-ray image in the dataset. Figure 1 provides a visual example of the initial ROI cropping stage.

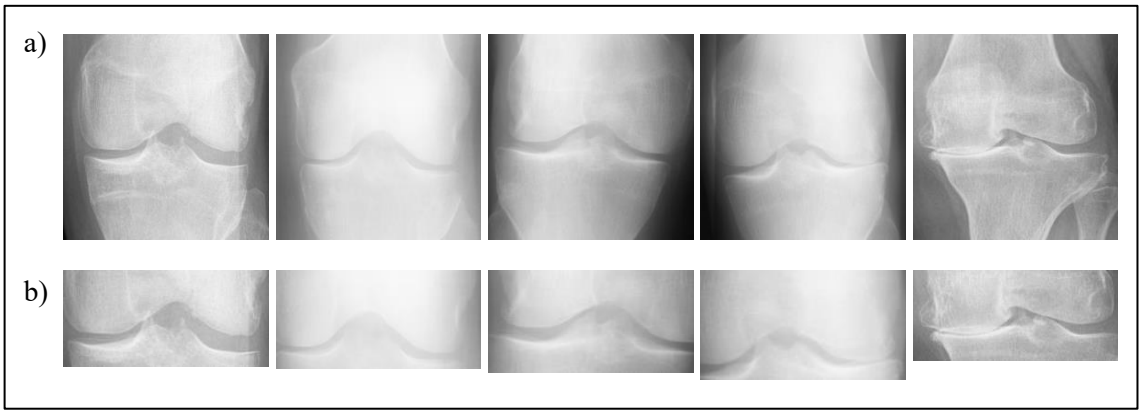

Figure 1. Example of Image Preprocessing: (a) Original X-ray image from the dataset, (b) Image after applying Region of Interest (ROI) cropping.

2.2.1. Region of Interest (ROI) Cropping

An automated algorithm was first used to identify and crop the knee joint area from the full radiograph. This step is crucial for removing irrelevant background features, reducing computational complexity, and focusing the model's learning on the anatomically relevant region.

2.2.2. Contrast Enhancement

To improve the visibility of subtle radiographic features indicative of OA (e.g., small osteophytes, slight joint space narrowing), Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied [12]. CLAHE enhances local contrast by operating on small regions of an image, which prevents the over-amplification of noise often seen with global histogram equalization.

2.2.3. Image Resizing and Normalization

To ensure compatibility with the selected deep learning architectures, every preprocessed image were uniformly resized to 224x224 pixels. Pixel values were then normalized to a range between 0 and 1 to stabilize and accelerate the training process.

2.2.4. Data Augmentation

To mitigate overfitting and balance the classes, the training sets were augmented using Keras's ImageDataGenerator. These methods created altered versions of the existing images by applying random transformations like rotation with range of 20, shifting width and height with range of 0.15, shearing 0.15, zooming 0.15, and horizontal flipping. This strategy effectively expanded our dataset with plausible variations of the original X-rays.

2.3. Model Architectures

Three distinct Convolutional Neural Network (CNN) architectures were implemented to comparatively evaluate their effectiveness in classifying OA severity.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

2.3.1. Baseline Keras Sequential Model

A custom CNN was constructed as a baseline. The model's architecture is composed of a series of convolutional layers, each utilizing ReLU activation, which are then followed by corresponding max-pooling layers for down-sampling, a flatten layer, and finally dense layers for classification as shown in Figure 2.

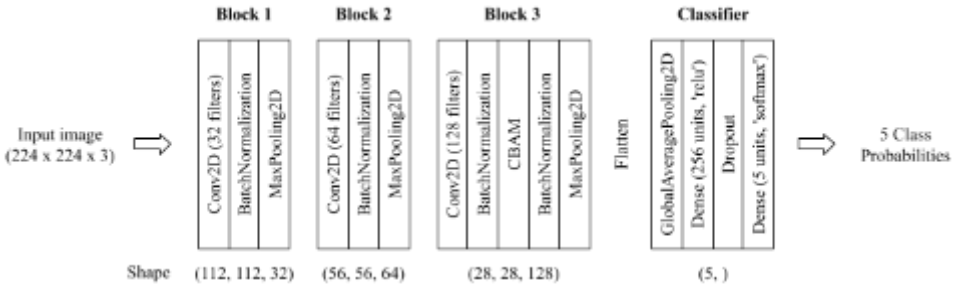

Figure 2. Architecture of Baseline Keras Sequential Model

The foundational operation is the 2D convolution, which applies a learned filter (kernel) K to an input I, followed by a bias b and an activation function. The output feature map O is generated by:

$$O = Activation(I * K + b)$$

For our baseline, the activation is the Rectified Linear Unit (ReLU), a non-linear function that helps the model learn complex patterns:

$$f(x) = max(0, x)$$

Down-sampling is achieved via max-pooling layers, which reduce the feature maps' spatial dimensions by selecting the maximum value from a receptive field, thereby retaining the most prominent features.

