RESEARCH ARTICLE

Knee Osteoarthritis Severity Detection Using Multimodal Data

Ayesha Malik[1] **and Rabbia Mahum**[1,*]

[1] Department of Computer Science, University of Engineering and Technology (UET) Taxila, Taxila 47050, Pakistan

Abstract

Osteoarthritis of the knee (KOA) is one of the main causes of disability; hence, it requires precise and early evaluation of the severity level of the disease. In this paper, we propose a multimodal deep learning architecture based on a self-supervised Swin transformer combined with a cross-modal attention mechanism (SWIN-MULTI-ATTEN) for combining radiological images and patients' information (age, sex, and BMI). This framework captures both structural and contextual information to achieve better classification accuracy. Experimental results on the Osteoarthritis Initiative (OAI) dataset show that the proposed model achieves an accuracy of 91.4% and a QWK of 0.903, which outperforms other CNN-based and transformer-based architectures.

Keywords: Deep learning, Transformers, Multi-attention, Osteoarthritis.

Submitted: 26 May 2025
Revised: 20 July 2026
Accepted: 21 July 2026
Published: 10 September 2026

Vol. 2, **No.** 4, 2026.
10.62762/JIAP.2026.275643

***Corresponding author:**
Rabbia Mahum
rabbia.mahum@uettaxila.edu.pk

1 Introduction

Osteoarthritis (OA) is the most prevalent musculoskeletal disorder of the knee, characterized by the gradual destruction of cartilage and bone tissues surrounding the joint [1]. OA is one of the most common causes of disability that impairs the ability to move around and perform routine activities. The condition causes discomfort, joint pain, and stiffness. Osteoarthritis is a common health problem worldwide according to WHO, and its prevalence will increase with an aging population and obesity in the coming years [2]. Early and precise detection is crucial for proper disease management and treatment strategy formulation. Currently, various clinical examinations and techniques including Computed Tomography (CT), MRI, and X-rays are applied to diagnose the presence of osteoarthritis. In particular, the use of X-rays is widely employed owing to their availability and lower costs compared to other radiological techniques. Usually, the severity of the disease is assessed by applying KL grading scales [3].

There are various advantages and disadvantages of imaging tests for osteoarthritis; in general, radiography is widely used because of its availability and affordability, whereas more modern imaging tests such as MRI give a better insight into soft tissues and joint lesions [4]. This issue highlights the necessity for the development of automated techniques.

Citation
Malik, A., & Mahum, R. (2026). Knee Osteoarthritis Severity Detection Using Multimodal Data. *ICCK Journal of Image Analysis and Processing*, 2(4), 206–217.

However, the traditional machine learning approaches make use of hand-crafted features that require domain knowledge and are unable to discover small-scale patterns within medical images [5]. In addition, these models do not scale well for large datasets and face challenges in processing high-dimensional data. Traditional methods are also inefficient in handling image variations, noise, and complex anatomy; these limitations result in inferior performance compared with deep learning-based models for osteoarthritis recognition, segmentation, and severity assessment [6].

Deep learning has emerged as an effective technology for medical image analysis and disease diagnosis. Convolutional Neural Network (CNN) models are designed for performing image analysis and have performed exceptionally well when dealing with complicated patterns in medical images. These models extract features hierarchically from raw data, eliminating the need for manual feature engineering. CNN models have been shown to match or even outperform human expertise in medical image analysis [7, 8]. A key advantage of CNNs in medical image analysis, including KOA recognition, is their ability to automatically learn spatial hierarchies of structural features without manual engineering. This contrasts with other algorithms like Support Vector Machines (SVM) or K-Nearest Neighbor (KNN) that rely on hand-crafted features to perform effectively on image data [9]. Moreover, CNNs exhibit translation equivariance through weight sharing, which helps them detect anatomical structures irrespective of their location within the image; robustness to changes in orientation is not inherent and is typically obtained through data augmentation. ResNet and VGGNet CNN models provide deep feature learning capabilities through residual connections and deep convolutional stacking, respectively [10, 11].

Despite CNNs showing better results, certain limitations still exist due to their high computational cost and the necessity of large amounts of labeled data [5, 8], as well as difficulties with the interpretability of trained models [12]. Nevertheless, current studies address these limitations through transfer learning, data augmentation, and interpretable AI techniques. Therefore, the motivation behind this research lies in the development of a technique for automatic osteoarthritis recognition and severity classification based on medical imaging and clinical data.

Over the last few years, there has been a considerable amount of development of automated KOA diagnosis through deep learning methods that include transformers and multimodal approaches. Several research works have proved the supremacy of the vision transformer architecture over conventional CNNs in the task of learning global context from medical images [13–15]. For example, hybrid vision transformer models have demonstrated higher levels of performance in grading KOA severity by utilizing hierarchical feature extraction and attention modules [15, 16]. Also, multimodal vision transformers were introduced to fuse medical images with patient characteristics for comprehensive analysis and prediction of disease development [16].

Even more recently, sophisticated Swin Transformer architectures have become widely used because of the possibility to capture local and global dependencies [17]. In 2026, researchers introduced an advanced Swin Transformer model with two kinds of attention mechanisms that surpassed other CNNs and transformer architectures in classifying KOA severity [18]. Also, several multimodal learning frameworks were developed to combine imaging and clinical data for further analysis, including clinically integrated multi-modal transformer approaches with cross-modal gated fusion for automated KL grading [19]. Interpretable transformer frameworks have also been proposed to enhance clinical transparency of KOA grading from X-ray images [20], while other works have combined image and language modalities for knee replacement prediction [21].

However, some challenges still need to be addressed. Most of the methods focus only on image modality without taking into consideration the patients' clinical information that plays an important role in their diagnosis. In addition, issues related to interpretability and generalization of the models are still relevant for clinical practice.

This paper introduces a new multimodal framework based on the combination of radiographs and clinical information with the help of cross-modal attention mechanisms. Key contributions of this research can be listed as follows:

- A novel multimodal architecture with a self-supervised Swin Transformer and clinical data for better KOA severity classification;
- A cross-modal attention mechanism that helps to

dynamically align individual clinical characteristics of the patients with their radiographic pattern;

- Self-supervised learning to improve feature extraction from scarce medical data labels;
- Extensive experiments proving that the proposed approach outperforms CNN and transformer baselines;
- Improved interpretability through attention-mechanism visualizations to assist clinical decision-making.

These contributions will allow for developing efficient computer-aided systems for KOA.

2 Related Work

The traditional method to detect KOA was X-ray images, which included human inspection and was not precise. Scientists worked on automated systems to detect KOA where machine learning (ML) algorithms have traditionally been utilized for X-ray image analysis via a systematic process that includes preprocessing, manual feature extraction, feature selection, and classification. First, techniques like noise reduction and contrast enhancement were applied to improve X-ray images. Domain-specific attributes—such as texture (e.g., Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP)), shape, and intensity patterns—were manually obtained to describe disease traits like joint space narrowing in KOA. The features were subsequently refined through techniques like Principal Component Analysis (PCA) prior to being input into classifiers including SVM, KNN, and Random Forests for the purpose of disease classification. Although these methods facilitated the initial automation of diagnosis, their effectiveness was largely reliant on the quality of manually created features, restricting their capacity to recognize complicated spatial patterns in medical images. Moreover, these pipelines were computationally slow and difficult to scale. The limitation of ML algorithms is that they only process small datasets of 2-dimensional (2D) X-ray images and cannot process multi-class data or 3D data like CT images [22]. Early efforts to overcome these constraints explored multimodal combinations of radiographic and clinical data, demonstrating that integrating patient-level information can improve prognostic accuracy even within classical machine learning pipelines [23].

To enhance classification accuracy, researchers have focused on the improvement of classification performance with deep learning algorithms like CNN-based architectures. The performance in KOA detection has improved with CNN models that are hybrid and a combination of deep features with traditional classifiers (e.g., SVM). To enhance the classification accuracy, Mahum et al. [24] proposed a hybrid CNN-based system, which integrates the deep features with the machine learning classifiers. Ensemble and hybrid CNN architectures have demonstrated superior performance compared with single-model approaches by integrating multiple feature representations, which enhances robustness and generalization in X-ray-based classification tasks [25]. It is evident that CNNs are important not only for enabling automated disease diagnosis but also for minimizing the subjectivity inherent in manual radiographic interpretation. CNN-based models such as ResNet, DenseNet, and EfficientNet have been able to display high efficacy in automatic extraction of spatial features such as joint space narrowing, osteophyte formation, and bone sclerosis in radiographic images [26]. Recent studies have also explored lightweight CNN architectures for automated KOA classification. For example, TurkerNeXtV2 was developed as a CNN-based framework for knee osteoarthritis classification, demonstrating the continuing relevance of computationally efficient convolutional architectures for automated KOA assessment [27]. Pan et al. [28] demonstrated that a hierarchical CNN-based classification approach can improve the KL grading process. Fusion-based deep learning approaches have also been explored to improve KOA detection and classification by combining complementary feature representations. For example, Ren et al. [29] proposed OA-MEN, a fusion deep learning framework for enhanced knee osteoarthritis detection and classification using X-ray imaging. In addition to CNN-based approaches, alternative deep feature learning methods have also been investigated for automated KOA assessment. For example, Amjad et al. [30] employed autoencoders for automated feature representation together with extreme learning machines for knee osteoarthritis detection and classification. Such approaches demonstrate that learned latent representations can provide an alternative to conventional handcrafted features for KOA analysis. Nevertheless, CNN- and transformer-based architectures remain particularly attractive for radiographic severity grading because they can directly model spatial and contextual information from knee X-ray images.

Another significant research trajectory concerns disease progression prediction and fine-grained severity grading. Incorporating multimodal MRI data, deep learning algorithms like DeepKOA can predict disease progression, allowing for early intervention strategies [31]. Nevertheless, multi-class classification like KL classification is much harder than binary classification because intermediate steps of disease progression are not so distinct [32].

Beyond classification and progression prediction, deep learning has also been successfully applied to knee structure segmentation. Cartilage and joint spaces can be segmented by using U-Nets and other models derived from the U-Net architecture. CNN-based segmentation models are shown to have high performance in the identification of anatomical structures, but the accuracy relies heavily on the quality of the datasets and the consistency of annotations [33]. Furthermore, to enhance the transparency of the models and clinical trust, explainable AI (XAI) methods are also being integrated, addressing one of the main constraints of deep learning models [32].

Transfer learning (TL) and 3D deep learning methods to process 3D scans have been widely investigated in addition to conventional CNN models. Yeoh et al. [34] trained a transfer learning-based 3D deep learning model that showed superior performance for KOA detection by pre-training of networks with a large dataset of OAI MRI data. Furthermore, multitask learning and hybrid learning models have been proposed to jointly analyze multiple features of OA. For example, Teoh et al. proposed a multi-task deep hybrid learning framework based on pretrained CNNs for feature extraction and machine learning classifiers for multi-feature diagnosis, enhancing overall diagnostic capability [35].