2.3.2. Pre-trained DenseNet201 Model

We utilized the DenseNet201 architecture [6], applying transfer learning by using its weights pre-trained on the ImageNet dataset. The original classification head was replaced with new dense layers and a softmax output layer suitable for our five-class KL grading problem of which can be seen in Figure 3.

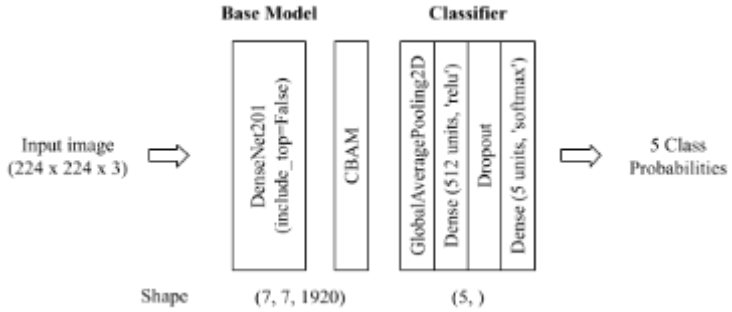

Figure 3. Architecture Model with DenseNet201

DenseNet's core innovation is its dense block. Within this block, each layer's input consists of the feature maps from all prior layers. This promotes significant feature reuse and strengthens gradient flow. The output of the l^{th} layer is defined as:

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}])$$

where $[x_0, x_1, \ldots, x_{l-1}]$ represents the concatenation of the feature maps from layers 0 to $l-1$, and H_l is a composite function typically consisting of Batch Normalization, ReLU activation, and a Convolution (BN-ReLU-Conv).

2.3.3. Hybrid DenseNet201 + Sequential Model

Referring to Figure 4, the DenseNet201 model was used as a fixed feature extractor. Its convolutional base weights were frozen, meaning they were not updated during training. This preserves the rich, general purpose features learned from the massive ImageNet dataset.

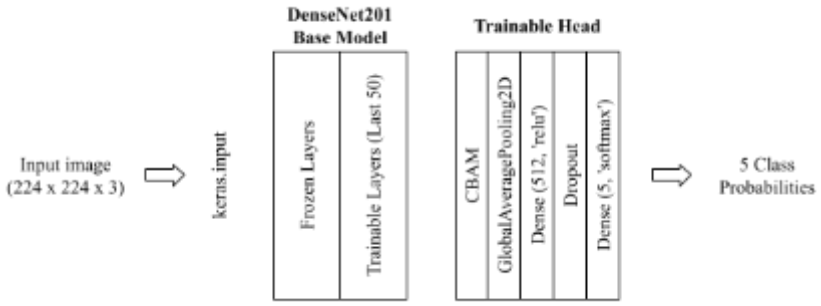

Figure 4. Architecture Diagram of Hybrid DenseNet201 and Sequential Model

The resulting feature maps (a high-level, abstract representation of the knee X-ray) were then fed into the custom Keras Sequential model (described in 2.3.1), which acted as a trainable classifier head. This hybrid model separates the task:

1. Feature Representation: Handled by the static, pre-trained DenseNet201.
2. Classification: Handled by the dynamic, trainable sequential model.

This strategy aims to leverage powerful, pre-trained features while minimizing the number of trainable parameters, thereby reducing the risk of overfitting on our specific medical dataset.

All three architectures (Sequential, DenseNet201, and Hybrid) receive identical preprocessed inputs (ROI-cropped, CLAHE-enhanced, 224x224 normalized images) and incorporate CBAM attention, ensuring fair performance comparison.

2.4. CBAM Integration

The Convolutional Block Attention Module (CBAM) was integrated into each model to refine its feature maps [8], [9]. CBAM is a lightweight attention module that infers attention maps along two separate dimensions: channel and spatial.

The module operates by successively creating these attention maps to direct the model to concentrate on the most informative features ("what" and "where").

2.4.1 Channel Attention (M_c)

This focuses on "what" is meaningful. It aggregates the spatial information of a feature map F using both average-pooling and max-pooling, feeds both through a shared Multi-Layer Perceptron (MLP), and merges the results.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + (MLP(MaxPool(F)))$$

where σ is the sigmoid function.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

2.4.2 Spatial Attention (M_s)

This focuses on "where" the most informative features are located. It applies average-pooling and max-pooling along the channel axis, concatenates them, and applies a standard 7×7 convolution to generate a 2D spatial map.