By fine-tuning feature extraction procedures, dual CNN frameworks optimized with metaheuristic algorithms, like the Gaussian Aquila optimizer, have shown increased detection and grading accuracy [36]. Systematic reviews also show that deep learning-based X-ray analysis, particularly the CNN-based method, is consistently superior in both the classification and the severity grading tasks, with some limitations like dataset imbalance and generalizability still being faced [26].

Despite all these successes that have been experienced with deep learning, certain issues still need to be addressed in the future. First, deep learning models lack generalizability between various datasets. Most studies rely on publicly available datasets such as the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST) [26, 32]. Other constraints include the requirement for huge datasets, availability of computing resources, and challenges in implementing the findings in real clinical practice.

To overcome these data limitations, transformer-based algorithms such as the Vision Transformer (ViT) have emerged as powerful alternatives to CNNs. The transformer algorithms employ self-supervised learning to detect global dependencies of the features within the dataset. Unlike CNNs that employ local feature detection methods, transformers have been found to work better on huge datasets. Therefore, because of the need for huge datasets and computing power, transformers have been considered unsuitable for medical image datasets that usually have very little data available. Thus, hybrid CNN-Transformer models are becoming more popular [26, 37].

In summary, while CNN-based deep learning has established strong baselines for KOA detection, classification, and segmentation, significant gaps remain in model generalization, interpretability, and clinical utility. These limitations motivate the development of more powerful and scalable multimodal solutions, as pursued in the present work.

3 Materials

The proposed study used the publicly available dataset Osteoarthritis Initiative (OAI) of knee radiographs with clinical data to study KOA. It includes visual data (X-rays) and tabular data (age, BMI, and sex).

In total, 8,000 knee radiographs with clinical metadata were selected from the OAI database for this study. Inclusion criteria required the availability of paired anteroposterior knee radiographs and complete clinical records (age, sex, and BMI); images with missing metadata or severe imaging artefacts were excluded. The selected radiographs were taken during regular follow-up examinations of enrolled participants over a longitudinal observation period. KOA severity in these images was graded according to the Kellgren–Lawrence (KL) classification system, which categorizes radiographs into five groups:

- Grade 0 – none (no radiographic features of OA)
- Grade 1 – doubtful
- Grade 2 – minimal (mild)

- Grade 3 – moderate
- Grade 4 – severe

The KL classification is the most common method for osteoarthritis severity evaluation [32].

4 Methodology

The proposed SWIN-MULTI-ATTEN framework combines a Swin Transformer with multimodal fusion and attention mechanisms for five-class KOA severity grading. The framework utilizes knee radiographs and corresponding clinical information from the Osteoarthritis Initiative (OAI) dataset. As illustrated in Figure 1, the overall pipeline consists of image and clinical data preprocessing, self-supervised visual feature extraction using the Swin Transformer backbone, cross-modal fusion of radiographic and clinical features, multi-level attention, and final classification into the five Kellgren–Lawrence (KL) grades.

4.1 Dataset Preparation

The dataset is separated into three categories: 70% training, 15% validation, and 15% test data. The split was performed at the patient level, so that all radiographs from a given patient appear in only one partition, preventing identity leakage across the training, validation, and test sets. Tabular clinical characteristics, including patient age (X_1), sex (X_2), and Body Mass Index (BMI) (X_3), were paired with images to enable multimodal analysis.

4.2 Preprocessing

4.2.1 Image Data Processing

- **Conversion:** Raw X-ray images were first converted into 8-bit greyscale images and a uniform resolution of 224×224.
- **Feature Extraction:** The tibiofemoral joint was defined as the Region of Interest (ROI) and localized using a YOLOv8 detector, leveraging its demonstrated effectiveness in anatomical region detection from musculoskeletal X-ray images [38]. This step removed extraneous regions and increased robustness.
- **Normalization:** Image intensity was normalized by calculating the mean and standard deviation.
- **Data Augmentation:** Data augmentation techniques, including horizontal flip, Gaussian blur, and rotation, were applied to enlarge the effective training distribution and improve the model's robustness to anatomical orientation variations in knee radiographs, consistent with the demonstrated benefits of augmentation for KOA X-ray classification [39].

4.2.2 Clinical Data Preprocessing

- **Normalization:** Z-score normalization was used to standardize continuous clinical information like age and BMI.
- **Encoding:** One-hot encoding was used to modify categorical variables like sex to facilitate successful multimodal integration. Using a linear projection layer, these processed characteristics were then projected into a high-dimensional embedded space, matching their dimensionality with visual feature tokens [40, 41].

4.3 Model Architecture

A multimodal fusion block, a multi-level attention module, and a hierarchical visual backbone made up the three main parts of the proposed SWIN-MULTI-ATTEN structure.

4.3.1 Self-Supervised Swin Backbone Architecture

The Swin Transformer Tiny (Swin-T) architecture extracts visual features. After ROI extraction, each image was rescaled to a uniform resolution prior to patch partitioning. Swin-T uses a hierarchical design with shifting window-based self-attention, which successfully captures multi-scale contextual information while achieving linear computational complexity in terms of image size [17]. Its effectiveness as a general-purpose image classification backbone has been further validated across diverse vision tasks [42].

- **Partitioning the image:** The input image was presented in disjoint patches of 4×4.
- **Window Shift:** Self-attention was computed within 7×7 local windows, with periodic window shifting to facilitate cross-window interactions. This mechanism is particularly effective for modeling global spatial dependencies essential for detecting subtle joint space narrowing.

The Swin-T backbone was pre-trained using Self-Supervised Learning (SSL) on unlabeled knee radiographs (drawn only from the training partition, so that no information from the validation or test sets leaks into pretraining) to improve feature representation. Specifically, a masked image modeling pretext task (SimMIM [43]) was adopted, in which a random subset of image patches is masked and the

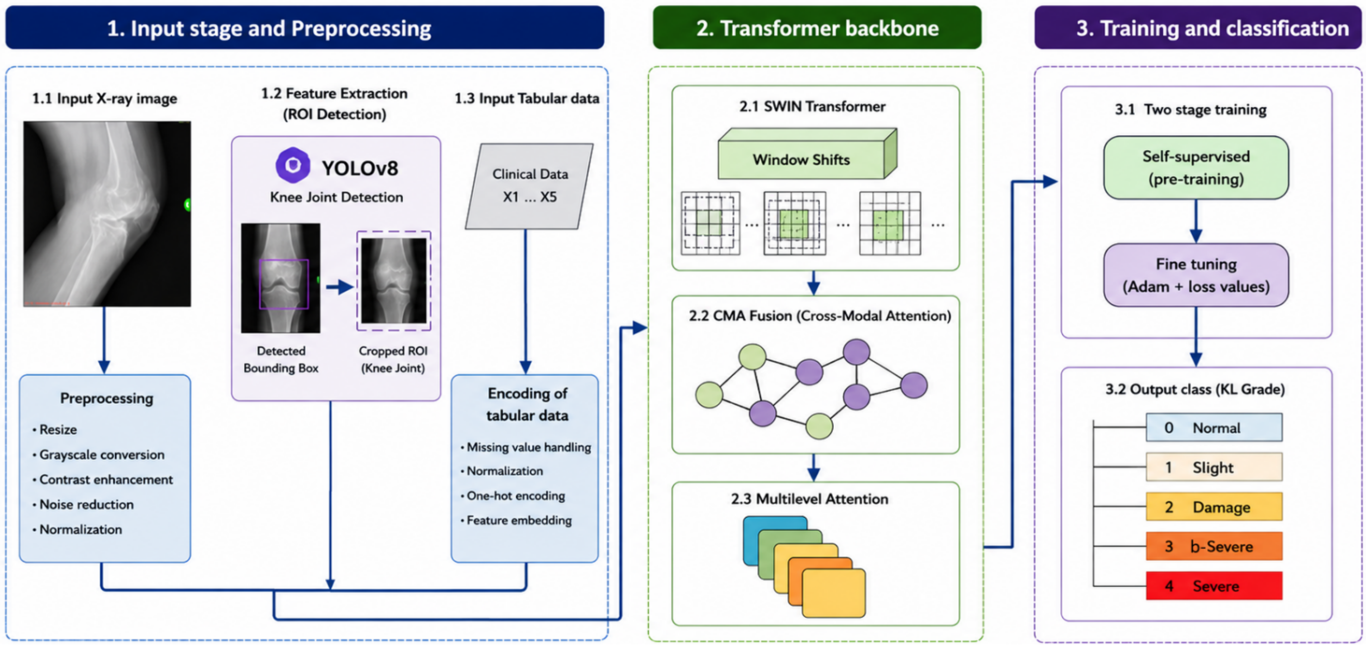


Figure 1. Flow diagram of knee osteoarthritis classification.

network is trained to reconstruct the missing raw-pixel content; this objective is well suited to the hierarchical Swin backbone. The model was able to learn rich, label-free representations of cartilage texture, bone density, and structural patterns according to this method, consistent with the strong representation learning capacity demonstrated by masked image modeling on hierarchical vision backbones [43].

4.3.2 *Multimodal Fusion via Cross Attention*

For efficient fusion, a Cross-Modal Attention (CMA) module was added in place of basic feature concatenation. Let $\mathbf{e}_c \in \mathbb{R}^{d_c}$ denote the embedded clinical vector (age, sex, BMI) and $\mathbf{Z} \in \mathbb{R}^{T \times d}$ denote the T visual tokens produced by the Swin backbone. The query, key, and value are obtained through learned linear projections rather than raw inputs: the query is derived from the clinical embedding, $F_q = \mathbf{e}_c \mathbf{W}_q$, whereas the key $F_k = \mathbf{Z}\mathbf{W}_k$ and the value $F_v = \mathbf{Z}\mathbf{W}_v$ are derived from the visual tokens, where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are trainable weight matrices. Here T is the number of visual tokens (keys/values) and d (i.e., d_k) is the dimension of the query and key vectors used to scale the dot product. The cross-modal attention is formulated as:

$$\mathrm{Attention}(F_q, F_k, F_v) = \mathrm{softmax}\left(\frac{F_q F_k^{\top}}{\sqrt{d}}\right) F_v. \quad (1)$$

Using the clinical embedding as the query, the module attends over all visual tokens through F_k and aggregates the corresponding values F_v, so that the model can dynamically emphasize the radiographic regions most associated with patient-specific clinical indications (e.g., increased focus on osteophyte detection in patients with higher age or BMI). The scaling factor $\sqrt{d}$ counteracts the growth of dot-product magnitudes in high dimensions and stabilizes the softmax gradients.

4.3.3 *Multi-level Attention*

A dual attention mechanism was added to further enhance the fused features:

- **Channel Attention (Squeeze-and-Excitation):** This adjusts the weighting of the various feature channels, giving more importance to those that represent textural patterns of high frequency that correspond to subchondral sclerosis (the thickness and hardness of the bone layer below the cartilage in a joint).
- **Spatial Attention:** It creates a 2D spatial weight map for the model to focus on clinically relevant areas, including the medial and lateral tibiofemoral compartments.

4.4 Two-Stage Training

The network is trained in two stages:

- **Phase I (SSL):** The model is pre-trained on unlabeled X-rays with a pretext task using the Swin backbone.
- **Phase II (Supervised):** The entire multimodal network is then fine-tuned using KL-labeled data.