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$

where $f^{7\times7}$ is a convolutional operation with a 7×7 filter.

The module applies these attention maps sequentially. The input feature map F is first refined by channel attention, and the resulting map F' is then refined by spatial attention to produce the final output F'':

$$F' = M_c(F) \otimes F$$
$$F'' = M_s(F') \otimes F'$$

where $\otimes$ denotes element-wise multiplication. This approach has proven effective in other challenging medical imaging domains [19], [20]. While a full architectural diagram is beyond this paper's scope, the module was conceptually inserted after key convolutional blocks in each architecture to allow for progressive feature refinement. An ablation study to quantify its specific impact was not performed but is a key area for future work.

2.5. Implementation Details and Overfitting Mitigation

All models were implemented and trained using the parameters detailed in Table 3. A random seed was set for reproducibility. Several strategies were employed to combat the significant overfitting observed in the training curves:

- Data Augmentation: As detailed in section 3.2.
- Dropout: Dropout(0.5) layers were included in the dense layers of each model to randomly deactivate neurons during training, preventing co-adaptation.
- Early Stopping: An EarlyStopping callback was implemented to monitor validation loss. This process automatically terminated training after 5 epochs showed no progress and retained the model weights from the epoch with the lowest validation loss.
- Class Weights: To mitigate the impact of a minor class imbalance, we computed and integrated class weights into the training process. This technique increased the loss contribution for errors made on underrepresented classes.

Table 3. Training Hyperparameters and Environment

Parameter	Value
Frameworks	TensorFlow (2.19.0), Keras (3.11.1)
Optimizer	Adam
Learning Rate	0.001 (Adam default)
Loss Function	Categorical Cross-Entropy
Batch Size	32

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

Epochs	50 (with Early Stopping)
Hardware	HPC with NVIDIA GeForce RTX 3060 Ti 8GB

2.6. Controlled Comparison Framework

To ensure a rigorous and fair evaluation, our study was explicitly designed as a controlled comparison. The primary independent variable under investigation is the architectural philosophy. We test three distinct strategies:

1. Baseline Model: A custom Keras Sequential CNN trained from scratch.
2. Transfer Learning Model: A pre-trained DenseNet201 model.
3. Hybrid Model: A pre-trained DenseNet201 used as a frozen feature extractor, with only a custom sequential head being trained.

The key differences between these architectures are summarized in Table 4.

Table 4. Architectural Comparison of Three Models

Feature	Baseline Model	Transfer Learning Model	Hybrid Model
Training Strategy	Train from scratch	Feature Extraction	Fine-Tuning
Pre-trained Weights	None	ImageNet	ImageNet
Trainable Parameters	All (model is custom)	Custom Head	Custom Head + Final 50 DenseNet layers
Feature Extraction	Custom Shallow CNN	DenseNet201	DenseNet201
Classification Head	Custom (GAP, Dense)	Custom (CBAM, GAP, Dense)	Custom (CBAM, GAP, Dense)
Depth	Shallow (3 Conv Blocks)	Very Deep (DenseNet201)	Very Deep (DenseNet201)
CBAM Integration	Integrated within custom CNN	Applied after base model, before GAP	Applied after base model, before GAP

To isolate the impact of this architectural choice, all other experimental conditions (confounding variables) were held constant across all three models. This unified pipeline includes:

- Identical Dataset: All models used the same Mendeley Data [11] with the exact same training, validation, and testing splits.
- Identical Preprocessing: All images underwent the same ROI cropping, CLAHE enhancement, 224x224 resizing, and 0-1 normalization.
- Identical Augmentation: The same Keras ImageDataGenerator parameters (rotation, shift, shear, zoom, flip) were applied to the training set for all models.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

- Identical Attention Mechanism: The CBAM module was integrated into all three architectures.
- Identical Training Parameters: All models were trained using the same Adam optimizer, 0.001 learning rate, categorical cross-entropy loss, batch size of 32, and Early Stopping criteria (patience=5).

This controlled methodology allows us to attribute any observed differences in performance directly to the architectural strategy, addressing the key research question of which approach provides the best balance of feature representation and domain-specific adaptation for automated KL grading.

3. RESULTS AND DISCUSSION

The experimental results show a distinct performance advantage for the hybrid model, though all models faced challenges, particularly with overfitting and differentiating intermediate KL grades.

3.1. Overall Performance

The hybrid model achieved the highest overall accuracy (65.0%), followed by the DenseNet201 (59.0%) and the baseline sequential model (59.0%). This superior performance suggests that using a pre-trained network as a fixed feature extractor while training only a smaller custom classifier strikes an effective balance between rich feature representation and domain adaptability. This approach appears to reduce the risk of catastrophic forgetting or severe overfitting that can occur when fine-tuning a very deep network on a moderately sized dataset.