During supervised fine-tuning, the model was trained for a maximum of 100 epochs using AdamW with a weight decay of 0.05 and a cosine-annealing learning-rate scheduler. Model performance was monitored on the validation set throughout training, and the checkpoint achieving the lowest validation loss was retained for final evaluation on the independent test set. The total loss is a weighted combination of Categorical Cross-Entropy $\mathcal{L}_{\text{CE}}$ and an Ordinal Logits Loss $\mathcal{L}_{\text{ord}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{ord}}, \tag{2}$$

where $\lambda \geq 0$ is a balancing hyperparameter selected on the validation set. The ordinal term better respects the ordinal nature of KL grades by applying lower penalties for adjacent-class misclassifications (e.g., KL-3 vs. KL-4) than for distant ones [44].

4.5 Classification Output

Following two-stage training, each test radiograph is classified by the model into one of five KL severity grades: none (KL-0), doubtful (KL-1), minimal (KL-2), moderate (KL-3), and severe (KL-4).

5 Experiments and Results

5.1 Model Evaluation

Multiple standard classification metrics suitable for multi-class and ordinal problems are applied to the test dataset for primary model evaluation, such as accuracy, macro-averaged precision, recall, and F1-score, and Quadratic Weighted Kappa (QWK). QWK was regarded as the most essential metric because it accounts for the ordinal character of Kellgren–Lawrence (KL) grades, penalizing distant misclassifications more significantly than nearby ones [45]. All reported metrics are computed on the held-out test set derived from the single patient-level split described in Section 4.1; the patient-level partitioning ensures that performance estimates are not inflated by within-patient correlations.

5.2 Qualitative Assessment

To qualitatively assess the model's decision-making process, attention-map visualizations and Grad-CAM++ were applied to the test set predictions. Using these approaches, clinicians can confirm whether the model is focusing on clinically significant pathological symptoms or not, such as joint space constriction, osteophyte formation, and subchondral sclerosis. These approaches generate intuitive heatmaps that highlight the most significant regions in the input X-ray images, as shown in Figure 2. The evaluated results are presented in Table 1.

Table 1. Evaluation metrics.

Metric	Value	Description
Accuracy	91.4%	Overall correct classification rate across all KL grades
Macro AUC-ROC	0.974	Area under the ROC curve (one-vs-rest)
Macro F1-Score	89.6%	Average F1-Score across all classes (KL-0 to KL-4)
Quadratic Weighted Kappa (QWK)	0.903	Primary metric; penalizes distant misclassification
Precision (Macro)	90.1%	Average precision across classes
Recall (Macro)	89.4%	Average recall across all classes

5.3 Performance Analysis and Class-wise Evaluation

The SWIN-MULTI-ATTEN model was tested using the reserved test set from the Osteoarthritis Initiative (OAI) dataset. The model demonstrated stable convergence and consistent classification performance across all KL grades.

5.3.1 Model Loss

The training and validation loss curves illustrate the optimization behavior of the model over 100 epochs. The model was trained using a composite loss function combining Categorical Cross-Entropy Loss and Ordinal Logits Loss. As shown in Figure 3, both the training and validation losses decreased substantially during the initial training stage. The validation loss began to stabilize after approximately 25 epochs and subsequently exhibited only minor fluctuations, while training was continued for the full 100 epochs. The model checkpoint with the lowest validation loss was retained for final evaluation on the test set. The close correspondence between the training and validation loss curves suggests stable optimization without substantial evidence of overfitting. These results also indicate the effectiveness of the AdamW optimizer with the cosine-annealing learning-rate scheduler and the benefit of self-supervised pretraining in Phase I.

5.3.2 Model Accuracy

The model showed rapid improvement during the early training epochs, which may be attributed in part to the initialization obtained through self-supervised pretraining. As shown in Figure 4, the training and validation accuracies increased progressively and approached a plateau during the later stages of training. By the end of the 100-epoch training schedule,

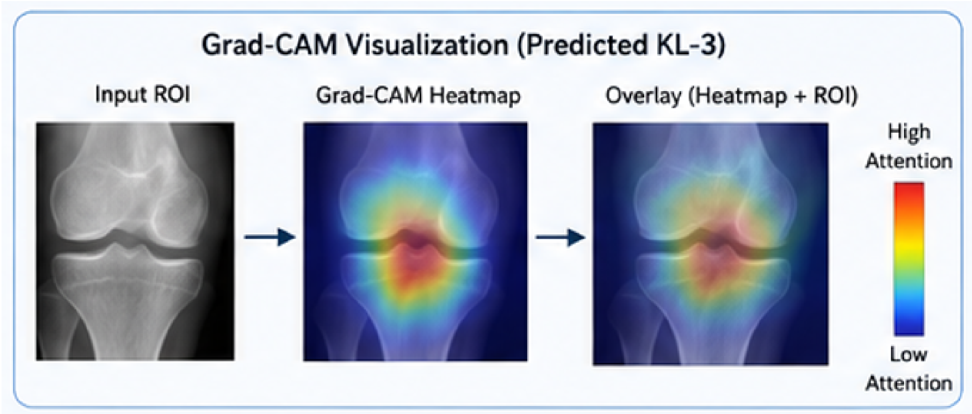


Figure 2. Grad-CAM++ visualization.

Table 2. Comparison with Transformer and CNN baselines. All baseline models were retrained on the same OAI patient-level data split using their official implementations with ImageNet-pretrained weights; hyperparameters were tuned on the validation set under the same training protocol as the proposed model.