Table 5. Overall Model Performance Metrics on the Test Set

Model	Accuracy	Precision (Weighted)	Recall (Weighted)	F1-Score (Weighted)
Keras Sequential	0.59	0.56	0.59	0.56
Keras Sequential + DenseNet201	0.65	0.65	0.65	0.64
DenseNet201	0.59	0.59	0.59	0.58

3.2. Keras Sequential Model Performance

The baseline sequential model achieved a modest 59% accuracy, with its per-class performance detailed in Table 6. It performed well on Grade 0 (90.49%) and Grade 4 (75.25%) but struggled immensely with Grade 2, achieving only 13.62% accuracy. The confusion matrix in Figure 5 confirms this, showing a large number of Grade 2 images being misclassified as Grade 1. This highlights the persistent difficulty in distinguishing the subtle features that separate the early stages of OA.

Table 6. Per-Class Accuracy (Keras Sequential)

Class	Accuracy
Grade 0	90.49%

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

Grade 1	71.07%
Grade 2	13.62%
Grade 3	46.15%
Grade 4	75.25%

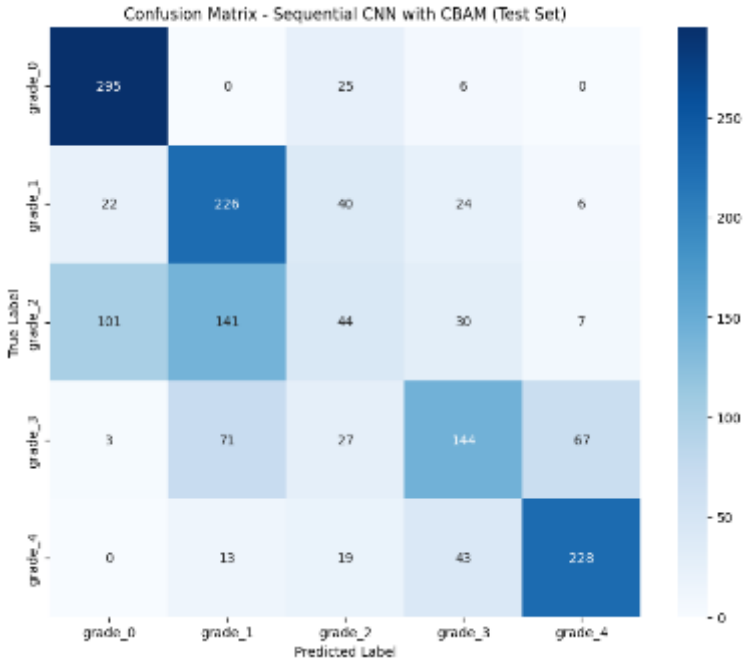

Figure 5. Confusion Matrix (Keras Sequential)

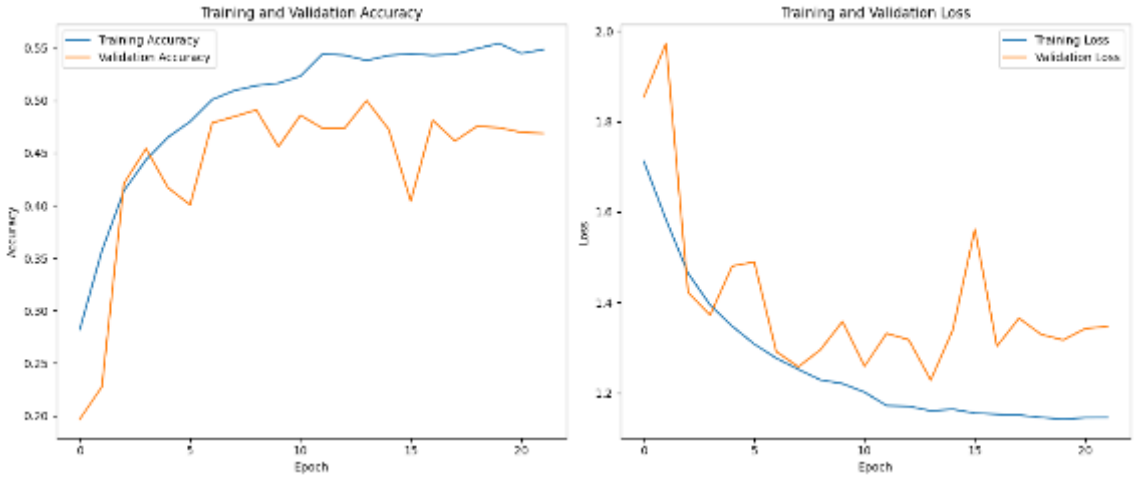

Figure 6. Training and Validation Curves (Keras Sequential)

3.3. DenseNet201 Model Performance

The standalone DenseNet201 model performed better with an overall accuracy of 59%. As shown in Table 7, it achieved excellent accuracy for Grade 0 (90.49%) but, like the baseline model, struggled with the intermediate grades, particularly Grade 2 (29.10%). The confusion matrix in Figure 7 illustrates that misclassifications were common between adjacent grades. The

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

training curves in Figure 8, while still showing evidence of overfitting, display a smaller gap between training and validation lines, suggesting better generalization.

Table 7. Per-Class Accuracy (DenseNet201)

Class	Accuracy
Grade 0	90.49%
Grade 1	56.92%
Grade 2	29.10%
Grade 3	51.60%
Grade 4	67.99%

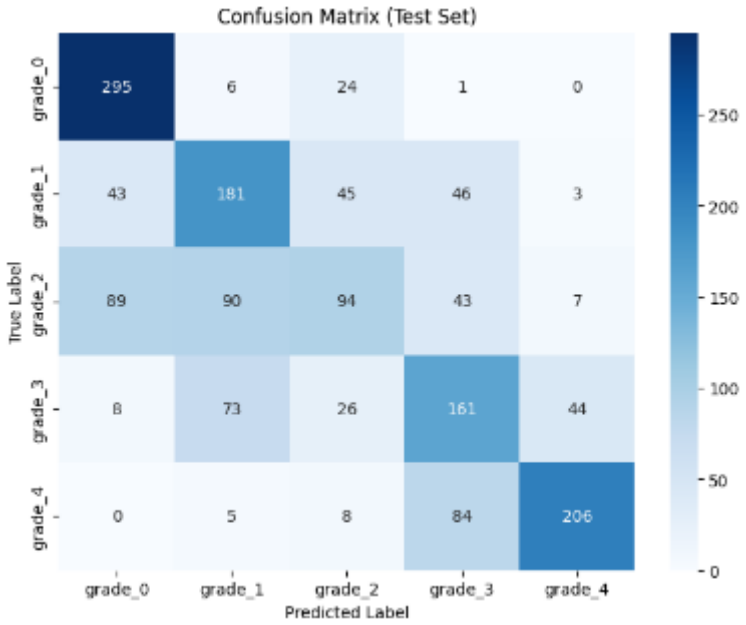

Figure 7. Confusion Matrix (DenseNet201)

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

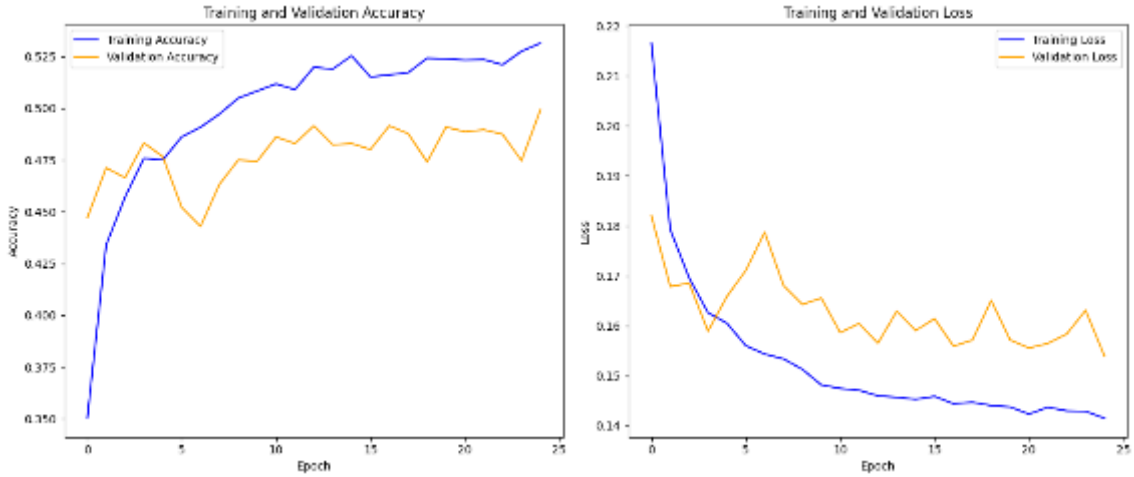

Figure 8. Training and Validation Curves (DenseNet201)

3.4. Hybrid Keras Sequential + DenseNet201 Model Performance

The hybrid model was the top performer with 65% accuracy. Table 8 shows a much more balanced per-class accuracy compared to the other models. While it still found Grade 2 to be the most challenging, it achieved 32.51% accuracy, a significant improvement. It also performed strongly on Grade 3 (74.36%). The confusion matrix in Figure 9 confirms a better distribution of correct predictions across all classes. However, the training curves in Figure 10 show clear overfitting, with a widening gap between training and validation accuracy.