Model Category	Model	Accuracy (%)	F1-Score (%)	QWK	AUC
CNN-based	ResNet50	82.4	80.1	0.812	0.935
CNN-based	DenseNet121	84.7	82.9	0.837	0.948
CNN-based	EfficientNet-B4	85.9	84.2	0.846	0.951
Transformer-based	ViT-Base	86.3	84.5	0.851	0.952
Transformer-based	Swin-T (Image-only)	88.1	86.7	0.874	0.961
Multimodal Transformer	**SWIN-MULTI-ATTEN (Proposed)**	**91.4**	**89.6**	**0.903**	**0.974**

Figure 3. Training and validation loss curves of the SWIN-MULTI-ATTEN model over 100 epochs.

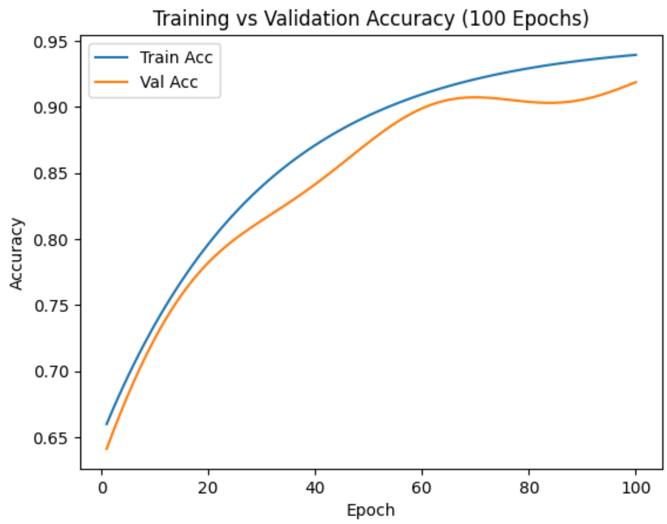


Figure 4. Training and validation accuracy curves of the SWIN-MULTI-ATTEN model over 100 epochs.

the training accuracy reached 93.8%, while the validation accuracy reached 91.7%. The gap between the training accuracy (93.8%) and the validation accuracy (91.7%) was 2.1 percentage points, indicating that the model maintained stable generalization on the internal validation set without significant overfitting. Furthermore, the absence of significant divergence between the training and validation curves indicates that the model did not overfit during training. The validation loss also became steady after convergence, demonstrating the robustness of the learned feature representations. These curves confirm the model's stable generalization and the positive impact of multi-level attention and cross-modal fusion.

5.4 Fusion Module and Multi-Level Attention Enhancement

The contribution of the Multi-Level Attention and Cross-Modal Attention (CMA) fusion mechanisms was measured by ablation research, as summarized in Table 3. Accuracy decreased by 2.8% and QWK decreased by 0.031 when the CMA module was removed. There was an additional 1.9% decrease in accuracy when the dual attention (channel + spatial) was disabled. These findings confirm that clinical characteristics that dynamically direct visual attention greatly improve performance, especially when it comes to identifying minute variations between neighboring KL grades (e.g., KL-2 vs. KL-3).

Table 3. Ablation study results on the OAI test set.

Configuration	Accuracy (%)	QWK
Full model (SWIN-MULTI-ATTEN)	91.4	0.903
w/o Cross-Modal Attention (CMA)	88.6	0.872
w/o Dual Attention (channel+spatial)	86.7	0.851

5.5 Comparative Study

5.5.1 Comparison of Machine Learning Techniques vs. Existing Deep Learning

Conventional ML techniques (such as SVM with manually engineered features, e.g., Haralick Texture/GLCM) usually get an accuracy of around 70 to 78% [46]. However, deep learning architectures normally attain higher accuracy levels, above 80%, by virtue of learning hierarchical structures automatically [11]. Advanced transformer architectures such as Swin-based models have achieved accuracy rates exceeding 88%, outperforming standard CNN baselines [13, 26].

5.5.2 Comparison of CNN, Transformers, and Our Proposed Methodology

As shown in Table 2, the proposed SWIN-MULTI-ATTEN model achieved the best overall performance among the evaluated CNN- and transformer-based baselines, with an accuracy of 91.4%, an F1-score of 89.6%, a QWK of 0.903, and an AUC of 0.974. Compared with the image-only Swin-T baseline, which achieved an accuracy of 88.1% and a QWK of 0.874, the proposed model improved accuracy by 3.3 percentage points and QWK by 0.029. These results suggest that integrating clinical information with radiographic features through cross-modal attention and multi-level attention provides additional predictive information for KOA severity grading.

6 Key Findings

- **Integration of multimodal data:** Combining tabular clinical data (age, BMI, sex) with X-ray image features through cross-attention improves performance over image-only models.
- **Hierarchical transformers:** This architecture excels over CNNs in KOA grading because it can capture both fine-grained textures and global joint structure.
- **Self-supervised pretraining and ordinal loss:** This combination proves highly effective for medical imaging tasks with limited labeled data and ordinal class structure.
- **Attention mechanisms:** These boost both classification performance and model explainability, making the proposed framework more suitable for clinical deployment.

In summary, the SWIN-MULTI-ATTEN model establishes a new benchmark for automated KOA severity assessment by effectively combining imaging and clinical data. The consistent improvements across quantitative evaluations, ablation studies, and visualizations support its potential as a reliable computer-aided diagnostic system.

7 Conclusion

The proposed multimodal deep learning architecture demonstrates the benefits of combining a hierarchical Swin Transformer with cross-modal attention mechanisms for diagnosing KOA severity. By effectively integrating radiographic and clinical data, the proposed framework achieves superior diagnostic performance relative to existing unimodal and CNN-based approaches. Advanced multimodal architectures hold considerable promise for broader applications in medical image analysis. Further research will focus on validating the proposed approach in real-life scenarios for early diagnosis of KOA.