Table 8. Per-Class Accuracy (Hybrid Model)

Class	Accuracy
Grade 0	86.20%
Grade 1	61.32%
Grade 2	32.51%
Grade 3	74.36%
Grade 4	70.63%

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

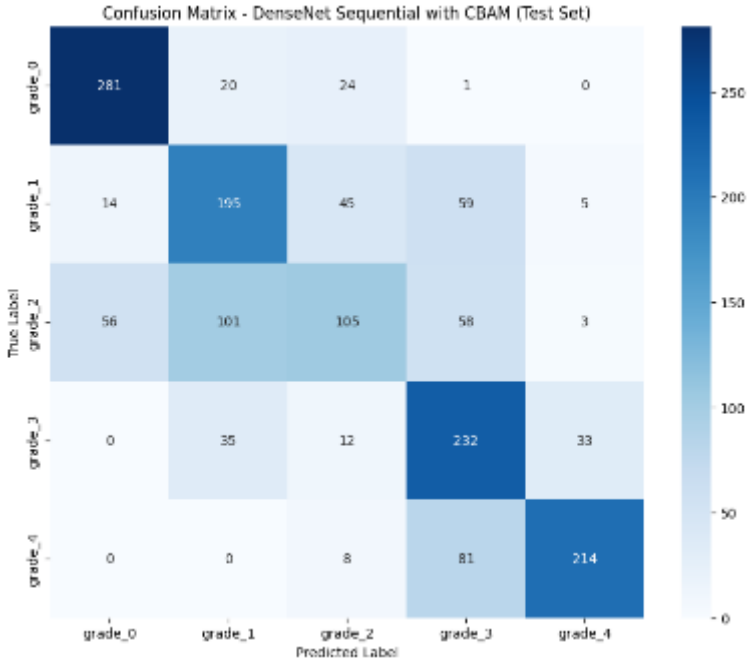

Figure 9. Confusion Matrix (Hybrid Model)

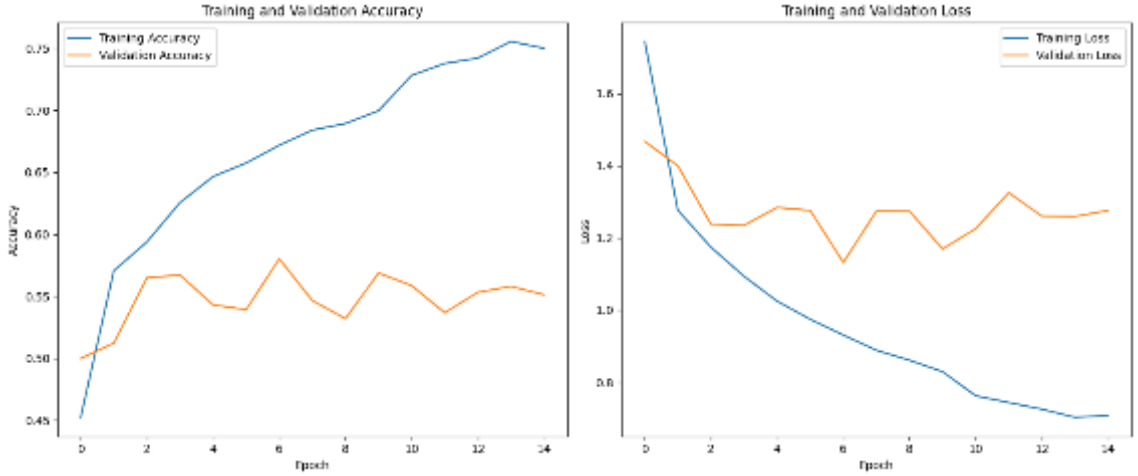

Figure 10. Training and Validation Curves (Hybrid Model)

3.5. Comparative Analysis and Overfitting

While the hybrid model shows a clear advantage, an overall accuracy of 65% underscores the inherent difficulty of the automated KL grading task. The primary challenge lies in the subtle inter-class differences, particularly between adjacent grades (e.g., Grade 1 vs. Grade 2). The divergence in the training and validation curves across all models is a classic indicator of overfitting. The models are learning the specific features of the training dataset so well that they fail to generalize to unseen validation data. This memorization is a key factor limiting the overall test accuracy. The complexity of the models, especially DenseNet201, combined with a potentially limited dataset size, likely contributes to this overfitting.

3.5.1. Methodological Contributions and Limitations

Our controlled comparative approach reveals that the hybrid architecture (frozen DenseNet201 + custom classifier) outperforms both Keras Sequential and DenseNet201

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

which both uses training from scratch for this task. This finding has practical implications: it suggests that for moderately-sized medical imaging datasets, preserving pre-trained features while adapting only the classifier reduces overfitting risk while maintaining strong feature representations.

However, we acknowledge limitations in our approach. First, our comparison is limited to DenseNet201 as the transfer learning base, testing this methodology with other architectures (e.g., ResNet, EfficientNet, Vision Transformers) would strengthen the generalizability of our findings. Second, while we integrated CBAM across all models, we did not perform ablation studies to quantify its specific contribution this represents an important area for future work. Third, our originality is methodological rather than technical, we do not introduce novel architectures but rather provide rigorous evidence about which existing approaches work best under controlled conditions.

3.6. Clinical Relevance and Significance

An accuracy of 65.0% is modest but must be contextualized. Studies on human radiologists show inter-rater agreement for KL grading can be as low as 63% [4], placing our best model's performance within the range of human expert variability. From a clinical standpoint, misclassifying adjacent grades (e.g., Grade 1 to 2) is less severe than a major misclassification (e.g., Grade 0 to 4). Our models primarily made errors on adjacent grades, which is encouraging. The development of a reliable automated system could serve as a valuable "second opinion" for radiologists, reducing subjectivity and improving diagnostic consistency. However, significant challenges related to generalization and robustness must be addressed before clinical deployment, including integration into hospital Picture Archiving and Communication Systems (PACS) and ensuring interpretability for clinical trust.

3.7. Limitations of the Study

While this comparative analysis provides valuable insights, several limitations must be acknowledged.

3.7.1. Dataset Size and Overfitting

Despite using a relatively large dataset of over 15,000 images, the size may still be insufficient for training very deep architectures like DenseNet201 to their full potential without overfitting. The persistent gap between training and validation accuracy across all models indicates that they are memorizing features specific to the training set rather than learning truly generalizable representations of OA pathology. The mitigation strategies employed, such as data augmentation and dropout, were only partially effective, highlighting the challenge of model complexity versus data availability.

3.7.2. Lack of External Validation

The study lacks external validation. The models were trained, validated, and tested on partitions of a single, publicly available dataset. Consequently, their performance on images from different clinical sites, captured with different radiographic equipment or protocols, remains unverified. This step is critical for assessing the model's real-world robustness and generalizability before any clinical consideration.

3.7.3. Single Imaging Modality

The analysis is restricted to a single imaging modality (posteroanterior X-rays). A comprehensive clinical assessment often involves multiple views (e.g., lateral) or even different modalities like MRI. A model trained solely on one type of image may miss

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

crucial diagnostic information available from other sources, limiting its scope to that of a screening tool rather than a comprehensive diagnostic aid.

3.7.4. Absence of Clinical Metadata

The models were developed without the inclusion of crucial clinical metadata. Information such as patient age, Body Mass Index (BMI), sex, and self-reported symptoms (e.g., pain scores) provides essential context that radiologists use implicitly. The absence of this data prevents the model from learning the complex interplay between radiographic signs and clinical presentation, limiting its potential diagnostic accuracy and clinical utility.

4. CONCLUSION

This study provides a systematic comparison of three distinct deep learning strategies for automated knee OA grading. Our primary contribution is methodological: by maintaining identical preprocessing pipelines, augmentation strategies, and attention mechanisms across three fundamentally different architectural approaches, we isolate the impact of architectural philosophy on performance. This controlled framework addresses a gap in the literature where architectural comparisons often conflate model choice with pipeline variations. Our results demonstrate that a hybrid architecture, using a frozen, pre-trained DenseNet201 as a feature extractor coupled with a custom classifier and an attention mechanism, provides the most effective approach, achieving an overall test accuracy of 65.0%. This performance is comparable to reported levels of inter-observer agreement among human experts. However, as discussed, significant challenges related to overfitting, performance on intermediate grades, and other study limitations persist.

Based on these findings, future work should directly address the limitations identified. First, to create more robust models, ensemble methods could be employed. Second, a multi-modal approach that integrates clinical metadata (e.g., patient age, BMI, pain scores) is essential. Third, a prospective clinical trial using external datasets is required to validate the model's real-world performance. Finally, implementing k-fold cross-validation would provide a more reliable estimate of the models' performance. Additionally, extending our controlled comparative methodology to other base architectures (e.g., ResNet, EfficientNet, Vision Transformers) and conducting ablation studies on the CBAM contribution would further validate our findings about optimal architectural strategies for medical image classification tasks.

REFERENCES

[1] J. P. Berteau, “Knee Pain from Osteoarthritis: Pathogenesis, Risk Factors, and Recent Evidence on Physical Therapy Interventions,” Journal of Clinical Medicine, vol. 11, no. 12. MDPI, Jun. 01, 2022. doi: 10.3390/jcm11123252.