Data Availability Statement

The raw knee radiograph data and clinical metadata analysed in this study are derived from the Osteoarthritis Initiative (OAI) database, which is publicly available at https://nda.nih.gov/oai/. The processed data subsets and model code generated during this study are not publicly available due to institutional policies but may be made available from the corresponding author upon reasonable request.

Funding

This work was supported without any funding.

Conflicts of Interest

Rabbia Mahum served as an Associate Editor of the *ICCK Journal of Image Analysis and Processing* at the time of manuscript submission. To ensure the integrity of the peer-review process, Rabbia Mahum was not involved in the editorial handling, peer review, or decision-making process for this manuscript, which was handled independently by another editor. The remaining authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

The Osteoarthritis Initiative (OAI) study was approved by the institutional review boards at each of the participating clinical sites and the coordinating center, and written informed consent was obtained from all original study participants prior to enrolment. The present study constitutes a secondary analysis of publicly available, de-identified OAI data and did not involve the recruitment of new participants or the collection of additional personal data; accordingly, separate ethical approval and informed consent were not required for this analysis.

References

[1] Hunter, D. J., & Bierma-Zeinstra, S. (2019). Osteoarthritis. *The Lancet*, 393(10182), 1745-1759. [CrossRef]

[2] World Health Organization. (2023). *Osteoarthritis*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/osteoarthritis

[3] Kellgren, J. H., & Lawrence, J. S. (1957). Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases, 16*(4), 494-502. [CrossRef]

[4] Guermazi, A., Hayashi, D., Roemer, F. W., & Felson, D. T. (2013). Osteoarthritis: A review of strengths and weaknesses of different imaging options. *Rheumatic Disease Clinics of North America, 39*(3), 567-591. [CrossRef]

[5] Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering, 19*, 221-248. [CrossRef]

[6] 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. [CrossRef]

[7] Thomas, K. A., Kidziński, Ł., Halilaj, E., Fleming, S. L., Venkataraman, G. R., Oei, E. H., ... & Delp, S. L. (2020). Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiology: Artificial Intelligence, 2*(2), e190065. [CrossRef]

[8] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., . . . & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60-88. [CrossRef]

[9] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging, 9*(4), 611-629. [CrossRef]

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 770-778). [CrossRef]

[11] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. [CrossRef]

[12] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247-278. [CrossRef]

[13] Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., ... & Hamamoto, R. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of medical systems, 48*(1), 84. [CrossRef]

[14] Wang, Z., Chetouani, A., & Jennane, R. (2023, September). Transformer with Selective Shuffled Position Embedding for Early Detection of Knee Osteoarthritis. In *Proceedings of the 20th International Conference on Content-based Multimedia Indexing* (pp. 261-264). [CrossRef]

[15] Jahan, M., Hasan, M. Z., Samia, I. J., Fatema, K., Rony, M. A. H., Arefin, M. S., & Moustafa, A. (2024). KOA-CCTNet: An enhanced knee osteoarthritis grade assessment framework using modified compact convolutional transformer model. *IEEE Access, 12*, 107719-107741. [CrossRef]

[16] Patel, U., Patel, S., Patel, D., Jain, N., Patel, S., & Gangavani, R. (2026). Hybrid CNN-transformer for knee osteoarthritis severity grading. *Journal of Innovative Image Processing, 8*(2), 617-643. [CrossRef]

[17] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021, October). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision* (*ICCV*) (pp. 9992-10002).

IEEE. [CrossRef]

[18] Sudha, K., & Kannan, A. R. (2026). Enhanced Swin transformer with dual attention for knee osteoarthritis severity grading from X-ray images. *Scientific Reports, 16*, 10617. [CrossRef]

[19] Huang, Y., Shao, Z., Wang, R., & Qiu, H. (2026). Clinically integrated multi-modal transformer framework with cross-modal gated fusion and clinical nomogram for automated Kellgren-Lawrence grading of knee osteoarthritis on x-ray images. *BMC Musculoskeletal Disorders, 27*(1), 528. [CrossRef]

[20] Roy, B., Das, J., Baidya, A., & Banik, D. (2026). Osteotrans-Net: An interpretable transformer framework for knee osteoarthritis severity grading in X-ray scans. *Biomedical Signal Processing and Control, 113*, 109096. [CrossRef]

[21] Zhang, C., Chen, S., Cigdem, O., Rajamohan, H. R., Cho, K., Kijowski, R., & Deniz, C. M. (2025). MR-Transformer: A Vision Transformer-based Deep Learning Model for Total Knee Replacement Prediction Using MRI. *Radiology: Artificial Intelligence, 7*(5), e240373. [CrossRef]

[22] Shamir, L., Ling, S. M., Scott, W., Hochberg, M., Ferrucci, L., & Goldberg, I. G. (2009). Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthritis and Cartilage, 17*(10), 1307-1312. [CrossRef]

[23] Tiulpin, A., Klein, S., Bierma-Zeinstra, S. M., Thevenot, J., Rahtu, E., Meurs, J. V., ... & Saarakkala, S. (2019). Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports, 9*(1), 20038. [CrossRef]

[24] Mahum, R., Rehman, S. U., Meraj, T., Rauf, H. T., Irtaza, A., El-Sherbeeny, A. M., & El-Meligy, M. A. (2021). A novel hybrid approach based on deep CNN features to detect knee osteoarthritis. *Sensors, 21*(18), 6189. [CrossRef]

[25] Pi, S. W., Lee, B. D., Lee, M. S., & Lee, H. J. (2023). Ensemble deep-learning networks for automated osteoarthritis grading in knee X-ray images. *Scientific Reports, 13*(1), 22887. [CrossRef]