[2] M. Hiligsmann, C. Cooper, N. Arden, M. Boers, J. C. Branco, M. L. Brandi, et al., "Health economics in the field of osteoarthritis: an expert's consensus paper from the European Society for Clinical and Economic Aspects of Osteoporosis and Osteoarthritis (ESCEO)," Seminars in arthritis and rheumatism, vol. 43, no. 3, pp. 303-313, Dec. 2013. doi: 10.1016/j.semarthrit.2013.07.003.

[3] J. H. Kellgren and J. S. Lawrence, "Radiological assessment of osteo-arthrosis," Ann. Rheum. Dis., vol. 16, no. 4, pp. 494–502, Dec. 1957.

[4] C. Kokkotis, S. Moustakidis, E. Papageorgiou, G. Giakas, and D. E. Tsaopoulos, “Machine learning in knee osteoarthritis: A review,” Osteoarthritis and Cartilage Open, vol. 2, no. 3. Elsevier Ltd, Sep. 01, 2020. doi: 10.1016/j.ocarto.2020.100069.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967

[5] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, “Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach,” Scientific Reports, vol. 8, no. 1, p. 1727, Jan. 2018. doi: 10.1038/s41598-018-20132-7.
[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700-4808.
[7] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, “Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss,” Computerized Medical Imaging and Graphics, vol. 75, pp. 84–92, 2019, doi: https://doi.org/10.1016/j.compmedimag.2019.06.002.
[8] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3-19.
[9] X. Li, “Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds”, IJNDI, vol. 2, no. 1, pp. 93–116, Mar. 2023. doi: 10.53941/ijndi0201006.
[10] J. Antony, K. McGuinness, N. O’Connor, and K. Moran, “Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks.” Oct. 2016. doi: 10.48550/arXiv.1609.02469.
[11] P. Chen, "Knee Osteoarthritis Severity Grading Dataset," Mendeley Data, v1, 2018. [Online]. Available: https://doi.org/10.17632/56rmx5bjcr.
[12] S. M. Pizer et al., "Adaptive histogram equalization and its variations," Comput. Vis. Graph. Image Process., vol. 39, no. 3, pp. 355-368, Sep. 1987.
[13] F. G. Gundersen, S. P. Ojanen, A. P. Reito, P. T. H. Lehenkari, and J. T. J. Lehtimäki, "A deep learning model for prediction of getting a knee replacement within 2 and 5 years for patients with knee osteoarthritis in the Oulu Knee Osteoarthritis Study cohort," BMC Musculoskelet. Disord., vol. 25, Art. no. 738, 2024
[14] X. Ren et al., “OA-MEN: a fusion deep learning approach for enhanced accuracy in knee osteoarthritis detection and classification using X-Ray imaging,” Frontiers in Bioengineering and Biotechnology, vol. 12, 2024, doi: 10.3389/fbioe.2024.1437188.
[15] R. Jose *et al.*, “Machine-learning-based diagnosis and progression analysis of knee osteoarthritis,” *Discover Data*, vol. 3, Oct. 2025, doi: 10.1007/s44248-025-00026-6.
[16] S. Olsson, E. Akbarian, A. Lind, A. Razavian, and M. Gordon, “Automating classification of osteoarthritis according to Kellgren-Lawrence in the knee using deep learning in an unfiltered adult population,” *BMC Musculoskeletal Disorders*, vol. 22, Oct. 2021, doi: 10.1186/s12891-021-04722-7.
[17] A. Khalid, E. Senan, K. Al-Wagih, M. Al-Azzam, and Z. Alkhraisha, “Hybrid Techniques of X-ray Analysis to Predict Knee Osteoarthritis Grades Based on Fusion Features of CNN and Handcrafted,” *Diagnostics*, vol. 13, p. 1609, Oct. 2023, doi: 10.3390/diagnostics13091609.
[18] D. Nasef, D. Nasef, V. Sawiris, P. Girgis, and M. Toma, “Deep Learning for Automated Kellgren–Lawrence Grading in Knee Osteoarthritis Severity Assessment,” *Surgeries*, vol. 6, p. 3, Oct. 2024, doi: 10.3390/surgeries6010003.
[19] C. Vanaja and P. Periasamy, “Convolutional block attention gate-based Unet framework for microaneurysm segmentation using retinal fundus images,” *BMC Medical Imaging*, vol. 25, Oct. 2025, doi: 10.1186/s12880-025-01625-0.
[20] R. Islam and S. Hossain, “Enhanced Brain Tumor Segmentation Using CBAM-Integrated Deep Learning and Area Quantification,” *International Journal of Biomedical Imaging*, vol. 2025, no. 1, 2025, doi: 10.1155/ijbi/2149042.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.14967