[26] 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 KL grades of knee osteoarthritis: a systematic review and meta-analysis. *European Radiology, 35*(1), 327-340. [CrossRef]

[27] Esmez, O., Deniz, G., Bilek, F., Gurger, M., Barua, P. D., Dogan, S., ... & Tuncer, T. (2025). TurkerNeXtV2: An Innovative CNN Model for Knee Osteoarthritis Pressure Image Classification. *Diagnostics, 15*(19), 2478. [CrossRef]

[28] Pan, J., Wu, Y., Tang, Z., Sun, K., Li, M., Sun, J., ... & Shen, B. (2024). Automatic knee osteoarthritis severity grading based on X-ray images using a hierarchical classification method. *Arthritis Research & Therapy, 26*(1), 203. [CrossRef]

[29] Ren, X., Hou, L., Liu, S., Wu, P., Liang, S., Fu, H., ... & Cheng, Y. (2025). 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, 12*, 1437188. [CrossRef]

[30] Amjad, J., Sajid, M. Z., Amjad, A., Hamid, M. F., Youssef, A., & Sharif, M. I. (2025). Knee osteoarthritis detection and classification using autoencoders and extreme learning machines. *AI, 6*(7), 151. [CrossRef]

[31] Hu, J., Zheng, C., Yu, Q., Zhong, L., Yu, K., Chen, Y., ... & Zhang, X. (2023). DeepKOA: a deep-learning model for predicting progression in knee osteoarthritis using multimodal magnetic resonance images from the osteoarthritis initiative. *Quantitative Imaging in Medicine and Surgery, 13*(8), 4852. [CrossRef]

[32] Lavanya, R., & Senthilkumar, S. (2026). Deep learning in knee osteoarthritis: A task-based systematic review of recent advances. *Biomedical Signal Processing and Control, 119*, 109771. [CrossRef]

[33] Rani, S., Rout, A., Soni, P., Gupta, M., Kumar, N., & Kumar, K. (2026). Review of CNN-Based Approaches for Preprocessing, Segmentation and Classification of Knee Osteoarthritis. *Diagnostics, 16*(3), 461. [CrossRef]

[34] Yeoh, P. S. Q., Lai, K. W., Goh, S. L., Hasikin, K., Wu, X., & Li, P. (2023). Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: Data from the osteoarthritis initiative. *Frontiers in Bioengineering and Biotechnology, 11*, 1164655. [CrossRef]

[35] Teoh, Y. X., Othmani, A., Lai, K. W., Goh, S. L., & Usman, J. (2023). Stratifying knee osteoarthritis features through multitask deep hybrid learning: data from the osteoarthritis initiative. *Computer methods and programs in biomedicine, 242*, 107807. [CrossRef]

[36] Subha, B., Jeyakumar, V., & Deepa, S. N. (2024). Gaussian Aquila optimizer based dual convolutional neural networks for identification and grading of osteoarthritis using knee joint images. *Scientific Reports, 14*(1), 7225. [CrossRef]

[37] Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., & Xie, Y. (2024). From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine, 37*(4), 1529-1547. [CrossRef]

[38] Ju, R. Y., & Cai, W. (2023). Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. *Scientific reports, 13*(1), 20077. [CrossRef]

[39] Prezja, F., Paloneva, J., Pölönen, I., Niinimäki, E., & Äyrämö, S. (2022). DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. *Scientific Reports, 12*(1), 18573. [CrossRef]

[40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[41] Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, *3*(1), 136. [CrossRef]

[42] Dihin, R. A., Al Shemmary, E. N., & Al-Jawher, W. A. M. (2023). Implementation of the swin transformer and its application in image classification. *Journal Port Science Research*, *6*(4), 318-331. [CrossRef]

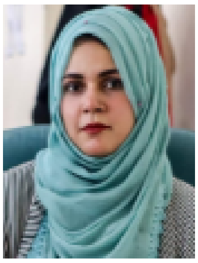
[43] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., ... & Hu, H. (2022, June). Simmim: A simple framework for masked image modeling. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 9643-9653). IEEE. [CrossRef]

[44] Ullah, Z., & Kim, J. (2026). VL-OrdinalFormer: Vision–language-guided ordinal transformers for interpretable knee osteoarthritis grading. *Mathematics*, *14*(6), 963. [CrossRef]

[45] Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. [CrossRef]

[46] Shamir, L., Ling, S. M., Scott, W. W., Bos, A., Orlov, N., Macura, T. J., ... & Goldberg, I. G. (2008). Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, *56*(2), 407-415. [CrossRef]

Ayesha Malik has a bachelor's degree in computer science from the University of Engineering and Technology, Taxila, in 2020. She has also been honored with a shield on account of her outstanding performance in her degree course. In addition to that, she has earned many certificates for taking part in curricular as well as extra-curricular activities during her education career. She is currently enrolled in an MS CS course at UET Taxila. (Email: 22-ms-csc-2@students.uettaxila.edu.pk)



Dr. Rabbia Mahum received her PhD degree in Computer Science from the University of Engineering and Technology (UET) Taxila in 2024. She earned her MS degree in Computer Science from the same institution in 2018 and obtained her B.Sc. degree in Computer Science from COMSATS University Islamabad, Wah Campus, in 2015. She was awarded a gold medal during her undergraduate studies and has received multiple academic scholarships throughout her educational journey. Her primary research interests include Computer Vision, Deepfake Audio Synthesis and Detection, and Medical Imaging. Additionally, she currently serves as an Academic Editor for the IECE Journal of Image Analysis and Processing. (Email: rabbia.mahum@uettaxila.edu.pk)